Outlier Robust Inference in a General Class of Instrumental Variable Models

Job Market Paper - Latest Version Here

Jens Klooster*

October 22, 2023

Abstract

We consider the problem of providing outlier robust inference in a general class of instrumental variable models that includes the linear instrumental variable model and the endogenous probit model. It is well known that classical instrumental variable regression tools can be unreliable in this context due to outliers. Therefore, we propose a framework to construct weak instrument robust testing procedures that are also robust to outliers. The framework is constructed upon M-estimators and we show that classical weak instrument robust tests, such as the Anderson-Rubin test and the conditional likelihood ratio test can be obtained as special cases. As it turns out that the classical tests are not robust to outliers, we show how to construct robust alternatives. We investigate the robustness properties of the robust test statistics and show that their asymptotic distributions are the same as the classical test statistics. The theoretical results are corroborated by a simulation study. Finally, we revisit three empirical studies affected by outliers and apply the robust tests to re-evaluate their results.

Keywords: Influence function; Robust inference; Outlier; Robust test; Weak instrument.

*Department of Econometrics, Econometric Institute, Erasmus University Rotterdam, The Netherlands. E-mail: Klooster@ese.eur.nl

1 Introduction

The instrumental variable (IV) model is recognized as an important tool that can be used to draw causal inferences in non-experimental data (Angrist et al., 1996). Its applicability spans over different fields, for example, to study the effect of chemotherapy for advanced lung cancer in the elderly (Earle et al., 2001), the effect of education on labor market earnings (Angrist and Krueger, 1991) or the effect of foreign media on authoritarian regimes (Kern and Hainmueller, 2009). In each of these examples, an endogeneity problem occurs where the explanatory variable is correlated with the error term, causing the Least Squares (LS) estimator to be biased. The researchers then introduce instrumental variables to resolve this problem. The instrumental variables should be correlated with the (endogenous) explanatory variable and uncorrelated with the error term. When instrumental variables are available, then reliable estimation (and inference) is possible using IV estimators, such as the Two-Stage Least Squares (2SLS) estimator.

In practice, it is difficult to find instruments that satisfy both conditions. In particular, the instruments that researchers propose are oftentimes only weakly correlated with the endogenous explanatory variable (Andrews et al., 2019). When the instruments are weak, then classical IV estimators, such as the 2SLS estimator, are biased and t-tests based on IV estimators can fail to control the size of the tests and their associated confidence intervals are incorrect (Nelson and Startz, 1990; Bound et al., 1995). Since IV estimators are biased when the instruments are weak, it is common to draw inference in the IV model using a two-step procedure. In the first step, the strength of the instruments is tested by means of an F-test. When the first-stage F statistic is above a certain threshold (motivated by Staiger and Stock 1997 a cutoff value of 10 is common), the instruments are considered to be strong. When the instruments are strong, then in the linear IV model estimation is done with a 2SLS estimator and inference with a t-test. When the instruments are weak, then weak instrument robust tests are used (Anderson and Rubin, 1949; Kleibergen, 2002; Moreira, 2003).

It is well known from the methodological literature that the two-step procedure described above is sub-optimal and it is typically advised to directly rely on weak instrument robust tests, or, use larger cutoff values for the first-stage F (Andrews et al., 2019; Lee et al., 2022; Keane and Neal, 2023). In particular, in the just-identified linear model, i.e., with one endogenous variable and one instrumental variable, Moreira (2009) shows that the Anderson-Rubin (AR) test (Anderson and Rubin, 1949) is the uniformly most powerful among the class of unbiased tests. When there is one endogenous variable and multiple instrumental variables, then Andrews et al. (2006) show that the conditional likelihood ratio (CLR) test introduced by Moreira (2003) enjoys good power properties in the (homoskedastic) linear IV model.

Although, weak instrument robust testing procedures, such as the CLR test, are widely used, their (outlier) robustness properties have not been studied[1]. Recently, Young (2022) pointed out that many empirical IV studies are affected by a few outliers or by small clusters of deviating observations. In the two-step procedure, the first-stage F-statistic is not robust against outliers (Ronchetti, 1982). Hence, even one outlier is enough to inflate the first-stage F statistic so that the researcher is under the impression that the instrument is strong, while it is weak (see Klooster and Zhelonkin 2023 for an example). Therefore the incorrect inference procedure is used in the second stage. It is well known that the classical IV estimators, such as the 2SLS and Limited Information Maximum Likelihood (LIML) estimators, are not robust to outliers (Zhelonkin et al., 2012; Freue et al., 2013; Sølvsten, 2020; Jiao, 2022) and robust estimators were proposed. However, the problem of (outlier) robust inference, in particular in combination with weak instruments, has not been fully addressed yet.

Therefore, in this article, we propose a framework that allows researchers to construct (outlier) robust versions of the AR, K (Kleibergen, 2002) and CLR tests. The tests allow outlier and weak identification robust inference in a large class of models including the linear instrumental variable model, endogenous probit model and the endogenous Tobit model. We

[1] A notable exception is the article by Klooster and Zhelonkin (2023) who study the robustness properties of the Anderson-Rubin test in the linear instrumental variable model.

characterize the robustness properties of the tests by using the influence function approach (Hampel, 1974). We formally show that the robustness properties of the tests fully depend on the robustness properties of the estimators they are constructed upon. The robust tests are constructed using a minimum distance approach that only requires estimation of the reduced form parameters for the construction of the tests (Magnusson, 2010). The reduced form model does not contain endogenous regressors and therefore conventional robust regression estimators can be used for the estimation. This allows us to directly use the large literature on robust statistics (Hampel et al., 1986; Huber and Ronchetti, 2009) and software implementations thereof for the construction of the tests making them easy to implement in practice.

Conceptually, we formalize the outlier contamination using the Huber (1964) gross-error model $F_t = (1-t)F + tG$, where t is a (typically small) contamination proportion. We are interested in drawing inference for the central distribution F, but we assume that it only holds approximately and the data that we observe comes from F_t. The distribution G is assumed to be completely unknown and it is the source of the contamination. This approach is closely related to the literature on local misspecification (Kitamura et al., 2013; Andrews et al., 2017, 2020; Bonhomme and Weidner, 2022; Ichimura and Newey, 2022). In our case, the outliers can be viewed as a specific type of local contamination and more general conclusions could be made. However, in this work, we focus on contamination by outliers. The goal is to draw inference that is valid at the central model F, but also remains stable and reliable when the data is generated according to F_t. The classical tests are valid when all the data is generated according to F, but, as we show both theoretically and by simulation, when the data is generated according to F_t, then the they can easily be distorted. Our approach benefits from the parametric structure, e.g., computational simplicity and interpretability, while being resistant to small but harmful deviations from the assumed model F.

When the model F does not hold even approximately, then the use of non-parametric weak instrument robust inference (Andrews and Soares, 2007; Andrews and Marmer, 2008) is

advised. Note, however, that non-parametric procedures are typically not designed to be robust to outliers. For example, the sample mean is a non-parametric estimator of the expectation, but it is not robust as one outlier can make it arbitrarily biased (see Huber and Ronchetti 2009, p. 6 for further discussion). Similarly, weak instrument robust quantile methods introduced by Chernozhukov and Hansen (2008) and Jun (2008) are also not designed to be robust against outliers.

The setup of the article is as follows. In Section 2, we introduce the model and the notation. In Section 3, we revisit the minimum distance approach to construct weak instrument robust tests (Magnusson, 2010). In Section 4, we study the robustness properties of the minimum distance robust tests. As it turns out that the classical minimum distance robust tests are not robust to outliers, we show how to construct robust alternatives in Section 5. Then, in Section 6, we use a simulation study to study the small sample properties of the robust tests and compare their performance to their classical counterparts. In Section 7, we revisit three empirical studies and show how the robust tests can be used in practice. Finally, in Section 8, we conclude.

2 Instrumental Variables Models

We assume that the data is generated according to a limited dependent instrumental variable regression model. The model consists of a structural equation (1) and a first-stage equation (2):

$$y^* = \beta x^* + w^\top \gamma_1 + u, \tag{1}$$

$$x^* = w^\top \gamma_2 + z^\top \pi + v, \tag{2}$$

where y^* and x^* are latent endogenous variables, z is a $k \times 1$ random vector of instrumental variables and w is a $p \times 1$ random vector of control variables. We assume that $\gamma_1, \gamma_2 \in \mathbb{R}^p$ are parameter vectors that both, when necessary, include an intercept. We assume $\pi \in \mathbb{R}^k$ and $\beta \in \mathbb{R}$. We assume that the errors (u, v) have mean zero, and we assume that the errors are

uncorrelated with the instruments z and control variables w.

We are interested in testing the hypothesis

$$H_0 \colon \beta = \beta_0 \text{ against } H_1 \colon \beta \neq \beta_0 \tag{3}$$

in the model (1) - (2). We do this without imposing further assumptions about the identification strength, i.e., the magnitude of the correlation between the instrument z and the endogenous regressor x^*. In practice, when this correlation is sufficiently small, then conventional approximations to the distribution of instrumental variable estimators, like the two-stage least squares estimator, are unreliable (Bound et al., 1995; Andrews et al., 2019). Consequently, this issue of weak identification can cause the estimators to be badly biased and make corresponding inferential procedures unreliable. For this reason, we construct a general class of outlier and weak identification robust tests based on a minimum distance approach (Magnusson, 2010). These tests are constructed upon estimators in the reduced form model. The reduced form model can be obtained by substituting (2) into (1). We obtain

$$y^* = z^\top \delta + w^\top \gamma + \epsilon.$$

with $\delta = \pi\beta$, $\gamma = \gamma_1 + \gamma_2\beta$ and $\epsilon = v\beta + u$. To further simplify the notation, we assume that $\gamma_1 = \gamma_2 = 0$ so that the control variables can be dropped from the model. Hence, we obtain the following reduced form equations

$$y^* = z^\top \delta + \epsilon, \tag{4}$$

$$x^* = z^\top \pi + v. \tag{5}$$

Define $\theta = (\delta^\top, \pi^\top)^\top$, then we assume that the model (4) - (5) is governed by F_θ. In practice, we assume we observe i.i.d. data $d_i = (f(y_i^*), h(x_i^*), z_i^\top)^\top$, which are random samples from $d = (f(y^*), h(x^*), z^\top)^\top$. We further define $y_i = f(y_i^*)$ and $x_i = h(x_i^*)$. The functions f and h are assumed to be known and allow us to model a general class of models. For example, when f and h are identity functions, then it is the linear instrumental variable model. When

$f(a) = \mathbf{1}(\{a > 0\})$, where $\mathbf{1}(\cdot)$ denotes the indicator function, and h is the identity function then we obtain the endogenous probit model as shown in Example 1. When $f(a) = \max(a, 0)$ and h is the identity function, then we obtain the endogenous tobit model as shown in Example 2. We define $Y, X \in \mathbb{R}^n$, which are vectors with entries y_i and x_i respectively, and $Z \in \mathbb{R}^{n \times k}$, the matrix with rows $z_i^\top$.

Example 1 (endogenous probit model)**.** *When we assume that* $f(a) = \mathbf{1}(\{a > 0\})$ *and* $h(a) = a$*, then we observe* $d_i = (y_i, x_i, z_i^\top)^\top$ *according to an endogenous probit model:*

$$y_i = \mathbf{1}(z_i^\top \delta + \epsilon_i > 0),$$
$$x_i = z_i^\top \pi + v_i,$$

where the error terms are realizations from a bivariate normal distribution, with variances $\sigma_v^2, \sigma_\epsilon^2 = 1$ *and correlation* ρ*.*

Example 2 (endogenous tobit model)**.** *When we assume that* $f(a) = \max(a, 0)$ *and* $h(a) = a$*, then we observe* $d_i = (y_i, x_i, z_i^\top)^\top$ *according to an endogenous tobit model:*

$$y_i = \max(z_i^\top \delta + \epsilon_i, 0),$$
$$x_i = z_i^\top \pi + v_i,$$

where the error terms are realizations from a bivariate normal distribution, with variances $\sigma_v^2, \sigma_\epsilon^2$ *and correlation* ρ*.*

3 Minimum Distance Approach

We construct outlier robust tests by using a minimum distance approach (Magnusson, 2010). The intuition behind this approach is that the structural parameter of interest β can be linked to the reduced form parameter θ via a distance function $r(\theta, \beta)$. Under the null hypothesis $H_0 \colon \beta = \beta_0$, we can use the reduced form equations (4) - (5) to show that

$$y^* - x^* \beta_0 = z^\top (\delta - \pi \beta_0) + \epsilon - \beta_0 v = u,$$

as $\delta - \pi\beta_0 = 0$ and $\epsilon = v\beta_0 + u$ under the null. Therefore, if we define $r(\theta, \beta_0) = \delta - \pi\beta_0$, then we can test whether the null hypothesis is true by checking whether $r(\theta, \beta_0) = 0$, which only requires estimation of the parameter θ. Thus, under Assumption 1, a test statistic can be constructed that tests whether $H_0 \colon \beta = \beta_0$ by checking if $r(\hat{\theta}, \beta_0) = 0$.

