Outlier Robust Inference in the Instrumental Variable Model With Applications to Causal Effects*

Jens Klooster†, Mikhail Zhelonkin‡

October 1, 2021

Abstract

The instrumental variable model is one of the central tools for the analysis of causal relationships in observational data. The Anderson and Rubin (1949) test is an important method that allows for reliable inference in the instrumental variable model when the instruments are weak. Yet, the robustness properties of this test have not been formally studied. As it turns out that the Anderson-Rubin (AR) test is not robust to outliers, we show how to construct an outlier robust alternative - the robust AR test. We investigate the robustness properties of the robust AR test and show that the robust AR statistic asymptotically follows a chi-square distribution. The theoretical results are illustrated by a simulation study. Finally, we apply the robust AR test to three different case studies that are affected by different types of outliers.

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

*We are very grateful to Andreas Alfons, Frank Kleibergen, Mikkel Sølvsten, Chen Zhou and members of the statistics group at the Econometric Institute for helpful comments and suggestions.

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

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

1 Introduction

The instrumental variable (IV) model is a standard tool for causal analysis in observational studies (Greenland, 2000; Sovey and Green, 2011; Bollen, 2012). A problem occurs when in the linear regression model a covariate is correlated with the error term, i.e., the covariate is endogenous. This can happen due to several reasons, for instance, omitted variables, measurement error and feedback relations between the variables. In this case, one can involve instrumental variables to solve the endogeneity problem. The instrumental variables should be uncorrelated with the error term, but correlated with the endogeneous covariate in the linear regression model. When instrumental variables are available, IV estimation and testing procedures allow for correct estimation and inference in the presence of an endogeneous covariate.

In practice, however, it is difficult to find valid instruments and the instruments that are used are often only weakly correlated with the endogeneous covariate (Andrews et al., 2019). When the instruments are "weak", this can lead to biased IV estimates and incorrect confidence intervals (Bound et al., 1995). For this reason, it is common to first conduct a pre-test, by means of an F-test, to measure the strength of the instruments. Then, depending on the strength of the instruments, a decision is made as to what procedure to follow in the statistical inference. If the instruments are strong, then the standard asymptotic theory for IV is valid (Greene, 2012, Chapter 10), and inference is typically done using a t-test based on a two-stage least squares (2SLS) estimate. However, when the instruments are weak, one has to rely on weak instrument robust procedures such as the Anderson-Rubin (AR) test (Anderson and Rubin 1949, see also Kleibergen 2002, Moreira 2003).

The classical parametric methods mentioned above such as the F-test, the 2SLS estimator and the AR test all rely heavily on distributional assumptions. When these assumptions fail or when they hold only approximately, the estimator and tests can break down. For instance, when there are outliers in the data, the 2SLS estimator and F-test can become arbitrarily

biased (see Freue et al. 2013 for the 2SLS estimator and Ronchetti 1982 for the F-test). In recent literature, Andrews et al. (2019) and Young (2021) raised concern that many results and conclusions in applied work were influenced by one or a small cluster of outliers. Young (2021) showed that in empirical work, the weak instrument pre-test based on the F-statistic is largely uninformative. A situation that can occur is that the instrument is "seemingly" strong due to an outlier that corrupts the first stage F-test by making the test statistic sufficiently large. Then, a researcher wrongfully assumes that it is legitimate to use a t-test based on a 2SLS estimate in the second stage instead of a weak instrument robust testing procedure, which then leads to incorrect inference. To illustrate this with an empirical example, we show a scatter plot of the first stage variables used in Ananat (2011) in Figure 1. We are interested in the strength of the linear relationship between the dependent variable y_1 and the instrument x_2, while controlling for the variable x_1. When we analyze Figure 1 it is clear that there is (at least) one point in the control variable x_1 that is an outlier. Removal of this outlier from the data results in the first stage F-statistic dropping from 19.32 to 1.97. Then, based on the typical threshold of 10 (Staiger and Stock, 1997), one concludes that the instrument is weak and inference should be done using the AR test. However, removing outliers and then using the AR test is not recommended for several reasons such as masking effects and underestimated variability and it is better to use a robust method from the start (Maronna et al., 2019; Chen and Bien, 2020).

Therefore, in this paper, we introduce a robust AR test that allows for reliable inference when the data might contain some outliers. We start by formally studying the robustness properties of the (classical) AR test under small data contamination. As it turns out that the AR test is not robust, we show how we can "robustify" the AR test by utilizing methods from the field of robust statistics (Hampel et al. 1986, Huber and Ronchetti 2009), which results in the robust AR test. Similar to the AR test, the robust AR test remains size correct irrespective of the strength of the instrument(s). Therefore, it is not necessary to rely on a (nonrobust)

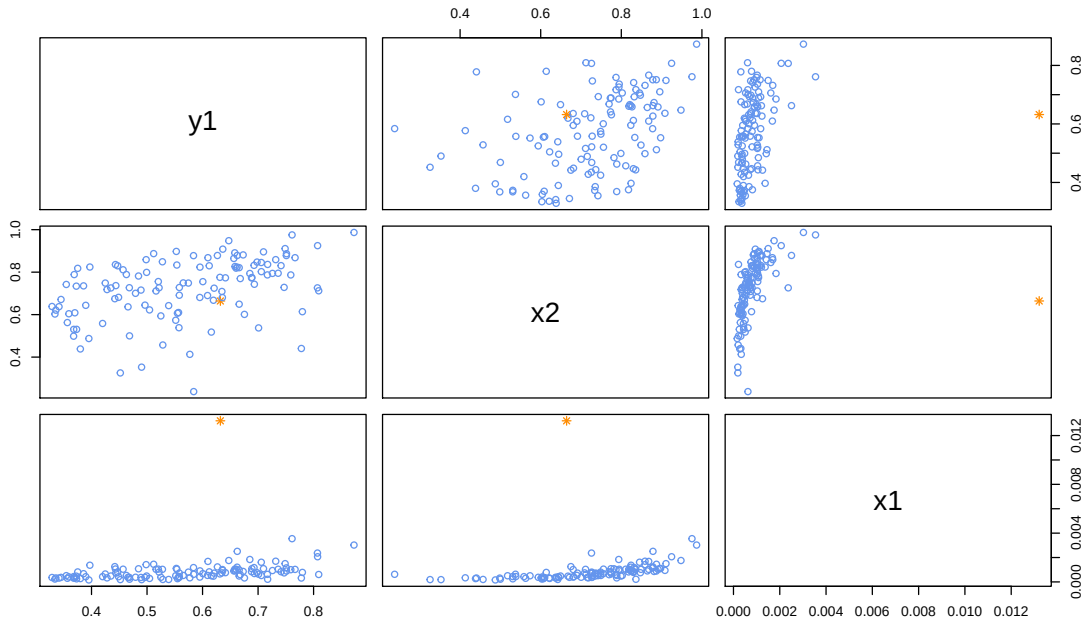

Figure 1: Scatter plot of the first stage variables in Ananat (2011). The outlier is denoted as an orange star.

pre-test to measure the strength of the instrument(s). In practice, we can start by applying the robust AR test and analyze its confidence set. It turns out that this procedure is weak instrument robust and robust to outliers. In case of the Ananat (2011) example, the robust AR test allows reliable inference from the start without relying on outlier detection methods as we further demonstrate in Section 5.

Throughout this paper we assume that the parametric model holds approximately. We use the Huber (1964) gross-error model $F_\epsilon = (1-\epsilon)F+\epsilon G$, where ϵ denotes a (small) contamination proportion. We are interested in the estimation and inference for the central model F, but we observe data from F_ϵ. The distribution G is assumed to be completely unknown. The desirable estimators and tests should be accurate at F and stable when the data comes from a model in the neighborhood F_ϵ. Our work is related to the growing literature on local misspecification, see Andrews et al. (2017), Andrews et al. (2020), Bonhomme and Weidner (2021), Ichimura and Newey (2021), Kitamura et al. (2013) and references therein. The closest approach is the one by Bonhomme and Weidner (2021), where the minimax robust estimators and confidence

intervals under misspecifications of specific aspects of the model are derived. We leave the misspecification component to be completely unspecified. This is natural in our case, since our primary goal is to be resistant to outliers, which can be generated from an arbitrary data generating process. This framework allows us to facilitate the practical case discussed by Young (2021) where there are a few outlying observations in the data.

Note that, if the researcher concludes that the model F does not hold even approximately, then the researcher would have to resort to semi- or nonparametric methods (in case of the AR test see for example Jun 2008, Andrews and Marmer 2008). However, nonparametric methods are usually not developed to be robust in the sense described above. A trivial example is the sample mean, which is a nonparametric estimator of the expectation, but it is not robust, since one outlier is sufficient to make it arbitrarily biased; see Huber and Ronchetti (2009, p. 6) for further discussion. The method we consider is parametric in nature and similar to the fully parametric case. The (robust) estimators on which the robust AR test relies identify exactly the same parameters as in the fully parametric case. Thus, it allows to enjoy the simplicity and interpretability of the parametric model. The robust estimators are slightly less efficient than the classical estimators when the model F holds exactly. This leads to a small loss of power of the robust AR test compared to the classical AR test. However, as we show in our simulation study, when there is a (small) portion of contamination, the classical AR test becomes less efficient than the robust AR test.

In Section 2 we present the IV model, set up the notation and introduce the classical AR test. In Section 3 we show that the classical Anderson-Rubin test is not robust and propose a robust alternative - the robust AR test. We show that the robust AR test is asymptotically chi-square distributed. Then we investigate the small sample properties of the tests by means of a simulation study in Section 4. Three case studies are presented in Section 5. Finally, Section 6 concludes the paper.

2 Instrumental Variable Model and Weak Instruments

We are motivated by the practically relevant instrumental variable setting with one endogeneous covariate and several instruments (Andrews et al., 2019). We consider the following IV model:

$$y_1 = \mathbf{x}_1^\top \gamma_1 + \mathbf{x}_2^\top \pi + \epsilon_1, \tag{1}$$

$$y_2 = \beta y_1 + \mathbf{x}_1^\top \gamma_2 + \epsilon_2, \tag{2}$$

where y_1 is an endogeneous regressor, $\mathbf{x}_2$ is a $p_2 \times 1$ vector of instrumental variables and $\mathbf{x}_1$ is a $p_1 \times 1$ vector of control variables. We assume that γ_1 and γ_2 are a $p_2 \times 1$ vectors of constants that both include an intercept, π is a $p_1 \times 1$ vector of constants and β is a scalar constant. The error terms (ϵ_1, ϵ_2) follow a symmetric distribution with mean zero and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$ and are assumed to be independent of the instruments and control variables. We assume that the control and instrumental variables $(\mathbf{x}_1, \mathbf{x}_2)$ are random variables with distribution K. Let $N \in \mathbb{N}$, then we assume that in practice we only observe independent realizations $\{(y_{1i}, y_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}) \mid i = 1, \ldots, N\}$ of the random variables $(y_1, y_2, \mathbf{x}_1, \mathbf{x}_2)$.