Assumption 1. *We assume* $\hat{\theta} \xrightarrow{p} \theta$ *and* $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Omega)$, *with* $\Omega = \begin{pmatrix} \Omega_{\delta\delta} & \Omega_{\delta\pi} \\ \Omega_{\pi\delta} & \Omega_{\pi\pi} \end{pmatrix}$, *a symmetric, positive definite covariance matrix.*

Proposition 1. *Given a hypothesized value* β_0, *we define*

$$S(\hat{\theta}, \beta_0) = n r(\hat{\theta}, \beta_0)^\top \Omega(\beta_0)^{-1} r(\hat{\theta}, \beta_0), \tag{6}$$

with $\Omega(\beta_0) = \Omega_{\delta\delta} - \beta_0(\Omega_{\delta\pi} + \Omega_{\pi\delta}) + \beta_0^2 \Omega_{\pi\pi}$. *Then, under the null hypothesis* $H_0 \colon \beta = \beta_0$ *and Assumption 1, it holds that* $S(\beta_0) \xrightarrow{d} \chi^2(k)$.

The limiting distribution of the statistic (6) does not depend on the nuisance parameter π and therefore it remains reliable irrespective of the strength of the identification. However, the test $\phi_S(\beta_0) = \mathbf{1}\{S(\beta_0) > \chi^2_{1-\alpha}(k)\}$ loses power when the number of instrumental variables k grows. This happens, because the degrees of freedom of the test increase, while the dimension of the parameter of interest β remains only one dimensional. A solution to this problem is to incorporate more information from $\hat{\pi}$ as it can show which deviations of $r(\hat{\theta}, \beta_0)$ from zero are actually informative. Unfortunately, this would cause the asymptotic distribution of the test to depend on the parameter π, which could lead to weak identification problems.

To circumvent this problem, Kleibergen (2002) and Moreira (2003) introduced tests that remain powerful when there more instrumental variables. The tests are constructed upon an alternative estimator of π defined as

$$D(\hat{\theta}, \beta_0) = \hat{\pi} - (\Omega_{\pi\delta} - \beta_0 \Omega_{\pi\pi}) \Omega(\beta_0)^{-1} r(\hat{\theta}, \beta_0).$$

Lemma 1. *Under the null hypothesis and Assumptions 1, it holds that*

$$\sqrt{n} \begin{Bmatrix} r(\hat{\theta}, \beta_0) \\ D(\hat{\theta}, \beta_0) \end{Bmatrix} \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{Bmatrix} \Omega(\beta_0) & 0 \\ 0 & \Lambda(\beta_0) \end{Bmatrix} \right], \tag{7}$$

with $\Lambda(\beta_0) = \Omega_{\pi\pi} - (\Omega_{\pi\delta} - \beta_0\Omega_{\pi\pi})\Omega(\beta_0)^{-1}(\Omega_{\delta\pi} - \beta_0\Omega_{\pi\pi})$.

Thus, as $r(\hat{\theta}, \beta_0)$ and $D(\hat{\theta}, \beta_0)$ are uncorrelated and normally distributed, they are asymptotically independent. Moreover, the nuisance parameter π only enters the limiting distribution via $D(\hat{\theta}, \beta_0)$. The conditional limiting distribution of $r(\hat{\theta}, \beta_0)$ given $D(\hat{\theta}, \beta_0)$ does not depend on π and is therefore not affected by the identification strength. This way, (conditional) test statistics can be constructed that incorporate information about $\hat{\pi}$ so that the test can remain powerful even when the number of instruments increase without being affected by potential identification failures. This brings us to the minimum distance versions of the K statistic (Kleibergen, 2002) and conditional likelihood ratio statistic (Moreira, 2003) defined as follows

$$K(\hat{\theta}, \beta_0) = nr(\hat{\theta}, \beta_0)^\top D(\hat{\theta}, \beta_0) \left\{ D(\hat{\theta}, \beta_0)^\top \Omega(\beta_0) D(\hat{\theta}, \beta_0) \right\}^{-1} D(\hat{\theta}, \beta_0)^\top r(\hat{\theta}, \beta_0), \tag{8}$$

$$W(\hat{\theta}, \beta_0) = nD(\hat{\theta}, \beta_0)^\top \Lambda(\beta_0)^{-1} D(\hat{\theta}, \beta_0),$$

$$CLR(\hat{\theta}, \beta_0) = \frac{1}{2}\left[S(\beta_0) - W(\beta_0) \; + \; \sqrt{\{S(\beta_0) - W(\beta_0)\}^2 + 4W(\beta_0)K(\beta_0)} \right]. \tag{9}$$

Proposition 2. *Under the null hypothesis* $H_0\colon \beta = \beta_0$ *and Assumption 1 it holds that* $K(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(1)$. *Under the null hypothesis* $\tilde{H}_0\colon \pi = 0$ *and* $\beta = \beta_0$ *and Assumption 1, it holds that* $W(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(k)$. *Under the null hypothesis* $H_0\colon \beta = \beta_0$ *and Assumption 1, then conditional on* $D(\hat{\theta}, \beta_0) = D$ *and with* $W = D^\top \Lambda(\beta_0)^{-1} D$, *it holds that*

$$CLR(\hat{\theta}, \beta_0) \xrightarrow{d} \frac{1}{2}\left[\chi^2(k-1) + \chi^2(1) - W + \sqrt{\{\chi^2(k-1) + \chi^2(1) + W\}^2 - 4W\chi^2(k-1)} \right], \tag{10}$$

where $\chi^2(k-1)$ *and* $\chi^2(1)$ *are independent chi-squared distributed random variables.*

Proposition 2 allows us to define the following tests

$$\phi_K(\beta_0) = \mathbf{1}\{K(\hat{\theta}, \beta_0) > \chi^2_{1-\alpha}(1)\},$$

$$\phi_W(\beta_0) = \mathbf{1}\{W(\hat{\theta}, \beta_0) > \chi^2_{1-\alpha}(k)\},$$

$$\phi_{CLR}(\beta_0) = \mathbf{1}\left[CLR(\hat{\theta}, \beta_0) > c_{1-\alpha}\{D(\hat{\theta}, \beta_0)\} \right],$$

where $c_{1-\alpha}\{D(\hat{\theta}, \beta_0)\}$ denotes the conditional $1-\alpha$ quantile of the (conditional) asymptotic distribution given in (10). The critical values and confidence sets of the CLR test can easily be computed using simulation and test inversion (Moreira, 2003; Andrews et al., 2019). That is, we find all the values of β_0 for which the data does not reject the null hypothesis. The confidence set is then $\{\beta_0 \mid \beta_0 \in \mathbb{R} \text{ and } \phi_{CLR}(\beta_0) = 0\}$.

In Example 3, we illustrate that if we use the LS estimator for $\hat{\theta}$, then $S(\hat{\theta}, \beta_0)$ is the Anderson-Rubin test statistic (Anderson and Rubin, 1949), $K(\hat{\theta}, \beta_0)$ is the K statistic (Kleibergen, 2002), and $CLR(\hat{\theta}, \beta_0)$ is the conditional likelihood ratio test statistic (Moreira, 2003).

Magnusson (2010) introduced extensions of these test statistics in limited dependent variable models (4) - (5). In such models, θ is estimated using an estimator tailored to the specific limited dependent variable model under consideration. For instance, in the case of the endogenous Tobit model (see Example 2), Magnusson (2010) uses the symmetrically censored least squares estimator (Powell, 1986) to estimate δ and the LS estimator for π.

To evaluate the robustness of these minimum distance test statistics, we restrict our focus to the class of minimum distance test statistics (6), (8) and (9) where θ is estimated with an M-estimator (Huber, 1964). This class encompasses numerous minimum distance tests, such as the classical Anderson-Rubin, K and CLR tests based on the LS estimator, but also many of the tests described in Magnusson (2010). In the following section, we provide a brief introduction to M-estimators.

Example 3. *When we estimate θ with the LS estimator, and we use*

$$\hat{\Omega} = (Z^\top Z)^{-1} \begin{pmatrix} \hat{\sigma}_\epsilon^2 & \hat{\sigma}_{\epsilon,v} \\ \hat{\sigma}_{v,\epsilon} & \hat{\sigma}_v^2 \end{pmatrix},$$

where $\hat{\sigma}_\epsilon^2, \hat{\sigma}_{\epsilon,v}$ and $\hat{\sigma}_v^2$ denote consistent estimates of $\sigma_\epsilon^2, \sigma_{\epsilon,v}$ and σ_v^2. Then we obtain

$$\begin{aligned} r(\hat{\theta}, \beta_0) &= (Z^\top Z)^{-1} Z^\top (Y - \beta_0 X), \\ D(\hat{\theta}, \beta_0) &= (Z^\top Z)^{-1} Z^\top \left\{ X - \frac{\hat{\sigma}_{u,v}}{\hat{\sigma}_u^2} (Y - X\beta_0) \right\}, \end{aligned}$$

where $\hat{\sigma}_u$ and $\hat{\sigma}_{u,v}$ denote consistent estimates of σ_u and $\sigma_{u,v}$. Plugging these into the statistics and yields the classical AR, K and Wald statistics:

$$
\begin{aligned}
S(\hat{\theta}, \beta_0) &= \frac{(Y - \beta_0 X)^\top Z (Z^\top Z)^{-1} Z^\top (Y - \beta_0 X)}{\hat{\sigma}_u^2}, \\
K(\hat{\theta}, \beta_0) &= \frac{1}{\hat{\sigma}_u^2} (Y - X\beta_0)^\top P_{ZD(\hat{\theta},\beta_0)} (Y - X\beta_0), \\
W(\hat{\theta}, \beta_0) &= \left\{ X - \frac{\hat{\sigma}_{u,v}}{\hat{\sigma}_u^2} (Y - X\beta_0) \right\}^\top \left\{ \left(\hat{\sigma}_v^2 - \frac{\hat{\sigma}_{u,v}}{\hat{\sigma}_u^2} \right) Z^\top Z \right\}^{-1} \times \\
&\quad \left\{ X - \frac{\hat{\sigma}_{u,v}}{\hat{\sigma}_u^2} (Y - X\beta_0) \right\},
\end{aligned}
$$

with $P_{ZD(\hat{\theta},\beta_0)} = ZD(\hat{\theta}, \beta_0) \left\{ \hat{D}(\hat{\theta}, \beta_0)^\top Z^\top Z D(\hat{\theta}, \beta_0) \right\}^{-1} D(\hat{\theta}, \beta_0)^\top Z^\top$. Plugging these statistics into (9) yields the CLR statistic introduced by Moreira (2003).

3.1 M-estimators

We construct minimum distance robust test statistics (6), (8) and (9) based on an M-estimator $\hat{\theta} = \begin{pmatrix} \hat{\delta}^\top & \hat{\pi}^\top \end{pmatrix}^\top$ that solves

$$
\frac{1}{n} \sum_{i=1}^{n} \Psi(d_i, \hat{\theta}) = 0. \tag{11}
$$

As we showed in Example 3, in case of the classical CLR statistic introduced by Moreira (2003), it is constructed upon LS estimators, which can be obtained by using the score function

$$
\Psi(d_i, \hat{\theta}) = \begin{Bmatrix} (y_i - z_i^\top \hat{\delta}) z_i \\ (y_i - z_i^\top \hat{\pi}) z_i \end{Bmatrix}. \tag{12}
$$

Under general regularity conditions $(A.1) - (A.8)$ stated in Assumption 2, M-estimators are consistent at the model and asymptotically normally distributed (Clarke, 1983, 1986). Note, when Assumption 2 holds and we construct a minimum distance robust test on an M-estimator $\hat{\theta}$, then the results presented in Proposition 1 and 2 also hold.

Assumption 2 (Adapted from Heritier and Ronchetti (1994))**.** *Let F be any arbitrary distribution on $\mathbb{R}^{2k}$ and define*

$$
K_F(\theta) = \int \Psi(d, \theta) dF.
$$

(A.1): $K_F(\theta)$ *exists at least on a (nondegenerate) open set* $\mathcal{O}$.

(A.2): *There exists* $\theta^* \in \mathcal{O}$ *satisfying* $K_F(\theta^*) = 0$.

(A.3): $\int \Psi(d,\theta) dF_\theta = 0$ *(Fisher consistency).*

(A.4): $\Psi(d,\theta)$ *is a* $2k \times 1$ *vector function that is continuous and bounded on* $\mathcal{D} \times \Theta$, *where* Θ *is some nondegenerate compact interval containing* θ^*.

(A.5): $\Psi(d,\theta)$ *is locally Lipschitz in* θ *about* θ^*, *i.e., there exists a constant* α *such that*

$$||\Psi(d,\theta) - \Psi(d,\theta^*)|| \leq \alpha ||\theta - \theta^*)||$$

uniformly in $d \in \mathcal{D}$ *and for all* θ *in a neighbourhood of* θ^*.

(A.6): *The generalized Jacobian* $\partial K_F(\theta)$ *is of maximal rank at* $\theta = \theta^*$.

(A.7): *Given* $\Delta > 0$, *there exists* $\epsilon > 0$ *such that for all distributions in a* ϵ *neighborhood of* F, $\sup_{\theta \in \Theta} ||K_G(\theta) - K_F(\theta)|| > \delta$ *and* $\partial K_G(\theta) \subset \partial K_F(\theta) + \Delta B$, *uniformly in* $\theta \in \Theta$, *where* B *is the unit ball of* $2k \times 2k$ *matrices.*

(A.8): $K_F(\theta)$ *has at least a continuous derivative* $(\partial/\partial\theta)K_F(\theta)$ *at* $\theta = \theta^*$.

More details on the Assumptions $(A.1) - (A.8)$ can be found in Heritier and Ronchetti (1994) and Clarke (1983, 1986). Here we briefly mention some important properties of the assumptions that are also explained by Heritier and Ronchetti (1994). The first two assumptions $(A.1) - (A.2)$ establish the existence of the functional $\theta^* = T(F)$ that defines an M-estimator through $T(F_n)$, where F_n denotes the empirical distribution function. Assumption $(A.3)$ makes sure the estimator is Fisher consistent. Fisher Consistency is a standard assumption in the robustness literature (Hampel et al., 1986). Assumptions $(A.4) - (A.8)$ make sure the estimator is Fréchet differentiable. When an estimator is Fréchet differentiable, then it has an influence function and the estimator is asymptotically normal. More details on the influence function and its significance are given in the next section.

4 Robustness properties

In this section, we study the robustness properties of test statistics (6), (8) and (9) when they are constructed upon an M-estimator $\hat{\theta}$ that solves (11). We investigate the robustness properties of the test statistics using the influence function approach (Hampel, 1974). For this reason, in Section 4.1, we first go over the definition of the influence function and explain its significance.

4.1 Influence function

To study the robustness properties of an estimator, we analyze its influence function (Hampel, 1974). Let T denote a Fisher consistent statistical functional, then the influence function is defined as

$$\mathrm{IF}(d; T, F) = \lim_{t \downarrow 0} \frac{T\left\{(1-t)F + t\Delta_d\right\} - T(F)}{t}, \tag{13}$$

where Δ_d is a point mass at d. The influence function describes the effect of an infinitesimal contamination at the point d on the functional, standardized by the mass of the contamination (Hampel et al., 1986). Therefore, when the influence function is unbounded, then we say that the functional is not (locally) robust. When a statistical functional $T(F)$ is sufficiently regular, then we can write out a von Mises (1947) expansion and obtain

$$T(G) = T(F) + \int \mathrm{IF}(d; T, F) d(G - F) + o(||G - F||_\infty), \tag{14}$$

where $||\cdot||_\infty$ denotes the supremum norm. If we now consider the family of distributions $F_t = (1-t)F + tG$ with $t \geq 0$ close to zero, then we see that the influence function can be used to linearize the bias of the functional $T(F)$ as we obtain the approximation

$$\sup_G ||T(F_t) - T(F)|| \approx t \sup_d ||\,\mathrm{IF}(d; T, F)||. \tag{15}$$

Hence, when the influence function of a statistical functional T is not bounded, then the maximum bias in a neighborhood of F can be infinite, even when the contamination proportion

t is small. Therefore, for a statistical functional to be (locally) robust, a bounded influence function is required.

Note, the influence function of an estimator is closely connected to its asymptotic variance (Hampel, 1974). When the functional T is sufficiently regular, then plugging in the empirical distribution function F_n in (14) and multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\{T(F_n) - T(F)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(d_i; T, F) + o(\sqrt{n}||F_n - F||_\infty).$$

Therefore, as $n \to \infty$, we have

$$\sqrt{n}\{T(F_n) - T(F)\} \xrightarrow{d} \mathcal{N}\{0, \int \mathrm{IF}(d; T, F)\, \mathrm{IF}(d; T, F)^\top dF\}.$$

4.2 Influence function of an M-estimator

We can write the estimator $\hat{\theta}$ in functional form as follows. Let F be an arbitrary distribution function on $\mathbb{R}^{2k}$. Define $\theta(F)$ the solution to

$$\int \Psi\{d, \theta(F)\} dF = 0.$$

Then we define $\hat{\theta} = \theta(F_n)$, where F_n denotes the empirical distribution function, which leads to (11). Using the functional form, we can derive the influence function as follows. Let $F_t = (1-t)F_\theta + t\Delta_d$, then we have

$$\int \Psi\{d, \theta(F_t)\} dF_t = 0 \implies t \int \Psi\{d, \theta(F_t)\} d(\Delta_d - F_\theta) + \int \Psi\{d, \theta(F_t)\} dF_\theta.$$

Taking a derivative with respect to t, and letting $t = 0$, we obtain

$$\mathrm{IF}(d, \theta(\cdot), F_\theta) = \left\{ - \int (\partial/\partial\theta) \Psi(d, \theta) dF_\theta \right\}^{-1} \Psi(d, \theta). \tag{16}$$

From (16) it becomes clear that the influence function of the M-estimator $\hat{\theta}$ is proportional to the general score function Ψ, the influence function of the M-estimator $\hat{\theta}$ is only bounded when the score function Ψ is bounded. In particular, as we show in Example 4, when we estimate θ with a LS estimator, then the influence function is not bounded.

Example 4. *The influence function of the LS estimator is*

$$IF(d_i, \theta_{LS}(\cdot), F_\theta) = \left\{ \int \begin{pmatrix} zz^\top & 0 \\ 0 & zz^\top \end{pmatrix} dF_\theta \right\}^{-1} \begin{Bmatrix} (y_i - z_i^\top \delta) z_i \\ (x_i - z_i^\top \pi) z_i \end{Bmatrix} \propto \begin{Bmatrix} (y_i - z_i^\top \delta) z_i \\ (x_i - z_i^\top \pi) z_i \end{Bmatrix},$$

which is unbounded in y_i, x_i and z_i. Therefore, an outlier in any variable can arbitrarily bias the LS estimator. Note, the asymptotic variance of the LS estimator is

$$\begin{aligned} \Omega &= \int IF(d, \theta_{LS}(\cdot), F_\theta) IF(d, \theta_{LS}(\cdot), F_\theta)^\top dF_\theta \\ &= \left\{ \int \begin{pmatrix} zz^\top & 0 \\ 0 & zz^\top \end{pmatrix} dF_\theta \right\}^{-1} \left\{ \int \begin{pmatrix} \epsilon^2 zz^\top & \epsilon v zz^\top \\ \epsilon v zz^\top & v^2 zz^\top \end{pmatrix} dF_\theta \right\} dF_\theta \left\{ \int \begin{pmatrix} zz^\top & 0 \\ 0 & zz^\top \end{pmatrix} dF_\theta \right\}^{-\top}, \end{aligned}$$

which is the usual "sandwich" form. When we further assume independence between the error terms and the instrumental variables, we obtain

$$\Omega = \left\{ \int \begin{pmatrix} zz^\top & 0 \\ 0 & zz^\top \end{pmatrix} dF_\theta \right\}^{-1} \int \begin{pmatrix} \epsilon^2 & \epsilon v \\ \epsilon v & v^2 \end{pmatrix} dF_\theta.$$

4.3 Influence function of the test statistics

To derive the influence function of the test statistics (6), (8) and (9), we first need to write them in functional form. This requires us to first write $r(\hat{\theta}, \beta_0)$ and $D(\hat{\theta}, \beta_0)$ in functional form. For an arbitrary distribution function F on $\mathbb{R}^{2k}$, we define

$$\begin{aligned} r(F, \beta_0) &= \delta(F) - \pi(F)\beta_0, \\ D(F, \beta_0) &= \pi(F) - (\Omega_{\pi\delta} - \beta_0 \Omega_{\pi\pi}) \Omega(\beta_0)^{-1} r(F, \beta_0), \end{aligned}$$

where $\delta(F) = \theta(F)_{(1)}$ and $\pi(F) = \theta(F)_{(2)}$, which denote the first and second k-dimensional components of $\theta(F)$. Then we obtain $r(\hat{\theta}, \beta_0) = r(F_n, \beta_0)$ and $D(\hat{\theta}, \beta_0) = D(F_n, \beta_0)$. The influence functions of the functionals $r(\cdot, \beta_0)$ and $D(\cdot, \beta_0)$ can easily be derived and are given by

$$\mathrm{IF}\{d, r(\cdot, \beta_0), F_\theta\} = \mathrm{IF}\{d, \delta(\cdot), F_\theta\} - \mathrm{IF}\{d, \delta(\cdot), F_\theta\}\beta_0, \tag{17}$$

$$\mathrm{IF}\{d, D(\cdot, \beta_0), F_\theta\} = \mathrm{IF}\{d, \delta(\cdot), F_\theta\} - (\Omega_{\pi\delta} - \beta_0 \Omega_{\pi\pi}) \Omega(\beta_0)^{-1} \mathrm{IF}\{d, r(\cdot, \beta_0), F_\theta\}. \tag{18}$$

Similarly, we can introduce functional versions $S(F,\beta_0), K(F,\beta_0)$ and $CLR(F,\beta_0)$ of the test statistics (6), (8) and (9) and study their robustness properties. However, the test statistics are not Fisher consistent. Consequently, when we derive the influence function of the test statistics, they will always be zero. To circumvent this problem, we can analyze the influence function of the square root of the statistics (Hampel et al., 1986). In Proposition 3, we derive the influence function of the minimum distance CLR statistic (9).

Proposition 3. *Under the null hypothesis $\beta = \beta_0$, and (A.1) - (A.8) of Assumption 2, the influence function of the CLR statistic, conditional on $D(\hat{\theta},\beta_0) = D$, is*

$$\mathrm{IF}\left\{d; \sqrt{CLR(\cdot,\beta_0)}, F_\theta\right\} = \begin{cases} \mathrm{IF}\left\{d; \sqrt{S(\cdot,\beta_0)}, F_\theta\right\}, & \text{if } D = 0, \\ \mathrm{IF}\left\{d; \sqrt{K(\cdot,\beta_0)}, F_\theta\right\}, & \text{if } D \neq 0. \end{cases}$$

The influence functions of the S and K statistics, conditional on $D(F_n,\beta_0) = D$, are given by

$$\begin{aligned} \mathrm{IF}\left\{d; \sqrt{S(\cdot,\beta_0)}, F_\theta\right\} &= \sqrt{\mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\}^\top \Omega(\beta_0)^{-1}\, \mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\}}, \\ \mathrm{IF}\left\{d; \sqrt{K(\cdot,\beta_0)}, F_\theta\right\} &= \sqrt{\mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\}^\top D \left\{D^\top \Omega(\beta_0) D\right\}^{-1} D^\top\, \mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\}}, \end{aligned}$$

where

$$\mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\} = \mathrm{IF}(d; \delta(\cdot), F_\theta) - \beta_0\, \mathrm{IF}(d; \pi(\cdot), F_\theta).$$

Proposition 3 shows that the influence function of the CLR statistic depends on the S and K statistics. This is unsurprising because when $D = 0$, then the CLR statistic is the S statistic so their influence functions must be the same in that case. When $D \neq 0$, then Moreira (2003) shows that the classical CLR test statistic converges to the Kleibergen (2002) K statistic, which is why we see that same behavior in the influence function.

In Proposition 3, we see that the influence functions of the minimum distance robust tests ultimately depend on the influence function of the estimator $\hat{\theta} = (\hat{\delta}^\top, \hat{\pi}^\top)^\top$. We immediately see that the influence function of the tests are only bounded, when the influence function of the estimators are bounded. In other words, the minimum distance robust tests inherit the

robustness properties of the estimators they are constructed upon. When the influence function of the estimator $\hat{\delta}$ or $\hat{\pi}$ is unbounded, only one observation can bias them and corrupt all the test statistics.

Note, instead of analyzing the influence function of the test statistic, it is also possible to analyze the behavior of the level (or power) of the test under contamination and compare it to the nominal level α_0 (Heritier and Ronchetti, 1994). We show how this works for the test ϕ_S. Define

$$F_{t,n}^{L} = \left(1 - \frac{t}{n}\right) F_\theta + \frac{t}{n}\Delta_d.$$

Then, the level of the test ϕ_S can be approximated by a functional of the form

$$\alpha(F_{t,n}^{L}) = 1 - H_k\{\eta_{1-\alpha_0}; S(F_{t,n}^{L}, \beta_0)\},$$

with $H_k\{\eta_{1-\alpha_0}; S(F_{t,n}^{L}, \beta_0)\}$ the cumulative distribution function of a $\chi^2(k)$ distribution, and $\eta_{1-\alpha_0}$ is the $1-\alpha_0$ quantile of the $\chi^2(k)$ distribution, where α_0 denotes the nominal level. Then, direct application of Proposition 5 in Heritier and Ronchetti (1994), gives

$$\begin{aligned}\lim_{n\to\infty} \alpha(F_{t,n}^{L}) &= \alpha_0 + t^2 \mu \,\mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\}^\top \Omega(F_\theta,\beta_0)^{-1}\,\mathrm{IF}\{d; r(\cdot,\beta_0), F_\theta\} + \mathsf{o}(t^2) \qquad (19)\\ &= \alpha_0 + t^2 \mu \,\mathrm{IF}\left\{d; \sqrt{S(\cdot,\beta_0)}, F_\theta\right\}^2 + \mathsf{o}(t^2),\end{aligned}$$

with $\mu = -[(\partial/\partial\delta) H_k(\eta_{1-\alpha_0}; \delta)]_{\delta=0}$. From (19) it becomes clear that if we want the level to be stable in a neighborhood around the hypothesis, then we need to bound the influence function of the test statistic and consequently the influence function of the estimator $\hat{\theta}$. Similar results can be obtained tests ϕ_K, ϕ_W and ϕ_{CLR}.

Intuitively, it is not surprising that the stability of the level and power of the minimum distance tests are closely to the robustness properties of the test statistic. When the influence function of the estimator $\hat{\theta}$ is not bounded, then an outlier can arbitrarily bias this estimator. This bias might result in a large movement of the test statistic and consequently large distortions in the empirical level or power.

4.4 Robustness properties in specific models

The result in Proposition 3 holds for a wide variety of models and ultimately depend on the robustness properties of the estimator $\hat{\theta}$. Therefore, in this section, we study the influence functions of different estimators $\hat{\theta}$ in several models that are used in practice (Finlay and Magnusson, 2009; Magnusson, 2010). Consequently, using Proposition 3, we can study the robustness properties of the (classical) test statistics (6), (8) and (9) in different instrumental variable models.