Equation (1) is referred to as the first stage and is a reduced form equation that explains the endogeneous regressor y_1 using the variables $\mathbf{x}_1$ and $\mathbf{x}_2$. Equation (2) is referred to as the second stage and is a structural form equation that explains y_2 by the endogeneous regressor y_1 and the variable $\mathbf{x}_1$. In the model (1)-(2) we are interested in testing the hypothesis

$$H_0 \colon \beta = \beta_0 \text{ vs. } H_1 \colon \beta \neq \beta_0, \tag{3}$$

which is typically interpreted as the causal effect of y_1 on y_2.

In general, the instruments $\mathbf{x}_2$ should be uncorrelated with the error terms and correlated with y_1, i.e., $\mathbf{x}_2$ needs to be a significant predictor of y_1. When $\mathbf{x}_2$ is only weakly correlated with y_1 it can lead to biased IV estimates and erroneous confidence intervals (Bound et al., 1995). Therefore, in the applied literature (e.g., Acconcia et al. 2014, Stephens Jr and Yang 2014),

inference on β is typically done in a sequential fashion. First, the strength of the instruments is determined by performing an F-test to test whether $\pi = \mathbf{0}$ in (1). When the first stage F-statistic is above 10 (see Staiger and Stock 1997 for a motivation on this threshold), then β is estimated using a 2SLS estimator and inference in (3) is done with a t-test. When the first stage F-statistic is lower than 10, the researcher has to resort to weak instrument robust testing procedures, see Anderson and Rubin (1949), Kleibergen (2002), Moreira (2003).

The idea of the AR test is that under the null hypothesis (3) the instruments $\mathbf{x}_2$ have no explanatory power for $y_2 - \beta_0 y_1$. That is, if the null hypothesis is true, then we must have $y_2 - \beta_0 y_1 = \mathbf{x}_1^\top \gamma_2 + \epsilon_2$. Therefore, we can test the null hypothesis (3) as follows: we consider the model

$$y_2 - \beta_0 y_1 = \mathbf{x}_1^\top \gamma_2 + \mathbf{x}_2^\top \delta + \epsilon_2 \tag{4}$$

and test whether

$$H_0^*\colon \delta = \mathbf{0} \text{ vs. } H_1^*\colon \delta \neq \mathbf{0}. \tag{5}$$

The AR test is

$$AR_N(\beta_0) = \frac{(\mathbf{Y}_2 - \beta_0 \mathbf{Y}_1)^\top \mathbf{P}_{\tilde{\mathbf{X}}_2} (\mathbf{Y}_2 - \beta_0 \mathbf{Y}_1)/p_2}{\hat{\sigma}_2^2(\beta_0)}, \quad \text{where} \tag{6}$$

$$\hat{\sigma}_2^2(\beta_0) = (\mathbf{Y}_2 - \beta_0 \mathbf{Y}_1)^\top \mathbf{M}_{\mathbf{X}} (\mathbf{Y}_2 - \beta_0 \mathbf{Y}_1)/(N - p_1 - p_2), \tag{7}$$

where $\mathbf{Y}_2$ and $\mathbf{Y}_1$ denote the $N \times 1$ vectors with entries y_{2i} and y_{1i} respectively, $\mathbf{X}_1 \in \mathbb{R}^{N \times p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{N \times p_2}$ denote the matrices with rows $\mathbf{x}_{1i}^\top$ and $\mathbf{x}_{2i}^\top$ respectively, and $\mathbf{M}_{\mathbf{X}} := \mathbf{I}_N - \mathbf{P}_{\mathbf{X}}$, where $\mathbf{P}_{\mathbf{X}} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ and $\mathbf{X}$ is the matrix with rows $(\mathbf{x}_{1i}^\top, \mathbf{x}_{2i}^\top)$ and $\tilde{\mathbf{X}}_2 = \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2$.

Under general regularity conditions (see Assumption M in Staiger and Stock 1997), the AR test rejects the null hypothesis (3) when the test statistic $p_2 AR_N(\beta_0)$ exceeds the $(1-\alpha)$-quantile of the $\chi^2_{p_2}$ distribution. To obtain the confidence set of the AR test we can make use of test inversion (Mikusheva, 2010). 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 $\left\{\beta_0 \mid AR_N(\beta_0) < \frac{\chi^2_{p_2,1-\alpha}}{p_2}\right\}$. It is well

known that under the stricter assumption of normally distributed error terms, the $AR_N(\beta_0)$ statistic is $F_{p_2,N-p_1-p_2}$ distributed. In the simulation study and the practical case studies, we do not rely on this assumption due to the recent doubts about this assumption in practice, as shown by Young (2021).

3 Robust Anderson-Rubin Test

In this section we investigate the robustness properties of the Anderson-Rubin test and propose a robust alternative.

3.1 Influence Function of the AR Test

We consider the parametric IV model (1)-(2) by focusing on the simple linear regression model (4). We assume that (4) is governed by F_θ, where $\theta := (\gamma_2, \delta)^\top$ lies in Θ, a compact subset of $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$. Let F_N denote the empirical distribution function with mass $1/N$ at each observation $\mathbf{w}_i = (\mathbf{x}_{1i}^\top, \mathbf{x}_{2i}^\top, y_{2i}, y_{1i})$ and let F denote the distribution of $\mathbf{w} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top, y_2, y_1)$.

To determine whether an estimator is robust, we can analyze its influence function (IF). For a statistical functional T, the influence function is defined (Hampel, 1974) as

$$IF(\mathbf{w}; T, F) = \lim_{\epsilon \downarrow 0} [T\{(1-\epsilon)F + \epsilon \Delta_{\mathbf{w}}\} - T(F)]/\epsilon, \tag{8}$$

where $\Delta_{\mathbf{w}}$ is a point mass at $\mathbf{w}$. The IF characterizes the standardized asymptotic bias of the estimator due to contamination ϵ. If the IF is unbounded then the worst possible bias in the neighborhood of F can be infinite. Hence, for an estimator to be (locally) robust, a bounded IF is required.

Similar to the influence function of an estimator, the test influence function describes the effect of an outlier on the test statistic that is used to do inference. The test influence function can be viewed as a generalization of the influence function when the estimator is not Fisher consistent (Rousseeuw and Ronchetti, 1981). In practice, to study the robustness properties of the test statistic it is sufficient to use (8) as it is proportional to the test influence function

up to a constant, even though the test statistic might not be Fisher consistent (Hampel et al., 1986, Chapter 3). We provide the expression of the IF of the AR statistic in the following proposition.

Proposition 1. *The IF of the AR statistic is given by*

$$IF(\mathbf{w}; \sqrt{AR(\beta_0)}, F_{\theta_0}) = \left| \frac{y_2 - \beta_0 y_1 - \mathbf{x}_1^\top \gamma_2}{\sigma_2} \right| \left\{ \begin{pmatrix} \mathbf{x}_1^\top & \mathbf{x}_2^\top \end{pmatrix} (\tilde{\mathbf{M}}^{-1} - \tilde{\mathbf{M}}^{+}) \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \end{pmatrix} / p_2 \right\}^{\frac{1}{2}}, \quad (9)$$

where $\tilde{\mathbf{M}}$ *and* $\tilde{\mathbf{M}}^{+}$ *are constant matrices that are defined as:*

$$\tilde{\mathbf{M}} = \int \begin{pmatrix} \mathbf{x}_1\mathbf{x}_1^\top & \mathbf{x}_1\mathbf{x}_2^\top \\ \mathbf{x}_2\mathbf{x}_1^\top & \mathbf{x}_2\mathbf{x}_2^\top \end{pmatrix} dK \quad \textit{and} \quad \tilde{\mathbf{M}}^{+} = \int \begin{pmatrix} \left(\mathbf{x}_1\mathbf{x}_1^\top\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} dK.$$

Proof. The result follows from (Ronchetti, 1982), since the AR test is a particular case of Ronchetti's τ-test with the function $\tau(\mathbf{x}, r) = \frac{r^2}{2}$ (see also Hampel et al. 1986, p.348). □

Note the square root of $AR(\beta_0)$ (denoting the functional corresponding to $AR_N(\beta_0)$) is taken to make sure the IF is nonzero. The IF (9) is unbounded in both the dependent variable $y_2 - \beta_0 y_1$ and the covariate space $(\mathbf{x}_1, \mathbf{x}_2)$ and therefore not (locally) robust. This implies that an outlier in either the first stage or second stage can break down the AR test, even when the outlier is only in the control variable.

3.2 Robust AR Test

Since the classical AR test is not robust we propose a robust alternative - the robust AR test. We use the robust F-test by Markatou and Hettmansperger (1990) to robustify the AR test. Similar to the AR test, we test (3) by testing $H_0^*\colon \delta = \mathbf{0}$ in $y_2 - \beta_0 y_1 = \mathbf{x}_1^\top \gamma_2 + \mathbf{x}_2^\top \delta + \epsilon_2$. However, instead of relying on the AR test, we use the following robust AR test:

$$W_N^2(\beta_0) := \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \Psi \left(\frac{y_{2i} - \beta_0 y_{1i} - \mathbf{x}_{1i}^\top \hat{\gamma}_2}{\hat{\sigma}_2}; c \right) \mathbf{x}_{2i} \right\}^\top \hat{\mathbf{U}}^{-1} \cdot \quad (10)$$
$$\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \Psi \left(\frac{y_{2i} - \beta_0 y_{1i} - \mathbf{x}_{1i}^\top \hat{\gamma}_2}{\hat{\sigma}_2}; c \right) \mathbf{x}_{2i} \right\},$$

where $\hat{\mathbf{U}}$ is a consistent estimate of the asymptotic variance-covariance matrix, formally defined in Proposition 2, $\hat{\gamma}_2$ is estimated using an MM-estimator (Yohai, 1987) in the null-restricted

model and $\hat{\sigma}_2$ is estimated robustly using an S-estimator in the null-restricted model. We use Tukey's Biweight function for Ψ

$$\Psi(r;c) := \begin{cases} r\left(1-\frac{r^2}{c^2}\right)^2, & \text{for } |r| \leq c, \\ 0, & \text{for } |r| > c, \end{cases} \tag{11}$$

where c denotes a tuning parameter that can be tuned to obtain a certain level of asymptotic efficiency (we choose $c = 4.68$ when $p_2 = 1$ and refer to Table 1 in Copt and Heritier (2007) for the optimal choices of c when $p_2 > 1$). The benefit of the Tukey Biweight function is that it downweights large outliers to zero. The function ω is a weight function for which several options are available. A first simple and practical (Cantoni and Ronchetti 2001) option that is the weight function $\omega(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \sqrt{1-h_i}$, where h_i is the i-th diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. It ensures a bounded IF, but it does not have a high breakdown point, i.e., it does not allow for a large proportion of outliers. A second more robust (in terms of the breakdown point) option is to use weights based on the robust Mahalanobis distance $d(x_1, x_2)$:

$$w(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 1, & \text{if } d(\mathbf{x}_1, \mathbf{x}_2) \leq \tilde{c}, \\ \tilde{c}/d(\mathbf{x}_1, \mathbf{x}_2), & \text{if } d(\mathbf{x}_1, \mathbf{x}_2) > \tilde{c}. \end{cases}$$

Since the squared Mahalanobis distance follows a χ^2-distribution under a normality assumption, a common choice is a 5% critical level for $\tilde{c}$. The expectation and covariance that are used to calculate the robust Mahalanobis distance are estimated by the Minimum Covariance Determinant (MCD), see Rousseeuw and Driessen (1999). However, since applications often contain covariate spaces with a lot of dummy variables, the computation of the MCD can become infeasible. For this reason, we rely on the first approach, when weighting is necessary. Note the MM-estimator we use to estimate γ_2 also uses Tukey's Biweight function. Furthermore, if we use weights for the test statistic, then we use the same weights also for the MM-estimator. In practice, we implement this using the function `lmrob` from the R package `robustbase` (Maechler et al., 2021).

In order to justify the use of the proposed test statistic $W_N^2(\beta_0)$ we need its asymptotic

distribution and boundedness of the IF. First, we introduce several new matrices:

$$\mathbf{M} = \int \frac{\omega(\mathbf{x}_1, \mathbf{x}_2)}{\sigma_2} \frac{\partial \Psi}{\partial r} \left(\frac{y_2 - \beta_0 y_1 - \mathbf{x}_1^\top \gamma_2}{\sigma_2}; c \right) \begin{pmatrix} \mathbf{x}_1 \mathbf{x}_1^\top & \mathbf{x}_1 \mathbf{x}_2^\top \\ \mathbf{x}_2 \mathbf{x}_1^\top & \mathbf{x}_2 \mathbf{x}_2^\top \end{pmatrix} dF,$$
$$\mathbf{Q} = \int \omega(\mathbf{x}_1, \mathbf{x}_2)^2 \Psi \left(\frac{y_2 - \beta_0 y_1 - \mathbf{x}_1^\top \gamma_2}{\sigma_2}; c \right)^2 \begin{pmatrix} \mathbf{x}_1 \mathbf{x}_1^\top & \mathbf{x}_1 \mathbf{x}_2^\top \\ \mathbf{x}_2 \mathbf{x}_1^\top & \mathbf{x}_2 \mathbf{x}_2^\top \end{pmatrix} dF,$$

and we denote the matrices $\hat{\mathbf{M}}$ and $\hat{\mathbf{Q}}$ as the matrices $\mathbf{M}$ and $\mathbf{Q}$ integrated with respect to the empirical distribution function F_N. For a matrix $\mathbf{A} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}$, let $\mathbf{A}_{11}$ denote the upper left $p_1 \times p_1$ part of the matrix $\mathbf{A}$, let $\mathbf{A}_{12}$ denote the upper right $p_1 \times p_2$ part of $\mathbf{A}$, let $\mathbf{A}_{21}$ denote the lower left $p_2 \times p_1$ part of the matrix $\mathbf{A}$ and let $\mathbf{A}_{22}$ denote the lower right $p_2 \times p_2$ part of the matrix $\mathbf{A}$. Using this notation, we define

$$\mathbf{U} = \mathbf{Q}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{Q}_{12} - \mathbf{Q}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12} + \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{Q}_{11}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}.$$

Proposition 2. *Assume that the regularity conditions* $(1) - (7)$ *and* $(a) - (b)$ *from the supplementary material hold. Under the null hypothesis in (5) the test statistic (10) converges in distribution to a* $\chi^2_{p_2}$*, where*

$$\hat{\mathbf{U}} = \hat{\mathbf{Q}}_{22} - \hat{\mathbf{M}}_{21}\hat{\mathbf{M}}_{11}^{-1}\hat{\mathbf{Q}}_{12} - \hat{\mathbf{Q}}_{21}\hat{\mathbf{M}}_{11}^{-1}\hat{\mathbf{M}}_{12} + \hat{\mathbf{M}}_{21}\hat{\mathbf{M}}_{11}^{-1}\hat{\mathbf{Q}}_{11}\hat{\mathbf{M}}_{11}^{-1}\hat{\mathbf{M}}_{12}, \tag{12}$$

and $\hat{\gamma}_2$ *and* $\hat{\sigma}_2$ *are null-restricted* M *and* S*-estimates.*

Proof. For a proof of this result in a general setting, we refer to Proposition 2 in Heritier and Ronchetti (1994). □

Analogous to the AR test, we use test inversion to obtain the confidence set of the robust AR test. That is, we find all values of β_0 for which the data does not reject the null hypothesis. The confidence set is then $\{\beta_0 \mid W_N^2(\beta_0) < \chi^2_{p_2,1-\alpha}\}$. In practice, we can find this confidence set by specifying a grid of β_0 values and checking for each value of β_0 whether the data rejects the null hypothesis. In a setting without outliers, this should lead to a confidence set that is very similar to the confidence set of the AR test as we see below.

Recall that in the model (1)-(2) we can make use of the moment conditions $\int \mathbf{x}_2\epsilon_2 \, dF = \mathbf{0}$. The numerator of the AR test (6) is a quadratic form that checks whether these moment conditions hold at β_0. Note, as Ψ is an odd function, we have

$$\int \mathbf{x}_2\epsilon_2 \, dF = \mathbf{0}, \text{ implies that } \int \omega(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2\Psi(\epsilon_2; c) \, dF = \mathbf{0}.$$

Hence, the robust AR test (10) checks whether these moment conditions hold at β_0. As Tukey's Biweight function is bounded, this limits the possible influence of outliers. However, only using the Tukey Biweight function is not sufficient to be robust since outliers can potentially be in the instrumental variable term $\mathbf{x}_2$ that is not downweighted by Tukey's Biweight function. For this reason, we also need to rely on the weight function ω. This is formalized in the following proposition based on Theorem 3.1 in Markatou and Hettmansperger (1990).

Proposition 3. *The IF of the robust AR statistic is*

$$IF(w; W(\beta_0), F_{\theta_0}) = \left|\omega(\mathbf{x}_1, \mathbf{x}_2)\Psi\left(\frac{y_2 - \beta_0 y_1 - \mathbf{x}_1^\top\gamma_2}{\sigma_2}; c\right)\right| \cdot \left\{\left(-\mathbf{x}_1^\top\mathbf{M}_{11}^{-1}\mathbf{M}_{12} + \mathbf{x}_2^\top\right)\mathbf{U}^{-1}\left(-\mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{x}_1 + \mathbf{x}_2\right)\right\}^{\frac{1}{2}}, \tag{13}$$

Proof. The proof of this result is similar to the proof of Proposition 1 in (Hampel et al., 1986, p.348). □

Markatou and Hettmansperger (1990) formally show that the IF (13) is bounded when the weight function based on the $\mathbf{H}$ matrix is used. If we do not use any weight function (i.e., every observation obtains the same weight) and only rely on Tukey's Biweight function to downweight the outliers, then the IF is unbounded. Hence, in general, to obtain a fully robust test using weights is necessary. However, in practice, using weights is not always needed. For example, when there is only a single large outlier in the variable y_2, y_1 or $\mathbf{x}_1$ then it is not needed to use weights as Tukey's Biweight function will downweight the outlier.

4 Simulation Study

In the simulation study we investigate the performance of the robust AR test (10) compared to the AR test (4) and the t-test based on the 2SLS estimator. We analyze the robustness and the performance of the different tests by analyzing their power curves in different (contaminated) settings.

4.1 Data Generation

We consider four different settings: one setting without contamination and three settings with contamination. We consider two settings with contamination by outliers and one setting with a distributional contamination in the error terms.

First, we generate uncontaminated data. We simulate data from the following model:

$$y_1 = x_1 + \pi x_2 + \epsilon_1, \tag{14}$$

$$y_2 = \beta y_1 + 2x_1 + \epsilon_2, \tag{15}$$

which is a particular case of the model (1)-(2). We generate one exogeneous instrument x_2 and one exogeneous control variable x_1 that both follow a standard normal distribution. The error terms ϵ_1 and ϵ_2 follow a bivariate normal distribution with variances equal to 1 and correlation $\rho = 0.25$. We consider two different cases for the parameter $\pi \in \{0.1, 1\}$. When $\pi = 0.1$ we are in the weak instrument setting and when $\pi = 1$ the instrument is strong. The sample size is $N = 250$ and we repeat the study 10000 times. In the simulation, we test (3) with $\beta_0 = 0$ at the 5% significance level.

In the first contamination, we simulate a setting with an outlier in the endogeneous variable y_2 by generating data from the uncontaminated model and then replacing the first data row by $(y_{21}, y_{11}, x_{11}, x_{21}) = (10, 2, 2, 3)$. In the second contamination, we simulate a setting with an outlier in the exogeneous control variable x_1 by generating data from the uncontaminated model and then replacing the first data row by $(y_{21}, y_{11}, x_{11}, x_{21}) = (2, 2, 10, 2)$. Finally, we simulate a third setting with a distributional contamination. We simulate the errors from a mixture

of a bivariate normal (as above) and a bivariate t-distribution, i.e., with 10% probability we simulate the error terms from a bivariate t-distribution with 2 degrees of freedom.