4.4.1 The linear instrumental variable model

When we assume that $y = y^*$ and $x = x^*$ in the model (4) - (5), then we are in the setting of a linear instrumental variable model. As we showed in Example 3, using the least squares estimator to estimate θ and plugging the estimate into (6), (8) and (9) results in the Anderson-Rubin, K and conditional likelihood ratio test statistics (Anderson and Rubin, 1949; Kleibergen, 2002; Moreira, 2003). From (12), we see that the score function of the least squares estimator is not bounded. As the influence function is proportional to the score function, we conclude that the least squares estimator does not have a bounded influence function. Therefore, using Proposition 3, we conclude that the classical AR, K and CLR statistics do not have a bounded influence function. For this reason, one outlying observation can bias the estimator $\hat{\theta}$ which will corrupt the test statistics. In Example 5, we explicity derive the influence function of the Anderson-Rubin test.

Example 5. *When we use the LS estimator for $\hat{\theta}$, then we obtain*

$$\text{IF}\{d_i, r(\cdot, \beta_0), F_\theta\} = \left(\int zz^\top dF_\theta \right)^{-1} (y_i - x_i\beta_0).$$

Then, under the null hypothesis $\beta = \beta_0$, the influence function of the statistic $S(\hat{\theta}, \beta_0)$ is

$$\text{IF}\{d_i, \sqrt{S(\cdot, \beta_0)}, F_\theta\} = |y_i - x_i\beta_0| \left\{ \left(\int zz^\top dF_\theta \right)^{-\top} \Omega(\beta_0)^{-1} \left(\int zz^\top dF_\theta \right)^{-1} \right\}^{1/2},$$

which is unbounded.

4.4.2 Endogenous probit, logit and Poisson models

We now study the robustness properties of the minimum distance robust tests in probit, logit and Poisson models with endogeneity. In this case, we assume that the first stage (5) is fully observed so that $x = x^*$. We assume that the second stage equation (4) can be modeled according to a generalized linear model where we assume that y is distributed according to a distribution that comes from the exponential family such that $\mathbb{E}[y] = \mu$ and $\mathbb{V}[y] = V(\mu)$ and

$$\eta = g(\mu) = z^\top \delta, \tag{20}$$

where g is the link function (Nelder and Wedderburn, 1972). In this case, we can estimate π via least squares estimation and δ with a quasi-likelihood estimator. The estimator $\hat{\theta}$ is a solution of the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(d_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \frac{(y_i - \mu_i)}{V(\mu_i)}\mu_i' \\ (x_i - z_i^\top \hat{\pi}) z_i \end{bmatrix} = 0, \tag{21}$$

where μ_i' denotes the derivative of μ_i with respect to δ. We obtain

$$\mathrm{IF}(d_i, \hat{\theta}, F_\theta) \propto \begin{bmatrix} \frac{(y_i - \mu_i)}{V(\mu_i)}\mu_i' \\ (x_i - z_i^\top \pi) z_i \end{bmatrix}.$$

The score function is unbounded in x_i and z_i and also potentially unbounded in y_i. Therefore, by Proposition 3, the influence functions of the minimum distance robust test will also not be bounded. In Example 6, we work out the details for the endogenous probit model introduced in Example 1.

Example 6 (Endogenous probit model (cont.))**.** *In case of the endogenous probit model, we have* $y = \mathbf{1}(z^\top \delta + \epsilon > 0)$*. So that*

$$\mathbb{E}[y] = \mathbb{E}[\mathbf{1}(z^\top \delta + \epsilon > 0)] = \mathbb{P}[\epsilon < z^\top \delta] = \Phi(z^\top \delta),$$
$$\mathbb{V}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 = \Phi(z^\top \delta)\left\{1 - \Phi(z^\top \delta)\right\}.$$

Therefore, $\mu = \Phi(z^\top \delta)$*, where* $\Phi(\cdot)$ *denotes the cumulative standard normal distribution. For*

the influence function we obtain

$$\text{IF}(d_i, \hat{\theta}, F_\theta) \propto \begin{bmatrix} \frac{\{y_i - \Phi(z_i^\top \delta)\}\mu_i'}{\Phi(z_i^\top \delta)\{1-\Phi(z_i^\top \delta)\}} \\ (x_i - z_i^\top \pi) z_i \end{bmatrix}.$$

As $y_i \in \{0, 1\}$, the influence function is only unbounded in x_i and z_i.

4.4.3 Endogenous tobit model

We now study the robustness properties of minimum distance robust test in the endogenous tobit model, see Example 2. The first-stage equation (4) is fully observed $x = x^*$. The second-stage equation (4) is censored, as $y = \max(0, y^*)$. To construct the minimum distance robust tests in this model, Magnusson (2010) uses the symmetrically censored least squares estimator (Powell, 1986) for the estimation of δ. The estimator $\hat{\theta}$ is a solution to the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(d_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \mathbf{1}(z_i^\top \delta > 0) \min(y_i - z_i^\top \delta, z_i^\top \delta) \\ (x_i - z_i^\top \pi) z_i \end{bmatrix} = 0.$$

The influence function is proportional to the score, so that

$$\text{IF}(d_i, \hat{\theta}, F_\theta) \propto \begin{bmatrix} \mathbf{1}(z_i^\top \delta > 0) \min(y_i - z_i^\top \delta, z_i^\top \delta) \\ (x_i - z_i^\top \pi) z_i \end{bmatrix},$$

which is not bounded. By Proposition 3, the influence functions of the minimum distance robust test statistics is also not be bounded. Therefore, only one outlying observations will be able to bias the estimator $\hat{\theta}$ and also corrupt the minimum distance robust tests.

5 Robust inference

In the previous section, we showed that the robustness properties of the minimum distance robust tests depend on the robustness properties of the estimators they are constructed upon. To construct outlier robust tests, we need to use a bounded influence estimator for the estimation of θ. For this purpose, we can use the large literature on robust M-estimation (Hampel et al., 1986; Huber and Ronchetti, 2009) and software implementations thereof. For example, in case of endogenous probit, logit, Poisson or tobit models, we can use robust M-estimators designed for those models (Peracchi, 1990; Cantoni and Ronchetti, 2001).

In the next section, we show how to construct outlier robust minimum distance tests in a wide variety of models such as the linear instrumental variable model, the endogenous probit/logit model and the endogenous Poisson model. We do so, by assuming that both the first and second-stage equations follow can be modeled according to a generalized linear model. This allows us to leverage robust estimation techniques designed for those models (Cantoni and Ronchetti, 2001). Note, in order to construct minimum distance robust tests, we need to be able connect the latent and observed variables via the known functions f and h as explain in Section 2. This connection between the latent and the observed variables via the functions f and h is not possible for all generalized linear models. Fortunately, however, this is possible for many empirically relevant models, such as the linear instrumental variable model, the endogenous probit/logit model (see Example 6) and the endogenous Poisson model.

5.1 Robust estimation

We assume that the observations y and x are distributed according to a distribution that comes from the exponential family such that $\mathbb{E}[y] = \mu_1, \mathbb{V}[y] = V(\mu_1)$, $\mathbb{E}[x] = \mu_2$, $\mathbb{V}[x] = V(\mu_2)$, $g_1(\mu_1) = z^\top\delta$ and $g_2(\mu_2) = z^\top\pi$, where g_1 and g_2 are link functions. We write $\mu_1 = g_1^{-1}(z^\top\delta)$ and $\mu_2 = g_2^{-1}(z^\top\pi)$ and let μ_1' and μ_2' denote their derivatives, with respect to δ and π, respectively.

We consider a general class of M-estimators of Mallows type (Cantoni and Ronchetti, 2001). The estimator $\hat{\theta} = (\hat{\delta}^\top, \hat{\pi}^\top)^\top$ solves the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(d_i,\mu_i) = \frac{1}{n}\sum_{i=1}^{n}\begin{Bmatrix}\Psi_\delta(y_i,\mu_{1i})\\ \Psi_\pi(x_i,\mu_{2i})\end{Bmatrix} = 0, \tag{22}$$

where

$$\Psi_\delta(y_i,\mu_{1i}) = \nu_1(y_i,\mu_{1i})w_1(z_i)\mu_{1i}' - a_1,$$

$$\Psi_\pi(x_i,\mu_{2i}) = \nu_2(x_i,\mu_{2i})w_2(z_i)\mu_{2i}' - a_2.$$

The functions ν_1, ν_2, w_1 and w_2 are weight functions (specified below) and the constants a_1 and

a_2 make sure the estimator is Fisher consistent. We have

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\{\nu_1(y, \mu_1)|z = z_i\} w_1(z_i)\mu'_{1i},$$
$$a_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\{\nu_2(x, \mu_2)|z = z_i\} w_2(z_i)\mu'_{2i}.$$

In Example 7, we show that the class of Mallows type M-estimators includes the least squares and classical quasi-likelihood estimators (Wedderburn, 1974).

Example 7. *Assume that* $w_1(z_i) = w_2(z_i) = 1$ *and that* $\nu_1(y_i, \mu_{1i}) = \frac{y_i - \mu_{1i}}{V(\mu_{1i})}$ *and* $\nu_2(x_i, \mu_{2i}) = \frac{x_i - \mu_{2i}}{V(\mu_{2i})}$. *Then,* $a_1 = a_2 = 0$ *so that we obtain the classical quasi-likelihood estimators. If we further assume that both* g_1 *and* g_2 *are the identity functions, then* $\mu_1 = g_1^{-1}(z^\top\delta) = z^\top\delta$ *and* $\mu_2 = g_2^{-1}(z^\top\pi) = z^\top\pi$, *and* $\mu'_1 = \mu'_2 = z$. *In this case,* $\hat{\theta}$ *is a solution to*

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(d_i, \mu_i) = \frac{1}{n}\sum_{i=1}^{n} \begin{Bmatrix} (y_i - z_i^\top\delta)z_i \\ (x_i - z_i^\top\pi)z_i \end{Bmatrix} = 0,$$

which is solved by the least squares estimators of δ *and* π.

When we analyze the influence function (16) it becomes clear that if we want to bound the influence function of the estimator $\hat{\theta}$, then we need to bound the (general) score function Ψ and consequently the functions ν_1, ν_2, w_1 and w_2 it depends upon. In the next section, we give several options for the functions ν_1, ν_2, w_1 and w_2 introduced in (22) that can be used in practice and result in a bounded influence function of the estimator.

5.2 Applied guidance

In practice, we suggest using the functions

$$\nu_1(y_i, \mu_{1i}) = \psi(r_{1i})\frac{1}{V^{1/2}(\mu_{1i})},$$
$$\nu_2(x_i, \mu_{2i}) = \psi(r_{2i})\frac{1}{V^{1/2}(\mu_{2i})},$$

with $r_{1i} = \frac{y_i - \mu_{1i}}{V^{1/2}(\mu_{1i})}$ and $r_{2i} = \frac{x_i - \mu_{2i}}{V^{1/2}(\mu_{2i})}$. The function $\psi(\cdot)$ is an odd and bounded downweighting function. A first good option for the function $\psi(\cdot)$ is to use the Huber function:

$$\psi_H(r; c) = \begin{cases} r, & |r| \leq c, \\ c \cdot \text{sgn}(r), & |r| > c. \end{cases}$$

The tuning parameter c is typically set to $c = 1.345$ to ensure a high level of asymptotic efficiency. Another good option is to use Tukey's biweight function instead of the Huber function. Tukey's biweight function is defined as

$$\psi_T(r; c) = \begin{cases} r\left(1 - \frac{r^2}{c^2}\right)^2, & \text{for } |r| \leq c, \\ 0, & \text{for } |r| > c. \end{cases} \tag{23}$$

The tuning parameter c is typically set to $c = 4.685$ to ensure a high level of asymptotic efficiency. Tukey's biweight function downweights large outlying observations to zero, which makes it a very robust option. However, in contrast to the Huber loss function, Tukey's loss function is not convex, which makes the optimization problem more difficult.

To downweight outliers in the covariate space, we suggest using the weight function $w_1(z_i) = w_2(z_i) = \sqrt{1 - h_i}$, where h_i denotes the i-th diagonal element of the "hat" matrix $H = Z(Z^\top Z)^{-1}Z^\top$. Weights based on the hat matrix are easy to compute and ensure a bounded influence function. However, using these weights does not result in an estimator with a high breakdown point, i.e., the estimator can only resist a few outlying observations. If the user wants a high breakdown point, we suggest using weights based on the inverse Mahalanobis distance $d(z_i) = \sqrt{(z_i - \hat{\mu}_Z)^\top \hat{\Sigma}_Z^{-1}(z_i - \hat{\mu}_Z)}$, where the multivariate location $\hat{\mu}_Z$ and scatter $\hat{\Sigma}_Z$ of the matrix Z are estimated with the Minimum Covariance Determinant (MCD) estimator (Rousseeuw and Driessen, 1999). We then define

$$w(z_i) = \begin{cases} 1, & \text{if } d(z_i) \leq \tilde{c}, \\ \tilde{c}/d(z_i), & \text{if } d(z_i) > \tilde{c}, \end{cases}$$

and let $w_1(z_i) = w_2(z_i) = w(z_i)$. Under a normality assumption, the squared Mahalanobis distance is asymptotically χ^2 distributed. Therefore, in practice it is common to use the square root of the 0.95-quantile of the χ^2 distribution with k degrees of freedom, where k denotes the number of columns of the matrix Z. Economic data often contains discrete data. In that case, computation of the MCD estimator can become infeasible and we suggest using the weights based on the hat matrix.