This particular simulation design serves two purposes. First, the goal of this simulation study is to show that in very simple situations, it is necessary to rely on the robust AR test for reliable inference. Second, our design mimics the situations that happen in the case studies in Section 5. The case study in Section 5.1 is affected by a large outlier in the control variable and the case study in Section 5.2 is affected by a large outlier(s) in the endogeneous variable. In the supplementary material, we also present an extension of the simulation study to a setting with multiple instruments and a higher correlation between the error terms.

4.2 Simulation Results

In Figure 2 we show the power curves of the three tests when there is no contamination. We see that when $\pi = 0.1$, corresponding to the weak instrument setting, both the AR test and the robust AR test have size correct power curves. As the power curve of the AR test is strictly above the power curve of the robust AR test, except when $\beta = 0$, we can conclude that the AR test is more powerful than the robust AR test when the instrument is weak. This is expected, as the robust AR test sacrifices some power to be more robust in contaminated settings, as we will see later. The t-test based on the 2SLS estimator is not size correct and its power curve is arbitrary due to the weak instrument. When $\pi = 1$, corresponding to the strong instrument setting, all tests are size correct and have comparable power curves. This is expected, as we are in a just-identified setting with a strong instrument. In this case, inference based on the 2SLS estimator is reliable and the AR test has optimal power properties (Andrews et al., 2006). In both the weak and strong instrument settings, we notice that the power curves of the robust AR test are only slightly lower than the power curves of the AR test, which implies that the extra robustness comes at a low cost.

In Figure 3 we show the power curves of the three tests when there is an outlier in the endogeneous variable y_2. In both the weak and strong instrument settings the power curves of

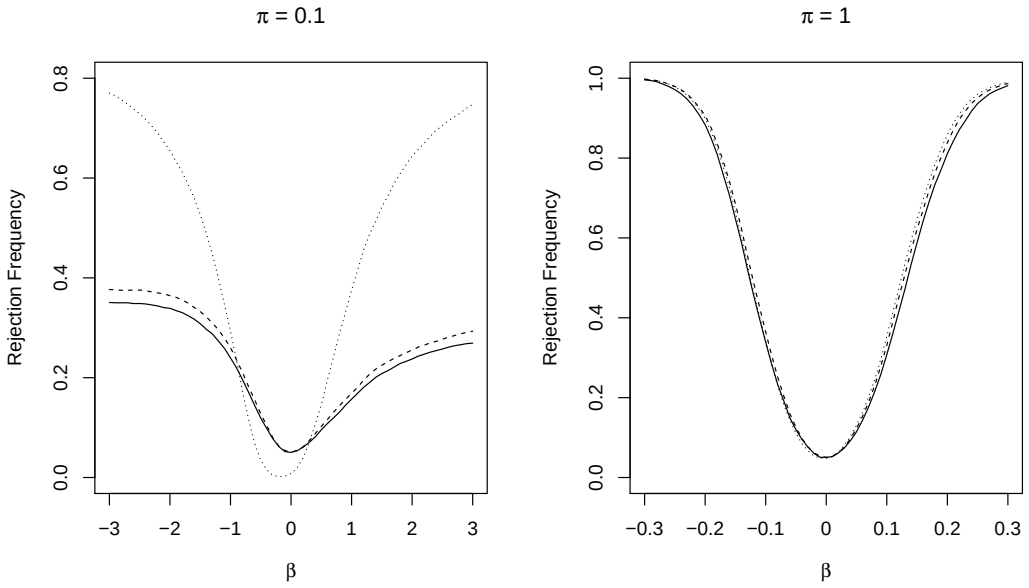

Figure 2: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with one instrument, $\rho = 0.25$ and no contamination.

the robust AR test are comparable to the power curves of the robust AR test in the setting without contamination. This shows that the outlier did not affect the robust AR test much. When we analyze the shape of the power curves of the AR test, we see that compared to the uncontaminated case, the minimum of the power curves shifted to the right. To understand how this happens, we consider the specific case $\beta_0 = 0$. As shown in Section 2, the AR test is testing the hypothesis $H_0^*\colon \delta = \mathbf{0}$ in $y_2 - \beta_0 y_1 = y_2 = \mathbf{x}_1^\top \gamma_2 + \mathbf{x}_2^\top \delta + \epsilon_2$ using an F-test. Hence, the AR test is testing whether there is a nonzero linear relationship between y_2 and $\mathbf{x}_2$, while controlling for $\mathbf{x}_1$. The outlier causes an upward bias in the OLS estimate of the $\mathbf{x}_2$ parameter in the unrestricted model used in the F-test. Subsequently, it happens more often that a positive linear relationship is detected due to the outlier, causing an overrejection at $\beta = 0$, and a shift to the right of the power curve. For the t-test this shift also happens, because the 2SLS estimator is biased upward due to the outlier. When the instrument is strong, the power curve of the t-test follows the power curve of the AR test. When the instrument is weak, the power curve is unreliable because of the weak instrument and the outlier.

In Figure 4 we show the power curves of the three tests when there is an outlier in the

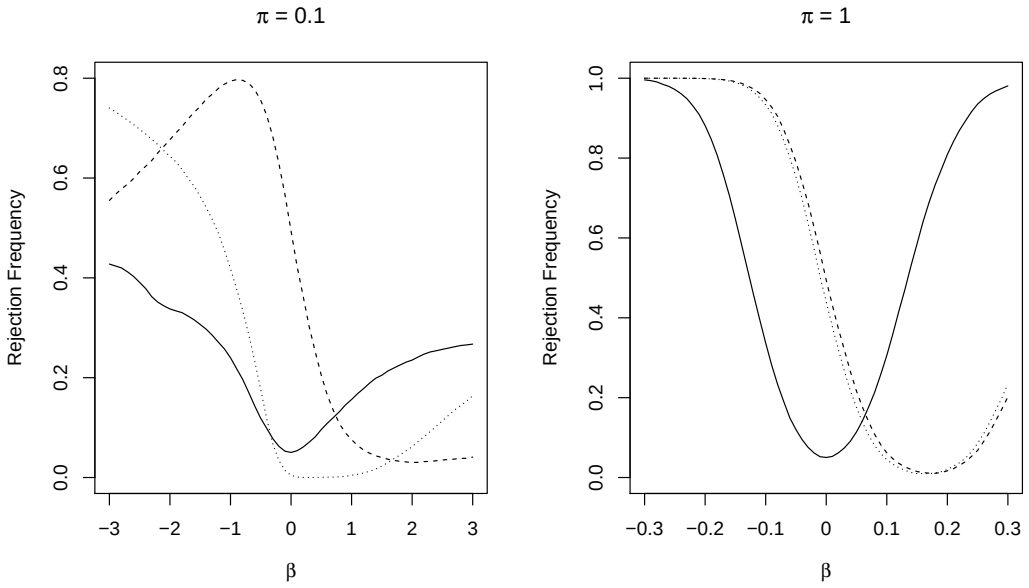

Figure 3: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with one instrument, $\rho = 0.25$ and an outlier in the endogeneous variable.

exogeneous control variable x_1. In both the weak and strong instrument settings the power curves of the robust AR test are comparable to the power curves of the robust AR test in the setting without contamination. This shows that the outlier did not affect the robust AR test much. When we analyze the shape of the power curves of the AR test, we see that compared to the uncontaminated case, the power curves have shifted to the right. To understand how the outlier in the control variable can break down the AR test we consider the often used technique of "partialling out" the control variable using the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963). When we partial out the control variable, we multiply the data matrices $\mathbf{Y}_2, \mathbf{Y}_1$ and $\mathbf{X}_2$ by the matrix $\mathbf{M}_{\mathbf{X}_1}$ and obtain $\tilde{\mathbf{Y}}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{Y}_2, \tilde{\mathbf{Y}}_1 = \mathbf{M}_{\mathbf{X}_1}\mathbf{Y}_1$ and $\tilde{\mathbf{X}}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$. Note that this procedure can transfer the outlier from the control variable to all other variables. Hence, when we use the AR test or 2SLS estimator, then an outlier in the control variable should be viewed as an outlier that is (possibly) in all variables. In Section 5 we illustrate how problematic the effect of such an outlier can be in a case study. The interested reader can already take a look at Figures 6 and 7 to get an idea how an outlier

in the control variable becomes an outlier in the other variables based on a real data example. When we consider the specific case $\beta_0 = 0$, then the AR test is testing whether $\tilde{H}_0^*\colon \delta = \mathbf{0}$ in $\tilde{y}_2 - \beta_0\tilde{y}_1 = \tilde{y}_2 = \tilde{\mathbf{x}}_2^\top\delta + \tilde{\epsilon}_2$ using an F-test. The outlier in the control variable, is now an outlier in the $\tilde{y}_2$ and $\tilde{\mathbf{x}}_2$ variables that causes an upward bias in the OLS parameter estimate of $\tilde{\mathbf{x}}_2$ in the unrestricted model. This causes an overrejection at $\beta = 0$, and a shift to the right of the power curve. We can use the same argument for the t-test based on the 2SLS estimate. The 2SLS estimator is biased upward due to the outlier. When the instrument is strong, the power curve of the t-test follows the power curve of the AR test. When the instrument is weak, the power curve is unreliable because of the weak instrument and the outlier.

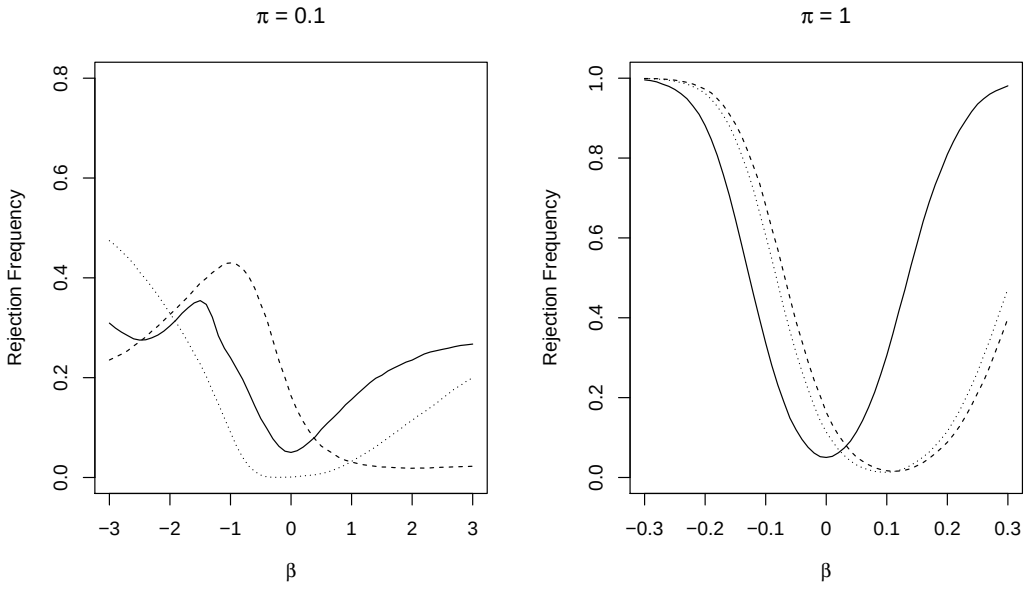

Figure 4: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with one instrument, $\rho = 0.25$ and an outlier in the exogeneous control variable.