In practice, the estimation of δ and π is typically done separately. In this case, the function

`glmrob()` from the R package `robustbase` can be used for robust estimation with weights as defined above in generalized linear models (Maechler et al., 2023) allowing us to obtain the estimates of $\pi, \delta, \Omega_{\delta\delta}$ and $\Omega_{\pi\pi}$. The user only needs to obtain an estimate of $\Omega_{\delta\pi}$. However, in a large class of empirically relevant models where $x = x^*$, a control function approach (Magnusson, 2010) can be used that does not require the direct estimation of $\Omega_{\delta\pi}$ (see Appendix B for details).

6 Simulation

In this section, we analyze the finite sample behavior of the outlier robust minimum distance tests and compare them to their classical counterparts. The outlier robust tests are based on Mallows type M-estimators using the Huber function and hat matrix weights as explained in Section 5.2 (see Appendix B for implementation details). We analyze the behavior of the tests in the linear instrumental variable model and the endogenous probit model. We consider a baseline environment without contamination and two contaminated environments, where we add outliers to the data.

In the baseline environment, we generate 10,000 random samples of $n = 250$ observations drawn from the following the following model:

$$x = z_1\pi_1 + 0 \cdot z_2 + 0.5 \cdot w + v, \tag{24}$$

$$y^* = \beta x + 0.3 \cdot w + u, \tag{25}$$

where $v = v_1 + 0.5 \cdot u$. We generate z_1, z_2, w, v_1 and u from independent standard normal distributions. The value of π_1 is set according to the first-stage F-statistic. We have $F \approx \frac{n\pi_1^2}{2\sigma_v^2}$ so that we let

$$\pi_1 = \sqrt{\frac{2\sigma_v^2 F^*}{n}},$$

allowing us to control the strength of the instruments in different simulation designs via F^*.

We study the behaviour of our robust tests in two different models. We consider the linear IV model in which we observe $y = y^*$ and the endogenous probit model in which we observe $y = \mathbf{1}(\{y^* > 0\})$. In the linear IV model, the classical tests are constructed upon the least square estimator, which leads to the classical AR, K and CLR tests (see Example 3). For the endogenous probit model, the classical test is constructed upon a quasi-likelihood estimate of δ and a least squares estimate of π. In both models, the robust tests are constructed upon a Mallows type M-estimator based on the Huber function and hat matrix weights as explain in Section 5.2. Note, we construct both the classical and robust tests using a control function approach as explain in Section 5.2 and Appendix B.

Next, we explain how we generate the environment that is contaminated by an outlier. In this environment, we generate data exactly the same as in the baseline environment. Then for each random sample, we change the first data row to $(y_1, x_1, z_{11}, z_{21}, w) = (25, 10, 3, 3, 3)$ for the linear IV model and for the endogenous probit model we change the first data row to $(y_1, x_1, z_{11}, z_{21}, w) = (0, 10, 3, 3)$ as the range of y is then limited to 0 or 1.

At last, we explain how we generate an environment with "distributional" contamination in the error terms. In this case, we replace the first 50 observations (20% of the observations) of v_1 and u with draws from a $t(3)$ distribution. This will sometimes generate large outlying observations in the error terms.

6.1 Sensitivity of the power

In Figure 1, we present the power curves of the robust CLR test compared to the classical CLR test in the two different models without outliers. We see that in each model, the classical test is more powerful than the robust test. This is expected, as the robust M-estimators are less efficient than the classical estimators when there are no outliers in the data. This leads to more powerful tests when there are no outliers in the data. However, we see in Figure 1 that the loss of power is not very large. Therefore, the trade-off between robustness and efficiency does not come at a large cost.

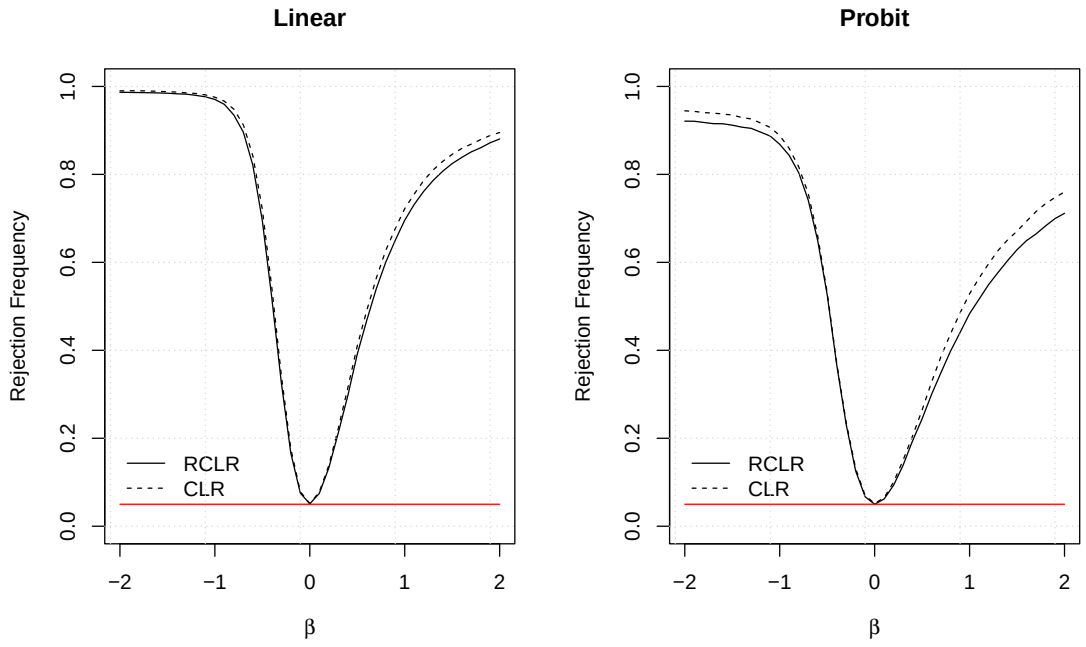

Figure 1: Power curves of the robust test and its classical version that tests $H_0\colon \beta = 0$ for various values of β in the linear instrumental variable model and the endogenous probit model. Baseline environment without contamination.

In Figure 2, we present the power curves of the robust test and the classical test in the two different models when there is an outlier in the data. First, we note that the power curves of the robust tests are almost not affected by the outlier when we compare them to the power curves in Figure 1. This is, however, not the case for the classical tests. In case of the linear IV model, the outlier biases the least squares estimators $\hat{\pi}$ and $\hat{\delta}$. As the classical test is constructed upon these estimates, it will corrupt the classical test. We have that $\beta_0 = 0$, so that we obtain $r(\hat{\theta}, \beta_0) = \hat{\delta} - \hat{\pi}\beta_0 = \hat{\delta}$. Therefore, only the biased estimator $\hat{\delta}$ will have an effect on $r(\hat{\theta}, \beta_0)$. If we assume, for simplicity, that we only have one instrumental variable, then we would have $\hat{\delta} = \delta + b$, where b denotes a scalar bias term. Therefore, we have $r(\hat{\theta}, \beta_0) = 0$ when $\delta + b = \pi\beta + b = 0$, which holds when $\beta = -b/\pi$. In our case, the outlier introduces a positive bias so that $r(\hat{\theta}, \beta_0) = 0$ at a negative β value, which is why the power curve of the classical test shifts to the left. For the robust tests, the bias is very close to zero, so that the we obtain $r(\hat{\theta}, \beta_0) = 0$ when $\beta \approx 0$. In case of the probit model, we do not have a large

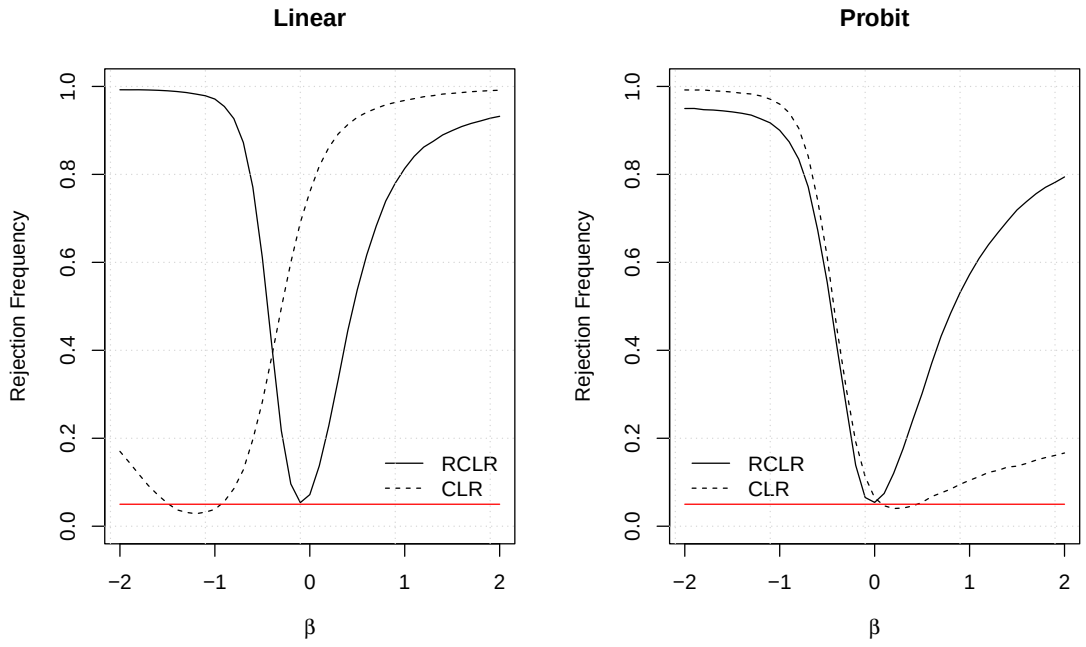

Figure 2: Power curves of the robust test and its classical version that tests $H_0\colon \beta = 0$ for various values of β in the linear instrumental variable model and the endogenous probit model. Environment with contamination by an outlier.

outlier in y so that the estimator $\hat{\delta}$ is not biased. Moreover, as $\beta_0 = 0$, the biased estimator $\hat{\pi}$ is not used to compute $r(\hat{\theta}, \beta_0)$. Therefore, we do not see a large shift of the power curve in the probit model. However, we do see that for positive values of β, the classical test is much less powerful than the robust test. When $\beta_0 \neq 0$, then we would also see a large overrejection at the null in the probit model as we show in the next section.

In Figure 3, we present the power curves of the robust test and the classical test in the two different models when there is "distributional" contamination in the error terms. In this case, we see that the robust tests are more powerful than the classical tests in both models. This happens, because the errors that are drawn from the $t(3)$ distribution sometimes generate large outliers. The robust tests can effectively downweight these outliers, which results in more power compared to the classical test. Note, the observations from the $t(3)$ distribution do not bias the classical estimators $\hat{\delta}$ and $\hat{\pi}$ as the $t(3)$ distribution is symmetric. Therefore, we do not see shifts in the power curves.

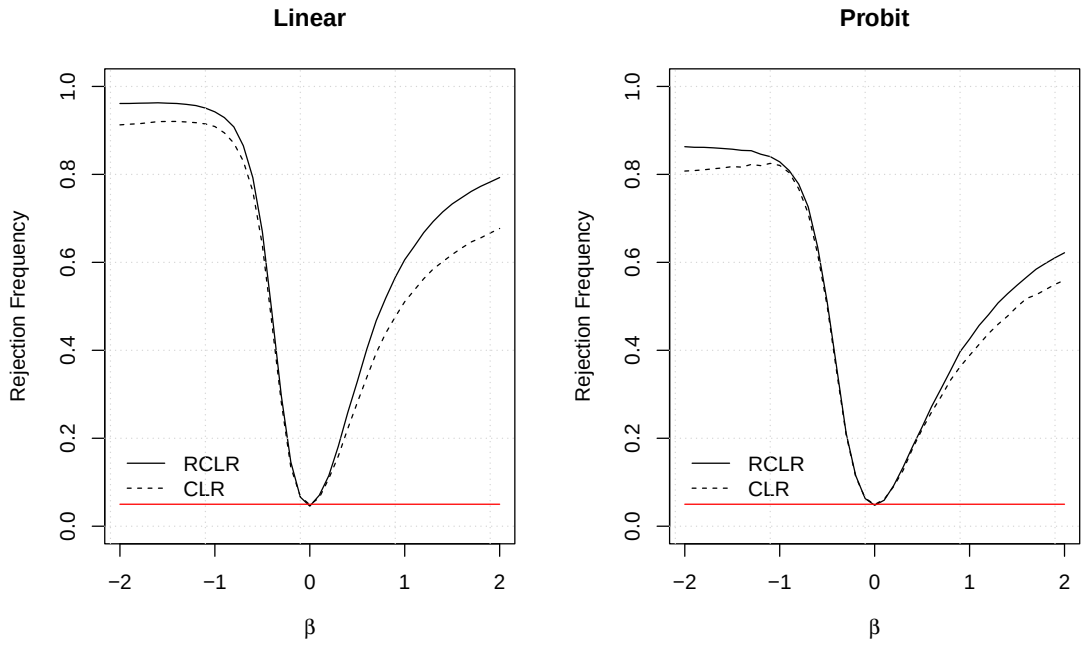

Figure 3: Power curves of the robust test and its classical version that tests $H_0\colon \beta = 0$ for various values of β in the linear instrumental variable model and the endogenous probit model. Environment with contamination in the error terms.

6.2 Sensitivity of the level

In Tables 1, 2 and 3 we present additional simulation evidence that shows good performance of the empirical size of the robust S, K and CLR tests when there is contamination by an outlier. We analyze the behavior of the level of the tests in the linear instrumental variable model and the endogenous probit model. The baseline and outlier-contaminated environments are exactly the same as before. The only difference is that we now consider two different values for F^*. We consider $F^* = 5$ and $F^* = 20$ to mimic a setting with weak and and strong instruments, respectively.

In Table 1, we present the results in the linear IV model. In this case, we see that both the robust and nonrobust tests remain size correct when there is no outlier in the data. However, when there is an outlier in the data we see that the nonrobust tests are not size correct anymore and largely overreject the null hypothesis. In contrast, the robust tests remain more reliable. The robust tests do overreject more compared to baseline environment. However, the

amount of overrejection is not very problematic and shows that the level remains stable in a neighborhood of the model.

Table 1: Size comparison (in percentage) $H_0\colon \beta = 0$, linear instrumental variable model with and without outlier.

	$F^* = 5$						$F^* = 20$					
	No outlier			Outlier			No outlier			Outlier		
Nominal level	10	5	1	10	5	1	10	5	1	10	5	1
S	10.22	5.18	1.18	98.89	95.98	76.68	10.22	5.18	1.18	98.89	95.98	76.68
K	10.13	4.90	1.20	70.77	59.55	35.36	10.13	4.98	1.12	81.16	68.00	34.58
CLR	10.01	4.91	1.23	88.10	80.69	56.65	10.10	4.92	1.14	85.45	74.00	42.99
RS	10.00	5.30	1.14	13.74	7.29	1.93	10.00	5.30	1.14	13.87	7.36	1.95
RK	9.95	5.02	1.13	12.24	6.74	1.81	10.04	5.00	1.16	12.46	6.92	1.73
RCLR	9.90	5.01	1.20	12.70	7.02	1.98	9.99	4.93	1.18	12.55	6.97	1.78

In Table 2, we present the results of the endogenous probit model. In this case, we see that the empirical size of the robust test remains reliable in all settings. We see that the nonrobust test is size correct in the setting without the outlier. In the setting with the outlier, the nonrobust only slightly overrejects in the probit model. This happens because the outlier that we introduced for the probit, in contrast to the linear IV model, does not bias the estimate of δ. The estimate of π is biased. However, as we are testing whether $\beta = 0$ so that $\beta_0 = 0$ we have that $r(\hat{\theta}, \beta_0) = \hat{\delta} - \hat{\pi}\beta_0 = \hat{\delta}$ so that even if the estimate of δ is biased it does not affect the size of the tests much as long as the estimate of δ is correct. This is, however, coincidental and if we would test $H_0\colon \beta = \beta_0$ with $\beta_0 \neq 0$ we would see a larger rejection rates for the nonrobust tests. In Table 3, we present the results of a simulation study where we test the null hypothesis $H_0\colon \beta = 0.2$ for the probit model. In this case, we see that the outlier has a large effect on the empirical size of the nonrobust tests causing larger overrejections.

7 Empirical examples

In this section, we show how the robust tests can be used in practice by revisiting three empirical studies. First, we consider the data and several specifications in Alesina and Zhuravskaya (2011) who examine the effect of segregation on the quality of government. Second, we revisit the main specifications considered in Ananat (2011) where the effect of (racial) segregation

Table 2: Size comparison (in percentage) H_0: $\beta = 0$, endogenous probit model with and without outlier.

	$F^* = 5$						$F^* = 20$					
	No outlier			Outlier			No outlier			Outlier		
Nominal level	10	5	1	10	5	1	10	5	1	10	5	1
S	10.48	5.00	0.94	12.53	6.58	1.39	10.48	5.00	0.94	12.47	6.49	1.36
K	10.31	5.55	1.10	12.92	6.85	1.69	10.39	5.35	1.08	12.54	6.57	1.59
CLR	10.29	5.41	1.06	12.71	6.87	1.75	10.14	5.45	1.11	12.47	6.65	1.56
RS	9.82	4.83	0.90	10.20	5.12	0.89	9.82	4.83	0.90	10.25	5.15	0.89
RK	10.21	5.12	0.95	10.44	5.31	0.99	10.10	4.97	0.99	10.48	5.27	1.01
RCLR	10.21	4.97	0.95	10.54	5.36	0.92	10.06	5.02	0.97	10.52	5.25	1.02

Table 3: Size comparison (in percentage) H_0: $\beta = 0.2$, endogenous probit model with and without outlier.

	$F^* = 5$						$F^* = 20$					
	No outlier			Outlier			No outlier			Outlier		
Nominal level	10	5	1	10	5	1	10	5	1	10	5	1
S	9.89	4.82	0.83	22.66	12.70	3.61	9.91	4.74	0.84	22.81	13.07	3.29
K	10.25	5.00	0.84	23.48	14.28	3.98	10.21	5.21	0.85	21.64	13.22	3.76
CLR	10.26	4.95	0.79	23.78	14.47	3.97	10.27	5.22	0.87	21.68	13.25	3.80
RS	9.40	4.39	0.64	9.75	4.52	0.67	9.32	4.46	0.74	9.64	4.74	0.76
RK	9.93	5.01	0.80	10.39	5.26	0.85	10.06	4.71	0.74	10.66	5.26	0.87
RCLR	9.92	4.94	0.76	10.38	5.08	0.80	10.03	4.74	0.75	10.57	5.37	0.89

on urban poverty and inequality is studied. Finally, we revisit the Staiger and Stock (1997) specifications for the Angrist and Krueger (1991) data where the effect of education on labor market earnings is studied.

7.1 Alesina and Zhuravskaya (2011)

Alesina and Zhuravskaya (2011) study the effect of segregation on the quality of government in a cross section of countries using a linear instrumental variable model. They find that ethnically and linguistically segregated countries have a lower quality of government. Furthermore, they find that there is no relationship between religious segregation and governance. To address endogeneity concerns caused by mobility and endogeneous internal borders, an instrument is constructed for segregation. For more information, data and the construction of the instrument we refer to Alesina and Zhuravskaya (2011).

In Section 5D Alesina and Zhuravskaya (2011) mention that they carefully examined

Table 4: Results using data from Alesina and Zhuravskaya (2011). Specifications correspond to the specifications in Panel D of Table 7 in Alesina and Zhuravskaya (2011). Confidence sets are given for the parameter belonging to endogeneous regressor (x) "Segregation" for six different specifications (y). The RAR confidence sets are calculated based on Mallows type estimator based on the Huber function and "hat" matrix weights as in Section 5.2.

Specification	Language					
	I	II	III	IV	V	VI
	Voice	Political stability	Government effectiveness	Regulatory quality	Rule of law	Control of corruption
95% AR confidence set	$[-6.04, -0.93]$	$[-5.73, -1.01]$	$[-3.70, 0.79]$	$[-4.86, 2.20]$	$[-3.62, 0.52]$	$[-3.48, 1.81]$
95% RAR confidence set	$[-7.40, -0.04]$	$[-6.10, 2.48]$	$[-3.41, 1.38]$	$[-2.32, 4.37]$	$[-3.32, 3.28]$	$[-3.84, 1.90]$
All control variables	Yes	Yes	Yes	Yes	Yes	Yes
No. of observations	92	92	92	92	92	92
First-stage F	17.22	17.22	17.22	17.22	17.22	17.22

whether a handful of influential observations drive their results. By exluding influential observations and recalculating their statistics they conclude that this is not the case. However, Alesina and Zhuravskaya (2011) mention that in the specifications of Panel D in Table 7 removing two influential observations leads to the first-stage F-statistic dropping from 17.22 to 7.82 making inference based on 2SLS estimator unreliable (Staiger and Stock, 1997). Alesina and Zhuravskaya (2011) solve this by also removing the most influential observation in the first stage from the data so that the instrument becomes "strong enough". Note, manually removing outliers from the data and then relying on classical statistical methods is not recommended for several reasons such as masking effects and underestimated variability and it is advisable to rely on a robust method from the start (Maronna et al., 2019, Section 4.3). It seems that the 2SLS estimator might be unreliable due to the outliers and the weak instrument. Therefore, to re-evaluate the robustness of the results, we apply our test to six specifications ("Voice", "Political stability", "Government effectiveness", "Regulatory Quality", "Rule of law" and "Control of corruption") of Panel D in Table 7 of Alesina and Zhuravskaya (2011). As there is only one instrument, we calculate the 95% confidence sets of the robust AR statistic and the (classical) AR statistic. The results are reported in Table 4.

From Table 4 we can see that in specifications II, III, IV and V that the confidence set of the robust AR confidence set is shifted compared to the AR confidence set. This suggests that outliers did have an effect in these regressions as we would expect the confidence set of the robust AR test to be wider than the confidence set of the AR test when there are no outliers, but not shifted. The shift of the confidence set suggests that the LS estimators the AR test is constructed upon are biased due to the outlier(s). Therefore, the confidence sets based on the robust AR are more reliable. Overall, the outliers do not seem to be very problematic as in all specifications, except specification II, the final decision whether to reject or not reject the null hypothesis $H_0 \colon \beta = 0$ remains the same. When we analyze specification II, we see that the robust confidence set would not reject the null hypothesis, while the classical confidence set would reject the null hypothesis. In this case, it does seem the outliers are the main drivers of the significant result.

7.2 Ananat (2011)

Ananat (2011) studies the effect of racial segregation on urban poverty and inequality using a linear instrumental variable model. To overcome endogeneity issues, a railroad division index is used to instrument for racial segregation. Using this instrumental variable, Ananat (2011) shows that segregation increases metropolitan rates of black poverty and overall black-white income disparities, while decreasing rates of white poverty and inequality within the white population. For more information, data and the construction of the instrument we refer to Ananat (2011).

Klooster and Zhelonkin (2023) show that an outlier in the control variable used in the main results of Ananat (2011) inflates the first-stage F-statistic from 1.83 to 19.32. As the outlier was not taken into account in the original study, it was assumed that the instrument was strong. Consequently, estimation was done with a 2SLS estimator and inference with a t-test. Due to the outlier (and the weak instrument) inference based on the 2SLS estimator might be unreliable. Therefore, to re-evaluate the robustness of the results, we apply our test to the four

Table 5: Results using data from Ananat (2011). Specifications correspond to the specifications in columns (3) and (4) in Table 2 in Ananat (2011). Confidence sets are given for the parameter belonging to endogeneous regressor (x) "Segregation" for four different specifications (y). The RAR confidence sets are calculated based on Mallows type estimator based on the Huber function and "hat" matrix weights as in Section 5.2.

Specification	I	II	III	IV
	Gini index whites	Gini index blacks	Poverty rate whites	Poverty rate blacks
95% AR confidence set	$(-0.64, -0.18)$	$(0.22, 2.15)$	$(-0.38, -0.09)$	$(0.00, 0.48)$
95% RAR confidence set	$(-\infty, -0.12)$ $\cup(1.62, \infty)$	$(-\infty, -3.79)$ $\cup(0.19, \infty)$	$(-\infty, -0.08)$ $\cup(0.90, \infty)$	$(-\infty, \infty)$
First-stage F	19.32	19.32	19.32	19.32
No. of observations	121	121	121	121

main specifications ("Gini index whites", "Gini index blacks", "Poverty rate whites", "Poverty rate blacks"), which can be found in columns (3) and (4) of Table 2 in Ananat (2011). As there is only one instrument, we calculate the 95% confidence sets of the robust AR statistic and the (classical) AR statistic. The results are reported in Table 5.

When we analyze Table 5, we find large differences between the robust and classical confidence sets. When there are no outliers in the data, we would expect that the robust confidence sets are only a bit wider than the classical confidence sets. However, in this case, the classical confidence sets are bounded convex sets, while the robust confidence sets are unbounded sets. The shape of the robust confidence sets do correctly suggest that the instrument is weak. In this example, the outlier does have a large effect and we recommend using the robust confidence sets for reliable inference.

7.3 Angrist and Krueger (1991)

Angrist and Krueger (1991) study the effect of education on labor market earnings using a linear instrumental variable model. To adress endogeneity issues, quarter of birth instruments are constructed for education. Using these instrumental variables they find a positive relationship

between years of education and labor market earnings. We revisit four specifications presented in Table 2 of Staiger and Stock (1997) based on the 1930 - 1939 cohort. For more information, data and the construction of the instrument, we refer to Staiger and Stock (1997) and Angrist and Krueger (1991).

Bound et al. (1995) showed that the relationship between the instruments and the endogeneous regressor is quite weak in certain specifications of Angrist and Krueger (1991). Furthermore, more recently, Sølvsten (2020) shows that the LIML residuals of the structural equation of a certain specification in Angrist and Krueger (1991) are distributed roughly like a normal distribution at the center with outlying errors that closely follow a $t(3)$-distribution (reminiscent of the "distributional" contamination scenario in Section 6). Due to the weak instruments (and possible outliers), inference based on the 2SLS estimator used by Angrist and Krueger (1991) might be unreliable. Moreover, due to outliers, weak instrument robust tests might be corrupted and/or inefficient. Therefore, to re-evaluate the robustness of the results we replicate the results reported in Panel A of Table 2 in Staiger and Stock (1997). For each specification, we give the 95% confidence set of the CLR and robust CLR, and the first-stage F-statistic. The results are reported in Table 6.

When we compare the 95% confidence sets of the CLR and RCLR statistics, we note that the RCLR confidence sets are smaller than the CLR confidence sets in every specification. This happens because the RCLR statistic effectively downweights outlying values in the residuals. Similar as in the "distributional" contamination scenario in Section 6 this results in better variance estimates and hence tighter confidence sets.