Finally, in Figure 5 we show the power curves of the three tests in a setting with distributional contamination. The thicker tails of the $t(2)$-distribution introduce outliers in the error terms. On average, this happens in a symmetric way. Therefore, we see that in the strong instrument setting all tests remain size correct. However, we see that the robust AR test has a higher power than the t-test and the AR test. This happens, because Tukey's Biweight

function downweights large outliers in the error terms. The same is true when the instrument is weak and we see that the robust AR test has a higher power than the AR test. We see that the t-test behaves arbitrarily again due to the weak instrument.

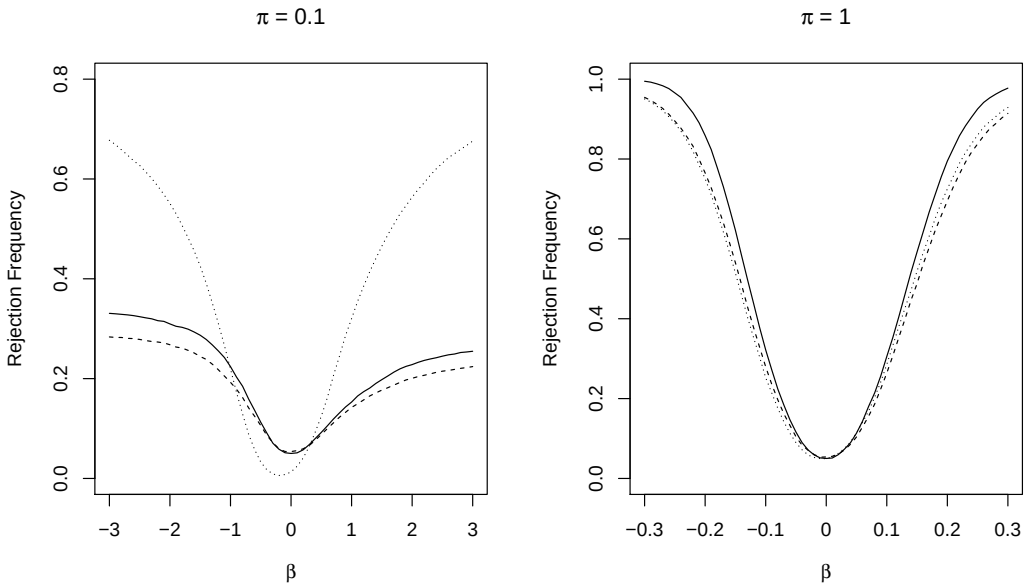

Figure 5: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with one instrument, $\rho = 0.25$ and contamination in the error term.

5 Case Studies

In this section we apply the robust AR test to several studies (Ananat 2011, Becker et al. 2011, Chodorow-Reich et al. 2012) that were also analyzed by Young (2021). These studies use the model (1)-(2) and the sequential procedure outlined in Section 1. From each study, we pick one IV regression and recalculate the first stage F-statistic and β confidence set used in the paper. Then we compute the confidence set of the robust AR test and compare it to the confidence set used in the paper and the confidence set of the AR test.

The three regressions we show are all affected by different types of outliers that are similar to the outliers we generated in the simulation study. In the first case study there are a few outliers in the control variable x_1, with one very large outlier. In the second case study the

outliers are mainly in the dependent variable y_2. Lastly, in the third study there is an outlier that is an outlier in every dimension. To detect these outliers, we run the BACON algorithm (Billor et al., 2000) with a 99^{th} percentile threshold. As all these case studies contain outliers, we recommend using the robust AR test. To further validate the results of the robust AR test, we remove the outliers from the data and compute the F-statistic, AR confidence set and robust AR confidence set again. When any large differences occur, we try to give a motivation of why this might have happened. Note that we only remove the outlier(s) for illustrative purposes to validate the results of the robust AR test. In general, removing outliers from the data and then using classical nonrobust methods, such as the AR test, is not recommended for several reasons. One reason is that after the data cleaning the data becomes non i.i.d., and the classical asymptotic results become invalid. Another reason is the masking effect, which happens when an outlier is found and removed, further ones can appear, and it is not clear when one should stop. For further discussion see Maronna et al. (2019, Section 4.3). Hence, it is recommended to rely on the robust AR test from the start.

5.1 The Causal Effect of Segregation on Economic Inequality

In this section we revisit the paper by Ananat (2011). The study is about the causal effect of racial segregation in cities within the United States on economic inequality. Due to possible endogeneity problems a function of nineteenth-century railroad configurations, conditional on total length of railroad, is used to instrument for the extent to which cities became segregated as they received inflows of African Americans during the twentieth century.

More specifically, the first stage explains segregation (y_1) by a railroad division index (x_2) and a control variable (x_1) that controls for the length of the railroad track. The second stage explains different poverty and inequality measures (y_2), using the endogeneous variable segregation (y_1) and the control variable (x_1). We pick one of the inequality measures and use the Gini coefficient for whites (y_2).

We start with an exploratory analysis of the data. In Figure 6 we present a scatter plot of

the main variables used in the first and second stage. We see that there are several outliers present in the data, most notable in the control variable x_1. To understand how the outlier in the control variable affects the outcome of the classical test statistics, we partial out the control variable and show the transformed data in Figure 7. Using Figure 7 it becomes easier to analyze how the outlier in the control variable affects the classical test statistics.

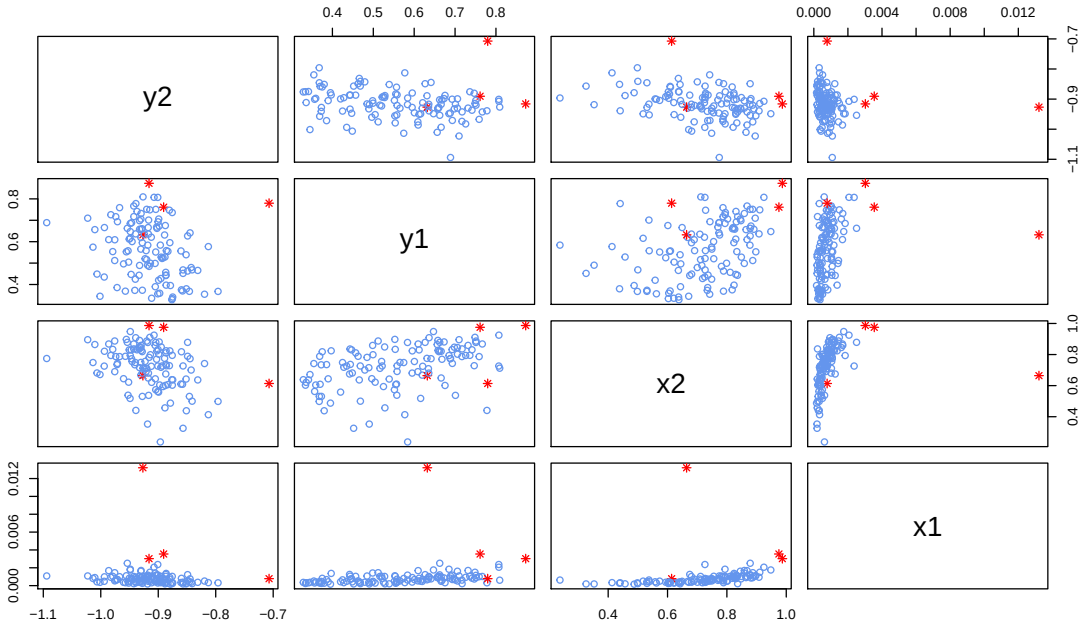

Figure 6: Scatter plot of the first and second stage variables in Ananat (2011). The red stars are flagged as outliers by the BACON algorithm.

In Table 1, we show the first stage F-statistic and β confidence set used in the paper. The first stage F-statistic is not explicitly given, but Ananat (2011) does mention that there is a strong first stage and refers to column 1 of Table 1. Using the regression results from this table, we calculate a first stage F-statistic of $(0.357/0.088)^2 = 16.45$, implying that the instrument is strong. Therefore, β is estimated using a 2SLS estimator and inference is done with a t-test. The result is $\hat{\beta} = -0.334$, with a (heteroskedasticity-robust) standard error of 0.099, implying that it is significant at the 5% level. The 95% confidence set is $(-0.53, -0.14)$. These results can be found in columns (1) (first stage) and (3) (second stage) of Tables 1 and 2 respectively in Ananat (2011).

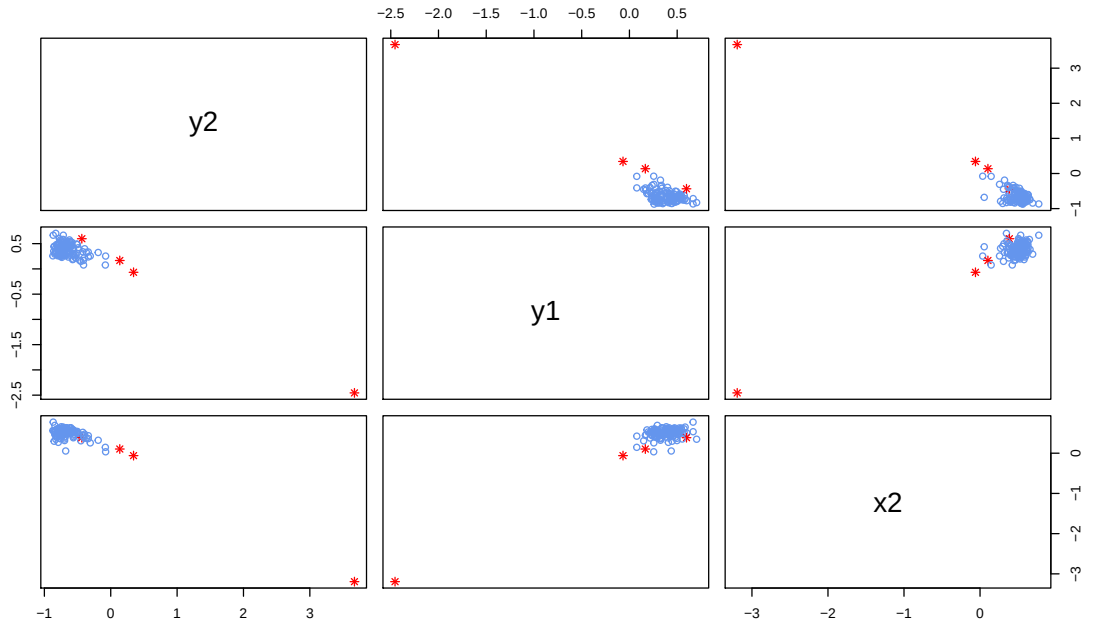

Figure 7: Scatter plot of the first and second stage variables in Ananat (2011), with the control variable partialled out. The red stars are flagged as outliers by the BACON algorithm.

First, we analyze the effect of the outlier on the first stage F-statistic. In Table 1 we show the first stage F-statistic when the outliers are in the data and when we remove the outliers from the data. When the outliers are in the data, the F-statistic is equal to 19.32, which implies that we can use the t-test based on the 2SLS estimate. Note that this first stage F-statistic is slightly different from the one that we calculated using the data from the paper because we did not use the heteroskedasticity-robust standard errors. When we remove the outliers from the data, we obtain an F-statistic of 1.83, which suggests the instrument is weak. This difference occurs because the first stage F-statistic depends on the linear relationship between the y_1 variable and the x_2 variable. When we analyze the blue open points in Figure 7 there does not seem to be a nonzero linear relationship between y_1 and x_2. However, when the outliers are not taken care of, the OLS estimate in the unrestricted model becomes biased toward the outlier(s). As a consequence, due to the outliers, there seems to be a strong relationship between y_1 and x_2. This phenomenon is also described in Ugarte Ontiveros and Verardi (2012) and Dehon et al. (2009).

Next, we analyze the effect of the outliers on the confidence set of the AR test. From Table 1 we see that when the outliers are in the data, then the confidence set of the AR test is an open convex set, which suggests the instrument is strong. When the outliers are removed, we see that the confidence set of the AR test changes to a union of two unbounded sets, which suggests the instrument is weak. To understand how this happens, we again analyze Figure 7. Let us consider the specific case where we test whether $\beta = 0$. In that case, as the AR test is an F-test, it tests whether there is a nonzero linear relationship between y_2 and x_2. We note that in this case, the outlier has a large effect as it biases the OLS estimate toward the outlier.

Finally, we study the confidence set of the robust AR test. As the outliers are most notable in the control variable, we do not use the weight function. In Table 1 we see that when the outliers are in the data, in contrast to the AR confidence set, the confidence set of the robust AR test is a union of two unbounded sets, which suggests the instrument is weak. When the outliers are taken out of the data, the confidence set of the robust AR test is a union of two unbounded sets that is only slightly different than the confidence set with the outliers.

Remark When we compute the confidence set with the outliers in the data the MM-estimator seems to get stuck in a local optimum for a few β_0 values close to the right boundary of the left open unbounded set $(-\infty, -0.14)$. We discovered this by plotting the p-values for each β_0 value used to invert the test statistic. For this reason, we also present a "weighted" robust confidence set, where we downweight the sixth observation, corresponding to the large outlier in the control variable, to zero. In this case, there does not seem to be a numerical problem and we see that the confidence set is only slightly different. We also used the weights based on the leverage matrix introduced in Section 3, but unfortunately these weights do not downweight the outlier enough to not get stuck in a local optimum for some β_0 values.

5.2 The Causal Effect of Education on Industrialization in Prussia

The second paper we revisit is the paper by Becker et al. (2011). They study the effect of education on industrialization in Prussia. As industrialization might cause changes in the

Table 1: In this table we present the results of the F-test, AR test and robust AR test based on Ananat (2011).

	With outliers	Without outliers
Results from Ananat (2011):		
First Stage F-test	16.45	-
95% conf. set	(-0.53, -0.14)	-
Our results assuming homoskedasticity:		
First Stage F-test	19.32	1.83
95% AR conf. set	(-0.67, -0.15)	$(-\infty, -0.19) \cup (1.30, \infty)$
95% robust AR conf. set	$(-\infty, -0.14) \cup (3.56, \infty)$	$(-\infty, -0.20) \cup (2.43, \infty)$
95% weighted robust AR conf. set	$(-\infty, -0.26) \cup (3.58, \infty)$	-

demand for education an endogeneity problem arises. To account for this problem education before the industrialization is used as an instrument.

More specifically, the first stage explains the endogeneous variable years of schooling in 1849 (y_1) by the school enrollment rate in 1816 (x_2) and several control variables ($\mathbf{x}_1$). The second stage explains the share of factory workers (for all factories) in the total population in 1849 (y_2) by the endogeneous variable years of schooling in 1849 (y_1) and three control variables ($\mathbf{x}_1$) that control for the share of population younger than 15, older than 65 and the size of the county area.

We start with an exploratory analysis of the first and second stage variables in Figure 8, where the three control variables are partialled out. At a first glance, there seems to be a strong linear relationship between the endogeneous variable y_1 and the instrument x_2, which suggests the instrument is strong. When we analyze the relationship between the variable y_2 and the variable y_1 we see that there are some outliers present in the dependent variable y_2.

In Table 2 we show the first stage F-statistic and the β confidence set used in the paper. The first stage F-statistic is 6207, implying that the instrument is strong. Therefore, β is estimated using a 2SLS estimator and inference is done with a t-test. The result is $\hat{\beta} = 0.132$,

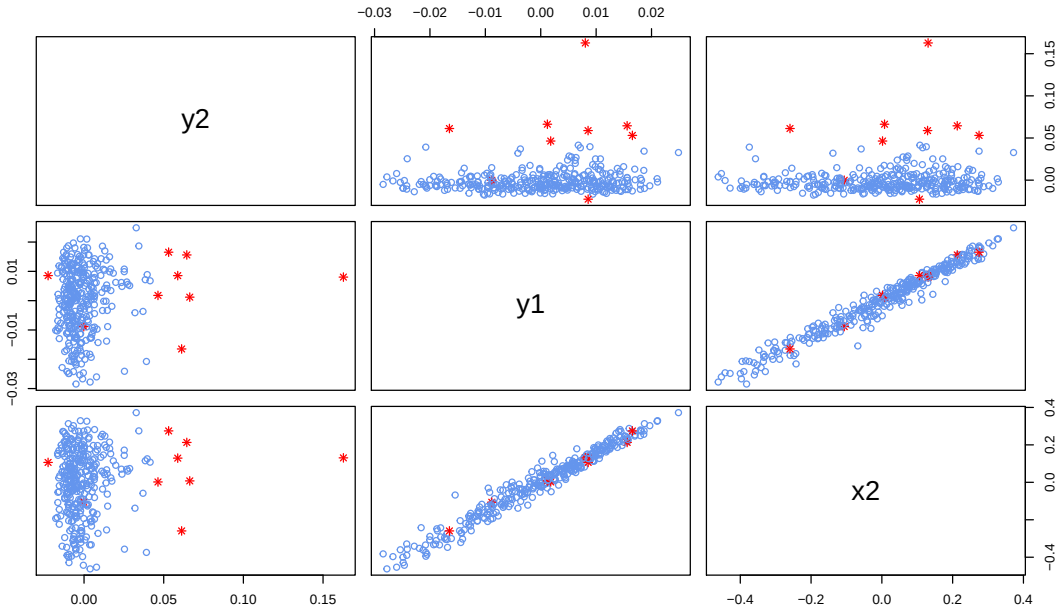

Figure 8: Scatter plot of the first and second stage variables in Becker et al. (2011) with the control variables partialled out. The red stars are flagged as outliers by the BACON algorithm.

with a standard error of 0.077, implying that it is significant at the 10% level. The 90% confidence set is $(0.01, 0.26)$. To obtain these results, clustered standard errors were used. The results can be found in columns (5) (first stage) and (6) (second stage) of Table 1 in Becker et al. (2011).

Next, we calculate the first stage F-statistic and the AR confidence set with and without the outliers. When the outliers are in the data the first stage F-statistic is 8379. Due to our homoskedasticity assumption, the first stage F-statistic is larger than the first stage F-statistic in the paper, but still leads to the same conclusion that the instrument is strong. When we remove the outliers from the data, the first stage F-statistic is 8039, which is very similar to the F-statistic when the outliers are in the data. This is not surprising, as the outliers seem to be most striking in the dependent variable y_2. The 90% AR confidence set is $(0.01, 0.27)$, which is very similar to the confidence set based on the t-test. When we remove the outliers from the data, the 90% AR confidence set is $(-0.01, 0.17)$. Based on this confidence set, we are unable to reject the hypothesis $\beta = 0$. Recall from the simulation study in Section 4 that

a large outlier in the y_2 variable can bias the OLS estimate of the x_2 parameter used in the AR test when analyzing the relationship between y_2 and x_2.

Finally, we calculate the 90% confidence set of the robust AR test. As the outliers are most notable in the dependent variable, we do not use the weight function. The robust confidence set is given by $(-0.02, 0.08)$, which is different than the AR confidence set and the t-test confidence set. When we remove the outliers from the data, the robust confidence set doesn't change. Without the outliers, the confidence sets of the robust AR test and the AR test are still different. This happens, because we use a strong cut-off value to remove the outliers from the data. When we analyze Figure 8 we see that if we remove the outliers and make a new scatter plot, we would see some "new" outliers. This is the masking effect problem, that we mentioned above. The robust method also downweights these points, albeit less so than the larger outliers.

In this example, we see that the classical methods suggest that we can reject the hypothesis that $\beta = 0$. However, the robust AR test does not lead to the same conclusion. It is likely that the significance in this regression is mainly driven by a few outliers in the data.

Table 2: In this table we present the results of the F-test, AR test and robust AR test based on Becker et al. (2011).