8 Conclusion

In this article, we proposed a general framework to construct weak instrument robust testing procedures that are also robust to outliers in a general class of limited dependent instrumental variable models. The framework is constructed upon M-estimators and we showed that the

Table 6: Results for Angrist and Krueger (1991) data. Specifications as in Table 2 of Staiger and Stock (1997), except for specification III (see the note below). Confidence sets are given for the parameter belonging to the endogeneous regressor (x) "Years of schooling" for four different specifications that all use the same dependent variable (y) "log weekly wages". The RCLR confidence sets are calculated based on Mallows type estimator based on the Huber function and "hat" matrix weights as in Section 5.2. QOB, YOB and SOB stand for quarter, year and state of birth.

Specification	I	II	III*	IV
95% CLR confidence set	$[0.042, 0.137]$	$[0.026, 0.116]$	$[-0.064, 0.279]$	$[-0.068, 0.266]$
95% RCLR confidence set	$[0.047, 0.122]$	$[0.032, 0.101]$	$[-0.038, 0.190]$	$[-0.048, 0.180]$
First-stage F	30.53	4.74	2.43	1.87
controls (w)				
Base controls	Yes	Yes	Yes	Yes
SOB	No	No	Yes	Yes
Age, Age^2	No	No	No	Yes
Instruments (z)				
QOB	Yes	Yes	Yes	Yes
QOB*YOB	No	Yes	Yes	Yes
QOB*SOB	No	No	Yes	Yes
No. of instruments	3	30	180	178
Observations	329,509	329,509	329,509	329,509

*This specification slightly different from specification III of Table 2 in Staiger and Stock (1997). Instead of using Age and Age^2 as control variables, we use the SOB controls instead. This was done as we encountered some small numerical difficulties when replicating the original specification leading to unusual confidence sets for both the CLR and RCLR tests.

classical weak instrument robust tests, such as the AR, K and CLR tests, can be obtained by specifying the M-estimators to be the LS estimators. We formally showed that influence function of the minimum distance test statistics are only bounded when the influence function of the estimators they are constructed upon is bounded. As all classical minimum distance robust tests are constructed upon estimators that do not have a bounded influence function, we showed how to construct robust alternatives. In particular, we showed how to construct minimum distance robust tests based on a Mallows type M-estimator that allows reliable inference in a wide variety of models, including the linear IV model and the endogenous probit model. By means of a simulation study, we documented good performance of our robust tests in different contaminated environments. Finally, we illustrated how the robust tests can be used in practice by revisiting three different empirical studies.

Appendix A

Proof of Proposition 1

Proof. Under Assumption 1, we have

$$\sqrt{n}\begin{pmatrix}\hat{\delta}\\ \hat{\pi}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left\{\begin{pmatrix}\delta\\ \pi\end{pmatrix}, \begin{pmatrix}\Omega_{\delta\delta} & \Omega_{\delta\pi}\\ \Omega_{\pi\delta} & \Omega_{\pi\pi}\end{pmatrix}\right\}.$$

We can use the continuous mapping theorem to show that, under the null hypothesis, we have

$$r(\hat{\theta}, \beta_0) \xrightarrow{p} r(\theta, \beta_0) = \delta - \pi\beta_0 = \pi\beta - \pi\beta_0 = 0.$$

We can write

$$r(\hat{\theta}, \beta_0) = \begin{pmatrix}I_k & -\beta_0 I_k\end{pmatrix}\begin{pmatrix}\hat{\delta}\\ \hat{\pi}\end{pmatrix}.$$

We obtain

$$\begin{aligned}\sqrt{n}r(\hat{\theta}, \beta_0) &\xrightarrow{d} \mathcal{N}\left\{0, \begin{pmatrix}I_k & -\beta_0 I_k\end{pmatrix}\begin{pmatrix}\Omega_{\delta\delta} & \Omega_{\delta\pi}\\ \Omega_{\pi\delta} & \Omega_{\pi\pi}\end{pmatrix}\begin{pmatrix}I_k\\ -\beta_0 I_k\end{pmatrix}\right\}\\ &= \mathcal{N}\left\{0, \Omega(\beta_0)\right\}\end{aligned}$$

where $\Omega(\beta_0) = \Omega_{\delta\delta} - \beta_0(\Omega_{\delta\pi} + \Omega_{\pi\delta}) + \beta_0^2\Omega_{\pi\pi}$.

We conclude that under the null hypothesis and Assumption 1, that

$$S(\hat{\theta}, \beta_0) = nr(\hat{\theta}, \beta_0)^\top \Omega(\beta_0)^{-1} r(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(k).$$

□

Proof of Lemma 1

Proof. Under Assumption 1, we have

$$\sqrt{n}\begin{pmatrix}\hat{\delta}\\ \hat{\pi}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left\{\begin{pmatrix}\delta\\ \pi\end{pmatrix}, \begin{pmatrix}\Omega_{\delta\delta} & \Omega_{\delta\pi}\\ \Omega_{\pi\delta} & \Omega_{\pi\pi}\end{pmatrix}\right\}.$$

We can use the continuous mapping theorem to show that, under the null hypothesis, we have

$$D(\hat{\theta}, \beta_0) \xrightarrow{p} D(\theta, \beta_0) = \pi.$$

We can write

$$\begin{Bmatrix} I_k & -\beta_0 I_k \\ -\left(\Omega_{\pi\delta} - \Omega_{\pi\pi}\beta_0\right)\Omega(\beta_0)^{-1} & \left(\Omega_{\pi\delta} - \Omega_{\pi\pi}\beta_0\right)\Omega(\beta_0)^{-1}\beta_0 + I_k \end{Bmatrix} \begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} = \begin{Bmatrix} r(\hat{\theta}, \beta_0) \\ D(\hat{\theta}, \beta_0) \end{Bmatrix}.$$

After some algebra, we can conclude that

$$\sqrt{n} \begin{Bmatrix} r(\hat{\theta}, \beta_0) \\ D(\hat{\theta}, \beta_0) \end{Bmatrix} \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{Bmatrix} \Omega(\beta_0) & 0 \\ 0 & \Lambda(\beta_0) \end{Bmatrix} \right],$$

with $\Lambda(\beta_0) = \Omega_{\pi\pi} - (\Omega_{\pi\delta} - \beta_0\Omega_{\pi\pi})\Omega(\beta_0)^{-1}(\Omega_{\delta\pi} - \beta_0\Omega_{\pi\pi})$. □

Proof of Proposition 2

Proof. From Lemma 1 it directly follows that under the null hypothesis $\pi = 0$ and $\beta = \beta_0$, we have $W(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(k)$.

To derive the asymptotic distribution of $K(\hat{\theta}, \beta_0)$, we follow similar arguments as in Kleibergen (2005). From Lemma 1, we have

$$\sqrt{n} r(\hat{\theta}, \beta_0) \xrightarrow{d} N \sim \mathcal{N}\{0, \Omega(\beta_0)\}$$

We denote the asymptotic distribution of $D(\hat{\theta}, \beta_0)$ by D, i.e., $\sqrt{n} D(\hat{\theta}, \beta_0) \xrightarrow{d} D \sim \mathcal{N}\{\pi, \Lambda(\beta_0)\}$. Then $\sqrt{n} D(\hat{\theta}, \beta_0)^\top \sqrt{n} r(\hat{\theta}, \beta_0) \xrightarrow{d} D^\top N$. The conditional distribution of $D^\top N$ given D reads

$$D^\top N \big| D \sim \mathcal{N}\{0, D^\top \Omega(\beta_0) D\}.$$

From Lemma 1, we know that D is independent of N as they are jointly normally distributed and uncorrelated. Therefore, we obtain an unconditional result by normalizing the expression by $\{D^\top \Omega(\beta_0) D\}^{-1/2}$,

$$\begin{aligned} &\{D(\hat{\theta}, \beta_0)^\top \Omega(\beta_0) D(\hat{\theta}, \beta_0)\}^{-1/2} D(\hat{\theta}, \beta_0)^\top \sqrt{n} r(\hat{\theta}, \beta_0) \\ = &\{\sqrt{n} D(\hat{\theta}, \beta_0)^\top \Omega(\beta_0) \sqrt{n} D(\hat{\theta}, \beta_0)\}^{-1/2} \sqrt{n} D(\hat{\theta}, \beta_0)^\top \sqrt{n} r(\hat{\theta}, \beta_0) \\ &\xrightarrow{d} \{D^\top \Omega(\beta_0) D\}^{-1/2} D^\top N \sim \mathcal{N}(0, 1). \end{aligned}$$

Therefore, $K(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(1)$.

At last, we derive the asymptotic distribution of the $CLR(\hat{\theta}, \beta_0)$ statistic, conditional on $D(\hat{\theta}, \beta_0) = D$. We define $\hat{U} = \Omega(\beta_0)^{-1/2} r(\hat{\theta}, \beta_0)$ and $\hat{R} = \Omega(\beta_0)^{-1/2} D(\hat{\theta}, \beta_0)$. By Lemma 1, the asymptotic distribution of $\sqrt{n}\begin{pmatrix}\hat{U} & \hat{R}\end{pmatrix}^\top$ is jointly normal with zero covariance. Let R denote the asymptotic distribution of $\sqrt{n}\hat{R}$ and U the asymptotic distribution of $\sqrt{n}\hat{U}$, then we know that U and R are independent. Moreover, we have $S(\hat{\theta}, \beta_0) = n\hat{U}^\top \hat{U}$ and $K(\beta_0) = n\hat{U}^\top P_{\hat{R}} \hat{U}$, where $P_{\hat{R}} = \hat{R}(\hat{R}^\top \hat{R})^{-1}\hat{R}$. We can write $S(\hat{\theta}, \beta_0) = K(\hat{\theta}, \beta_0) + J(\hat{\theta}, \beta_0)$, with

$$J(\hat{\theta}, \beta_0) = n\hat{U}^\top (I_k - P_{\hat{R}})\hat{U}$$

It holds that $(I_k - P_{\hat{R}})^{-1/2}\sqrt{n}\hat{U} \xrightarrow{d} (I_k - P_R)^{-1/2} U$, with $U \sim \mathcal{N}(0, I_k)$. The conditional distribution of $(I_k - P_R)^{-1/2}U$ given $R = \Omega(\beta_0)^{1/2} D$ reads

$$(I_k - P_R)^{-1/2} U \big| R \sim \mathcal{N}(0, I_k - P_R).$$

We know that R is independent of N by Lemma 1. Therefore, the result also holds unconditionally. Furthermore, as the rank of $I_k - P_R$ is $tr(I_k - P_R) = k - 1$, we have

$$J(\hat{\theta}, \beta_0) \xrightarrow{d} \chi^2(k-1).$$

The asymptotic distribution of $K(\hat{\theta}, \beta_0)$ and $J(\hat{\theta}, \beta_0)$ are independent, as $J(\hat{\theta}, \beta_0)$ projects on the orthogonal complement of $\hat{R}$. Now the result follows. □

Proof of Proposition 3

Proof. We denote the functional form of the minimum distance statistics as $S(F), K(F), W(F)$ and $CLR(F)$, where for simplicity we suppressed the dependency on β_0 in all the test statistics.

Let G be an arbitrary distribution function and define $W = D^\top \Lambda(\beta_0)^{-1} D$. The functional form of the CLR statistic, conditional on D is

$$CLR(G) = \frac{1}{2}\left[S(G) - W + \sqrt{\{S(G) - W\}^2 + 4W \cdot K(G)}\right].$$

To simplify the notation, we write

$$A(G) = \{S(G) - W\}^2 + 4W \cdot K(G),$$

so that

$$CLR(G) = \frac{1}{2}\left\{S(G) - W + \sqrt{A(G)}\right\}.$$

We start with the case $D \neq 0$. In this case it must hold that $W > 0$ and as $A(F_\theta) = W^2$, we have $CLR(F_\theta) = 0$. To calculate the first derivative, note that

$$\frac{\partial}{\partial t}CLR(F_t)\Big|_{t=0} = \frac{1}{2}\left\{\frac{\partial}{\partial t}S(F_t)\Big|_{t=0} + \frac{1}{2\sqrt{A(F_\theta)}} \cdot \frac{\partial}{\partial t}A(F_t)\Big|_{t=0}\right\}.$$

We know that $\frac{\partial}{\partial t}S(F_t)\big|_{t=0} = 0$, and

$$\begin{aligned}\frac{\partial}{\partial t}A(F_t)\Big|_{t=0} &= 2\left\{S(F_\theta) - W\right\}\frac{\partial}{\partial t}S(F_t)\Big|_{t=0} + 4W\frac{\partial}{\partial t}K(F_t)\Big|_{t=0} \\ &= 0,\end{aligned}$$

as $\frac{\partial}{\partial t}S(F_t)\big|_{t=0} = 0$, and $\frac{\partial}{\partial t}K(F_t)\big|_{t=0} = 0$. Furthermore, as $W > 0$, we have

$$A(F_\theta) = \left\{S(F_\theta) - W\right\}^2 + 4WK(F_\theta) = (0 - W)^2 + 0 = W^2.$$