	With outliers	Without outliers
Results from Becker et al. (2011):		
First stage F-test	6207	-
90% conf. set:	(0.01, 0.26)	-
Our results assuming homoskedasticity:		
First Stage F-test	8379	8039
90% AR conf. set:	$(0.01, 0.27)$	$(-0.01, 0.17)$
90% robust AR conf. set:	$(-0.02, 0.08)$	$(-0.02, 0.08)$

5.3 The Causal Effect of State Fiscal Relief on Employment

The third paper we revisit is the paper by Chodorow-Reich et al. (2012). They study the effect of state fiscal relief on employment in the United States. As state fiscal relief outlays are endogeneous to a state's economic environment an instrumental variable is used. In this case, the instrument is a state's prerecession Medicaid spending.

More specifically, the first stage explains the endogeneous variable Federal Medical Assistance Percentages (FMAP) (y_1) by a state's Medicaid spending in 2007 (x_2). The second stage explains the seasonally adjusted change in total employment in state and local government, health, and education per individual 16+ in a state, from December 2008 to July 2009 (y_2) by the endogeneous variable FMAP. In this baseline specification no controls were used.

We begin with an exploratory analysis of the first and second stage variables in Figure 9. At a first glance, there seems to be a positive linear relationship between the endogeneous variable y_1 and the instrument x_2, which suggests the instrument is strong. We do notice there might be some "good leverage points" in the data (Dehon et al., 2009). The two good leverage points in the first stage do not have a negative effect on the slope estimate in the first stage. However, these good leverage points can become problematic in the second stage and should be analyzed with care.

In Table 3 we show the first stage F-statistic and the β confidence set used in the paper. Chodorow-Reich et al. (2012) mention that the first stage F-statistic is above 260, implying that the instrument is strong. Therefore, β is estimated using a 2SLS estimator and inference is done with a t-test. The result is $\hat{\beta} = 0.99$ with a (heteroskedasticity-robust) standard error of 0.54, implying that the estimate is significant at the 10% level. The 90% confidence set is $(0.08, 1.90)$. This result can be found in column (4) of Table 4 in Chodorow-Reich et al. (2012).

Next, we assume homoskedasticity and calculate the first stage F-statistic and the AR confidence set with and without the outlier. When the outlier is in the data the first stage F-statistic is 261, which implies that the instrument is strong. When we remove the outlier

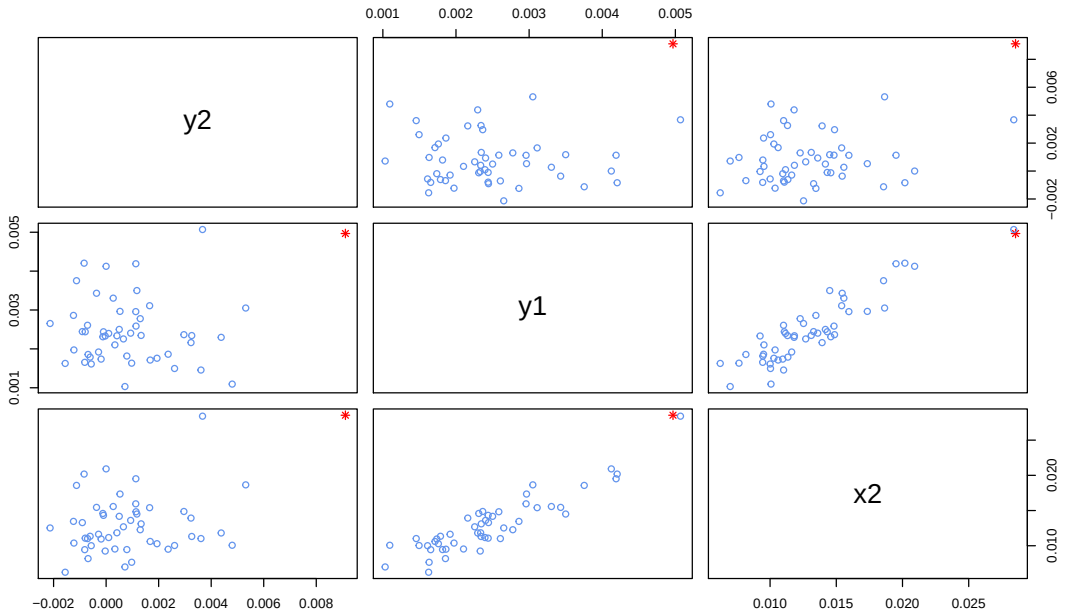

Figure 9: Scatter plot of the first and second stage variables in Chodorow-Reich et al. (2012). The red star is flagged as an outlier by the BACON algorithm.

from the data, the first stage F-statistic is 215, which leads to the same conclusion that the instrument is strong. The 90% AR confidence set is $(0.44, 1.62)$, which is a bit smaller than the confidence set reported in the paper. This difference occurs because of the homoskedasticity assumption we make. When we remove the outlier from the data, the 90% confidence set is $(-0.13, 1.00)$. Based on this confidence set we cannot reject the hypothesis $H_0\colon \beta = 0$. This happens because when we test this null hypothesis with the AR test, we test whether there is a nonzero linear relationship between y_2 and x_2. When we study Figure 9, we see that one of the good leverage points in the first stage is now a bad leverage point. At a first sight, when we analyze the data without the outlier, there does not seem to be a positive linear relationship between y_2 and x_2. However, when we do not take care of the outlier, it causes a positive bias in the OLS estimate so that the AR test rejects the hypothesis that $\beta = 0$.

Finally, we calculate the 90% confidence set of the robust AR test. As the outlier is in all dimensions, we do use the weights based on the leverage matrix. The robust confidence set is $(-0.38, 1.87)$, which is wider than the AR confidence set. The most striking difference is that

zero is in the confidence set of the robust AR test. When we remove the outlier from the data, the robust confidence set is $(-0.38, 0.98)$. Based on the confidence set of the robust AR test, we are not able to reject the hypothesis that $\beta = 0$. Again, it is likely that the significance of the result is mainly driven by one outlier in the data.

Table 3: In this table we present the results of the F-test, AR test and robust AR test based on Chodorow-Reich et al. (2012).

	With outlier	Without outlier
Results from Chodorow-Reich et al. (2012):		
First stage F-test	260	-
90% conf. set:	(0.08, 1.90)	-
Our results assuming homoskedasticity:		
First Stage F-test	261	215
90% AR conf. set:	(0.42,1.64)	(-0.13, 1.00)
90% robust AR conf. set:	(-0.38, 1.87)	(-0.38, 0.98)

6 Conclusion

In this paper we showed how to obtain a robust alternative to the Anderson-Rubin test - the robust AR test. The robust AR test allows for reliable inference in the instrumental variable model in the presence of small but harmful deviations from the model's assumptions. In particular, it (also) allows for reliable inference when the instruments are weak. We studied the robustness properties to small data contamination of the robust AR test in two ways. We started by showing that the influence function of the robust AR test is bounded. Then, by means of a thorough simulation study we showed that the robust AR test is more reliable than the classical AR test in different contaminated settings.

The robust AR test can be helpful for applied researchers in several ways. First, it can serve as an extra robustness check. Both the classical AR test and the robust AR test are asymptotically χ^2-distributed and must be close to each other if the model F (our model (1)-(2)) holds exactly. In the case studies, the confidence set of the robust AR set was sometimes shifted, or a lot wider, compared to the confidence set of the AR test or t-test. This is an indication that the classical distributional assumptions do not hold and cannot be trusted. On the other hand, if the confidence set of the robust AR test is very similar to the confidence set of the classical AR test, it indicates the validity of the distributional assumptions. Second, if the researcher has some doubts about the modeling assumptions, for example, due to valid but outlying observations in the data, then, the robust AR test can be used as a stand-alone method.

References

Acconcia, A., Corsetti, G., and Simonelli, S. (2014), "Mafia and Public Spending: Evidence on the Fiscal Multiplier from a Quasi-Experiment," *American Economic Review*, 104, 2185–2209.

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," *The 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, 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.

Becker, S. O., Hornung, E., and Woessmann, L. (2011), "Education and Catch-up in the Industrial Revolution," *American Economic Journal: Macroeconomics*, 3, 92–126.

Billor, N., Hadi, A. S., and Velleman, P. F. (2000), "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators," *Computational Statistics & Data Analysis*, 34, 279–298.

Bollen, K. A. (2012), "Instrumental Variables in Sociology and the Social Sciences," *Annual Review of Sociology*, 38, 37–72.

Bonhomme, S. and Weidner, M. (2021), "Minimizing Sensitivity to Model Misspecification," *Technical Report*.

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.

Chen, S. and Bien, J. (2020), "Valid Inference Corrected for Outlier Removal," *Journal of Computational and Graphical Statistics*, 29, 323–334.

Chodorow-Reich, G., Feiveson, L., Liscow, Z., and Woolston, W. G. (2012), "Does State Fiscal Relief During Recessions Increase Employment? Evidence From The American Recovery and Reinvestment Act," *American Economic Journal: Economic Policy*, 4, 118–45.

Copt, S. and Heritier, S. (2007), "Robust Alternatives to the F-test in Mixed Linear Models Based on MM-Estimates," *Biometrics*, 63, 1045–1052.

Dehon, C., Gassner, M., and Verardi, V. (2009), "Beware of 'Good' Outliers and Overoptimistic Conclusions," *Oxford Bulletin of Economics and Statistics*, 71, 437–452.

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.

Frisch, R. and Waugh, F. V. (1933), "Partial Time Regressions as Compared With Individual Trends," *Econometrica*, 1, 387–401.

Greene, W. H. (2012), *Econometric Analysis*, Upper Saddle River, NJ: Prentice–Hall, 7th ed.

Greenland, S. (2000), "An Introduction to Instrumental Variables for Epidemiologists," *International journal of epidemiology*, 29, 722–729.

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," *The 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. (2021), "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, to Appear.

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

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.

Lovell, M. C. (1963), "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58, 993–1010.

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. (2021), *Robustbase: Basic Robust Statistics*, R package available at *http://robustbase.r-forge.r-project.org/*.

Markatou, M. and Hettmansperger, T. P. (1990), "Robust Bounded-Influence Tests in Linear Models," *Journal of the American Statistical Association*, 85, 187–190.

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.

Mikusheva, A. (2010), "Robust Confidence Sets in the Presence of Weak Instruments," *Journal of Econometrics*, 157, 236–247.

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

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.

Rousseeuw, P. J. and Ronchetti, E. (1981), "Influence Curves of General Statistics," *Journal of Computational and Applied Mathematics*, 7, 161–166.