Hence, $A(F_\theta) > 0$ so that we are not dividing by zero. It thus follows that

$$\frac{\partial}{\partial t}CLR(F_t)\Big|_{t=0} = 0.$$

Next, we calculate the second derivative. We have

$$\frac{\partial^2}{\partial t^2}CLR(F_t)\Big|_{t=0} = \frac{1}{2}\left\{\frac{\partial^2}{\partial t^2}S(F_t)\Big|_{t=0} + \frac{1}{2W}\frac{\partial^2}{\partial t^2}A(F_t)\Big|_{t=0}\right\},$$

where we used the results that $\frac{\partial}{\partial t}A(F_t)\big|_{t=0} = 0$ and $A(F_\theta) = W^2$. We continue and obtain

$$\begin{aligned}\frac{\partial^2}{\partial t^2}A(F_t)\Big|_{t=0} &= 2\left\{\frac{\partial}{\partial t}S(F_t)\Big|_{t=0}\right\}^2 - 2W\frac{\partial^2}{\partial t^2}S(F_t)\Big|_{t=0} + 4W\frac{\partial^2}{\partial t^2}K(F_t)\Big|_{t=0} \\ &= -2W\frac{\partial^2}{\partial t^2}S(F_t)\Big|_{t=0} + 4W\frac{\partial^2}{\partial t^2}K(F_t)\Big|_{t=0}.\end{aligned}$$

Substituting this back into the previous equation, it follows that

$$\begin{aligned}\frac{\partial^2}{\partial t^2}CLR(F_t)\Big|_{t=0} &= \frac{1}{2}\left\{\frac{\partial^2}{\partial t^2}S(F_t)\Big|_{t=0} + \frac{4W\frac{\partial^2}{\partial t^2}K(F_t)\Big|_{t=0} - 2W\frac{\partial^2}{\partial t^2}S(F_t)\Big|_{t=0}}{2W}\right\} \\ &= \frac{\partial^2}{\partial t^2}K(F_t)\Big|_{t=0}.\end{aligned}$$

Therefore, using L'Hôpital's rule twice, the influence function of the CLR statistic, given $\tilde{W} > 0$, is

$$\begin{aligned}
\mathrm{IF}\{d; \sqrt{CLR}, F\} &= \lim_{t\to 0} \left\{ \sqrt{CLR(F_t)} - \sqrt{CLR(F)} \right\} / t \\
&= \left\{ \lim_{t\to 0} CLR(F_t)/t^2 \right\}^{1/2} \\
&= \left\{ \frac{1}{2} \frac{\partial^2}{\partial t^2} CLR(F_t)\Big|_{t=0} \right\}^{1/2} \\
&= \left\{ \frac{1}{2} \frac{\partial^2}{\partial t^2} K(F_t)\Big|_{t=0} \right\}^{1/2} \\
&= \mathrm{IF}(d; \sqrt{K}, F).
\end{aligned}$$

Next, we assume $D = 0$. In this case $W = 0$, so that

$$CLR(G) = S(G).$$

Hence,

$$\mathrm{IF}(d; \sqrt{CLR}, F) = \mathrm{IF}(d; \sqrt{S}, F),$$

and the result follows.

We continue with deriving the influence function of the S and K statistics, conditional on $D(F_n, \beta_0) = D$. We have $\frac{\partial}{\partial t} S(F_t)\big|_{t=0} = 2\,\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\}^\top \Omega(\beta_0)^{-1} r(F_\theta, \beta_0) = 0$, as $r(F_\theta, \beta_0) = 0$ due to Fisher consistency. The second derivative gives

$$\frac{\partial^2}{\partial t^2} S(F_t)\Big|_{t=0} = 2\,\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\}^\top \Omega(\beta_0)^{-1}\,\mathrm{IF}\{d; r(\cdot, \beta_0\}, F_\theta).$$

Using L'Hôpital's rule twice, we obtain

$$\begin{aligned}
\mathrm{IF}(d; \sqrt{S}, F_\theta) &= \lim_{t\to 0} \left\{ \sqrt{S(F_t)} - \sqrt{S(F)} \right\} / t \\
&= \left\{ \lim_{t\to 0} S(F_t)/t^2 \right\}^{1/2} \\
&= \left\{ \frac{1}{2} \frac{\partial^2}{\partial t^2} S(F_t)\Big|_{t=0} \right\}^{1/2} \\
&= \sqrt{\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\}^\top \Omega(F_\theta, \beta_0)^{-1}\,\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\}}.
\end{aligned}$$

□

The K statistic, conditional on $D(F_n, \beta_0) = D$, follows exactly the same arguments so we omit this derivation. We obtain

$$\mathrm{IF}(d; \sqrt{K}, F) = \sqrt{\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\}^\top D \left\{D^\top \Omega(\beta_0) D\right\}^{-1} D^\top \mathrm{IF}\{d; g(\cdot, \beta_0), F_\theta\}}.$$

At last, we have

$$\mathrm{IF}\{d; r(\cdot, \beta_0), F_\theta\} = \mathrm{IF}\{d; \delta(\cdot), F_\theta\} - \beta_0 \, \mathrm{IF}\{d; \pi(\cdot), F_\theta\}.$$

Appendix B

Details of the practical implementation

For the implementation, we follow the algorithm presented in Section 4 of the Appendix in Magnusson (2010). Specifically, we use a control function approach. In this case, we consider the model

$$\begin{cases} y^* &= \beta x + \alpha v + \epsilon \\ x &= z^\top \pi + v \end{cases} \qquad \begin{cases} y^* &= z^\top \delta + \delta_v v + \epsilon \\ x &= z^\top \pi + v, \end{cases} \tag{26}$$

with $\delta_v = \alpha + \beta$ and $\epsilon = u - v\alpha$. Magnusson (2010) shows that in this case, it holds that $\Omega_{\pi\delta} = \Omega_{\pi\pi}(\delta_v \otimes I_k)$. This is beneficial, as built-in software package typically are able to provide an estimate of the matrix $\Omega_{\pi\pi}$, but not of $\Omega_{\pi\delta}$.

For the (robust) tests that we consider in the simulation study in Section 6, we use the following steps:

1. Estimate π and $\Omega_{\pi\pi}$ using a robust M-estimator. We use a Mallows type M-estimator with "hat" matrix weights and using the Huber downweighting function. We use the `rlm()` function from the R package MASS (Venables and Ripley, 2002) with the default settings to obtain the estimates $\hat{\pi}$ and $\hat{\Omega}_{\pi\pi}$. Moreover, we keep the residuals from the robust regression and denote them by $\hat{v}$.

2. Estimate δ, δ_v and $\Omega_{\delta\delta}$ from the following equation

$$y = f(z^\top \delta + \delta_v \hat{v} + \tilde{\epsilon}),$$

where $\tilde{\epsilon} = \epsilon - (\hat{v} - v)\delta_v$ and f is the known function. For the estimation, we use a Mallows type M-estimator with "hat" matrix weights and we use the Huber downweighting function. Note, to compute the hat matrix, we also include the residuals $\hat{v}$.

- In case of the linear IV model, we use the function `rlm()` from the R package `MASS` (Venables and Ripley, 2002) with the default settings to obtain the estimates $\hat{\delta}$, $\hat{\delta}_v$ and $\hat{\Omega}_{\delta\delta}$.
- In case of the probit, logit and Poisson model, we use the function `glmrob()` from the R package `robustbase` (Maechler et al., 2023).

3. We estimate $\hat{\Omega}_{\pi\delta} = \hat{\delta}_v \hat{\Omega}_{\pi\pi}$. Using the estimates $\hat{\delta}, \hat{\pi}, \hat{\Omega}_{\pi\pi}, \hat{\Omega}_{\delta\delta}$ and $\hat{\Omega}_{\pi\delta}$ we follow Magnusson (2010) and construct

$$\begin{aligned}
\hat{\Omega}(\beta_0) &= \hat{\Omega}_{\delta\delta} + (\hat{\delta}_v - \beta_0)^2 \hat{\Omega}_{\pi\pi}, \\
D(\hat{\theta}, \beta_0) &= \hat{\pi} - (\hat{\delta}_v - \beta_0)^2 \{\hat{\Omega}(\beta_0)\}^{-1} \hat{\Omega}_{\pi\pi}, \\
\hat{\Lambda}(\beta_0) &= \hat{\Omega}_{\pi\pi} - (\hat{\delta}_v - \beta_0)^2 \hat{\Omega}_{\pi\pi} \{\hat{\Omega}(\beta_0)\}^{-1} \hat{\Omega}_{\pi\pi},
\end{aligned}$$

which allows us to construct all the (robust) test statistics.

References

Alesina, A. and Zhuravskaya, E. (2011), "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 101, 1872–1911.

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, 34–66.

Anderson, T. W. and Rubin, H. (1949), "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Statistics*, 20, 46–63.

Andrews, D. W. and Marmer, V. (2008), "Exactly Distribution-Free Inference in Instrumental Variables Regression With Possibly Weak Instruments," *Journal of Econometrics*, 142, 183–200.

Andrews, D. W., Moreira, M. J., and Stock, J. H. (2006), "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–752.

Andrews, D. W. and Soares, G. (2007), "Rank Tests for Instrumental Variables Regression With Weak Instruments," *Econometric Theory*, 23, 1033–1082.

Andrews, I., Gentzkow, M., and Shapiro, J. M. (2017), "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132, 1553–1592.

— (2020), "On the Informativeness of Descriptive Statistics for Structural Estimates," *Econometrica*, 88, 2231–2258.

Andrews, I., Stock, J. H., and Sun, L. (2019), "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. D., Imbens, G. W., and Robun, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. D. and Krueger, A. B. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

Bonhomme, S. and Weidner, M. (2022), "Minimizing Sensitivity to Model Misspecification," *Quantitative Economics*, 13, 907–954.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Cantoni, E. and Ronchetti, E. (2001), "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Association*, 96, 1022–1030.

Chernozhukov, V. and Hansen, C. (2008), "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398.

Clarke, B. R. (1983), "Uniqueness and Fréchet Differentiability of Functional Solutions to Maximum Likelihood Type Equations," *Annals of Statistics*, 1196–1205.

— (1986), "Nonsmooth Analysis and Fréchet Differentiability of M-functionals," *Probability Theory and Related Fields*, 73, 197–209.

Earle, C. C., Tsai, J. S., Gelber, R. D., Weinstein, M. C., Neumann, P. J., and Weeks, J. C. (2001), "Effectiveness of Chemotherapy for Advanced Lung Cancer in the Elderly: Instrumental Variable and Propensity Analysis," *Journal of Clinical Oncology*, 19, 1064–1070.

Finlay, K. and Magnusson, L. M. (2009), "Implementing Weak-Instrument Robust Tests for a General Class of Instrumental-Variables Models," *The Stata Journal*, 9, 398–421.

Freue, G. V. C., Ortiz-Molina, H., and Zamar, R. H. (2013), "A Natural Robustification of the Ordinary Instrumental Variables Estimator," *Biometrics*, 69, 641–650.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Heritier, S. and Ronchetti, E. (1994), "Robust Bounded-Influence Tests in General Parametric Models," *Journal of the American Statistical Association*, 89, 897–904.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Statistics*, 35, 73–101.

Huber, P. J. and Ronchetti, E. M. (2009), *Robust Statistics*, New York: Wiley, 2nd ed.

Ichimura, H. and Newey, W. K. (2022), "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, 13, 29–61.

Jiao, X. (2022), "A Simple Robust Procedure in Instrumental Variables Regression," Tech. rep.

Jun, S. J. (2008), "Weak Identification Robust Tests in an Instrumental Quantile Model," *Journal of Econometrics*, 144, 118–138.

Keane, M. and Neal, T. (2023), "Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice," *Journal of Econometrics*.

Kern, H. L. and Hainmueller, J. (2009), "Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes," *Political Analysis*, 17, 377–399.

Kitamura, Y., Otsu, T., and Evdokimov, K. (2013), "Robustness, Infinitesimal Neighborhoods, and Moment Restrictions," *Econometrica*, 81, 1185–1201.

Kleibergen, F. (2002), "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

— (2005), "Testing Parameters in GMM Without Assuming That They Are Identified," *Econometrica*, 73, 1103–1123.

Klooster, J. and Zhelonkin, M. (2023), "Outlier Robust Inference in the Instrumental Variable Model With Applications to Causal Effects," *Journal of Applied Econometrics (forthcoming)*.

Lee, D. S., McCrary, J., Moreira, M. J., and Porter, J. (2022), "Valid t-Ratio Inference for IV," *American Economic Review*, 112, 3260–90.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and Anna di Palma, M. (2023), *robustbase: Basic Robust Statistics*, r package version 0.99-0.

Magnusson, L. M. (2010), "Inference in Limited Dependent Variable Models Robust to Weak Identification," *The Econometrics Journal*, 13, 56–79.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019), *Robust Statistics: Theory and Methods (With R)*, New York: Wiley, 2nd ed.

Moreira, M. J. (2003), "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

— (2009), "Tests With Correct Size When Instruments Can Be Arbitrarily Weak," *Journal of Econometrics*, 152, 131–140.

Nelder, J. A. and Wedderburn, R. W. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135, 370–384.

Nelson, C. and Startz, R. (1990), "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One," *The Journal of Business*, 63, 5125––5140.

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

Powell, J. L. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 1435–1460.

Ronchetti, E. (1982), "Robust Testing in Linear Models: The Infinitesimal Approach," Ph.D. thesis, ETH Zürich.

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

Sølvsten, M. (2020), "Robust Estimation With Many Instruments," *Journal of Econometrics*, 214, 495–512.

Staiger, D. and Stock, J. H. (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586.

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

von Mises, R. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *Annals of Statistics*, 18, 309–348.

Wedderburn, R. W. (1974), "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss—Newton Method," *Biometrika*, 61, 439–447.

Young, A. (2022), "Consistency Without Inference: Instrumental Variables in Practical Application," *European Economic Review*, 147, 104–112.

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