Sovey, A. J. and Green, D. P. (2011), "Instrumental Variables Estimation in Political Science: A Readers' Guide," *American Journal of Political Science*, 55, 188–200.

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

Stephens Jr, M. and Yang, D. Y. (2014), "Compulsory Education and the Benefits of Schooling," *American Economic Review*, 104, 1777–92.

Ugarte Ontiveros, D. and Verardi, V. (2012), "Supposedly Strong Instruments and Good Leverage Points," Tech. rep., University of Namur, Department of Economics.

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

Young, A. (2021), "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application," *Technical Report*.

Supplement to "Outlier Robust Inference in the Instrumental Variable Model With Applications to Causal Effects"

Jens Klooster*, Mikhail Zhelonkin†

October 1, 2021

Supplementary Material

Regularity Conditions and Assumptions

When we use the robust Anderson-Rubin test, we first estimate the parameters γ_2 and σ_2 in the null-restricted model. We use robust M-estimators of regression and scale. Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, let $\mathbf{1}(\cdot)$ denote the indicator function and let $||\cdot||_2$ denote the Euclidean norm. We consider solutions (γ_2, σ_2) of the system

$$\int \phi\left(\mathbf{x}, (y_2 - y_1\beta_0 - \mathbf{x}_1^\top\gamma_2)/\sigma_2\right)\mathbf{x}_1 dF = \mathbf{0}, \tag{1}$$

$$\int \chi(|y_2 - y_1\beta_0 - \mathbf{x}_1^\top\gamma_2|/\sigma_2)dF = 0, \tag{2}$$

where $\phi\colon \mathbb{R}^{p_1+p_2} \times \mathbb{R} \to \mathbb{R}$, and $\chi\colon \mathbb{R} \to \mathbb{R}$. We assume that the following assumptions hold (taken from Maronna and Yohai 1981).

1. For each $\mathbf{x}$, $\phi(\mathbf{x}, \cdot)$ is odd and uniformly continuous, and $\phi(\mathbf{x}, u) \geq 0$ for $u \geq 0$.

2. The function $\mu(\mathbf{x}, u) = \phi(\mathbf{x}, u)/u$ is nonincreasing for $u > 0$, and there exists $u_0 > 0$ such that $\mu(\mathbf{x}, u_0) > 0$ for all $\mathbf{x}$.

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

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

3. χ is nondecreasing, continuous, and bounded. Let $\chi(0) = -a, \chi(\infty) = b$, with $a, b \in (0, \infty)$.

4. χ is strictly increasing in the interval $\{u \mid \chi(u) < b\}$.

5. $\int ||\mathbf{x}||_2 \sup_u |\phi(\mathbf{x}, u)| dF < \infty$.

6. $\sup \left\{ \int \mathbf{1}(\mathbf{x}_1^\top \gamma_2 = 0) dF \mid \gamma_2 \neq \mathbf{0} \right\} < \chi(u_0)/(a + \chi(u_0))$, with u_0 defined in 2.

7. $\sup \left\{ \int \mathbf{1}(\alpha(y_2 - y_1\beta_0) + \mathbf{x}_1^\top \gamma_2 = 0) dF \mid \alpha \in \mathbb{R}, \gamma_2 \in \mathbb{R}^{p_1}, |\alpha| + ||\gamma_2||_2 \neq 0 \right\} < b/(a+b)$, with a, b, as in 3.

As shown in Theorem 2.1 of Maronna and Yohai (1981), these assumptions make sure there exists a solution to the estimating equations (1)-(2). In this paper, we use $\phi(\mathbf{x}, r) = \omega(\mathbf{x})\Psi(r; c)$ or, when we do not use weights in the test, $\phi(\mathbf{x}, r) = \Psi(r; c)$, where Ψ denotes Tukey's Biweight function. We use $\chi(r) = \rho(r; c)$, where ρ is equal to Tukey's loss function defined as

$$\rho(r; c) := \begin{cases} \frac{c^2}{6} \left\{ 1 - \left(1 - \frac{r^2}{c^2} \right) \right\}^3, & \text{for } |r| \leq c, \\ \frac{c^2}{6}, & \text{for } |r| > c. \end{cases}$$

To ensure asymptotic normality of the estimators and test statistic, we further assume that

a) $\int ||\mathbf{x}||_2^3 dK < \infty$.

b) The matrix

$$\mathbf{M} = \int \frac{\omega(\mathbf{x}_1, \mathbf{x}_2)}{\sigma_2} \frac{\partial \Psi}{\partial r} \left(\frac{y_2 - \beta_0 y_1 - \mathbf{x}_1^\top \gamma_2}{\sigma_2}; c \right) \begin{pmatrix} \mathbf{x}_1 \mathbf{x}_1^\top & \mathbf{x}_1 \mathbf{x}_2^\top \\ \mathbf{x}_2 \mathbf{x}_1^\top & \mathbf{x}_2 \mathbf{x}_2^\top \end{pmatrix} dF$$

is nonsingular.

These last two assumptions make sure that assumptions $(C1) - (C6)$ introduced in Theorem 4.1 of Maronna and Yohai (1981) hold in our specific setting.

Simulations Supplement

We consider four different settings: one setting without contamination and three settings with contamination. We consider two settings with contamination by outliers and one setting with a distributional contamination in the error terms.

First, we generate uncontaminated data. Let $\mathbb{1} = (1, 1, 1, 1, 1)$. We simulate data from the following model:

$$y_1 = x_1 + \pi \mathbf{x}_2^\top \mathbb{1} + \epsilon_1, \tag{3}$$

$$y_2 = \beta y_1 + 2x_1 + \epsilon_2. \tag{4}$$

We generate five exogeneous instruments $\mathbf{x}_2 = (x_{21}, x_{22}, x_{23}, x_{24}, x_{25})$ and one exogeneous control variable x_1. We draw each $x_{2j}, j = 1, \ldots, 5$, and x_1 from a standard normal distribution. The error terms ϵ_1 and ϵ_2 follow a bivariate normal distribution with variances equal to 1 and correlation $\rho = 0.90$. We consider two different cases for the parameter $\pi \in \{0.1, 1\}$. When $\pi = 0.1$ we are in the weak instrument setting and when $\pi = 1$ the instrument is strong. The sample size is $N = 250$ and we repeat the study 10000 times. In the simulation, we test $H_0 \colon \beta = 0$ at the 5% significance level.

In the first contamination, we simulate a setting with an outlier in the endogeneous variable y_2 by generating data from the uncontaminated model and then replacing the first data row by $(y_{21}, y_{11}, x_{11}, x_{211}, x_{221}, x_{231}, x_{241}, x_{251}) = (10, 2, 2, 3, 3, 3, 3, 3)$. In the second contamination, we simulate a setting with an outlier in the exogeneous control variable x_1 by generating data from the uncontaminated model and then replacing the first data row by $(y_{21}, y_{11}, x_{11}, x_{211}, x_{221}, x_{231}, x_{241}, x_{251}) = (2, 2, 10, 2, 2, 2, 2, 2)$. Finally, we simulate a third setting with a distributional contamination. We simulate the errors from a mixture of a bivariate normal (as above) and a bivariate t-distribution, i.e., with 10% probability we simulate the error terms from a bivariate t-distribution with 2 degrees of freedom.

Simulation Results of Extension

In Figures 1, 2, 3 and 4 we present the results of the extended simulation study. As the results are very similar to the results in the one instrument setting, we only comment on the key differences.

In Figure 1 we see that in the strong instrument setting, the t-test based on the 2SLS estimator is strictly more powerful than the AR and robust AR test. This happens because the AR and robust AR test lose power when more instruments are used. However, the robust AR test remains useful as we can see in Figures 2 and 3. Here we see that the different outliers completely break down the t-test and AR test, while the robust AR test remains size correct and almost as powerful as in the uncontaminated setting. At last, in Figure 4, we see that in a setting with contamination in the error term the robust AR test is strictly more powerful than the AR test in both weak and strong instrument settings. However, in contrast to the one instrument setting, when the instrument is strong, the t-test based on the 2SLS estimator remains more powerful than the robust AR test. This happens because of the power loss of the robust AR test compared to the t-test due to the extra instruments.

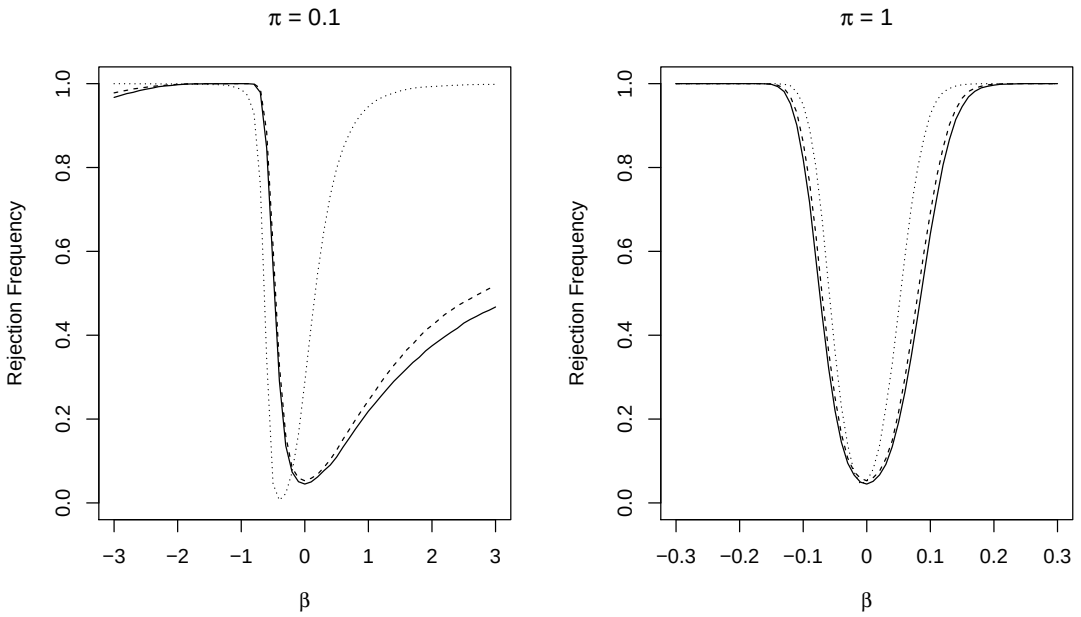

Figure 1: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with five instruments, $\rho = 0.90$ and no contamination.

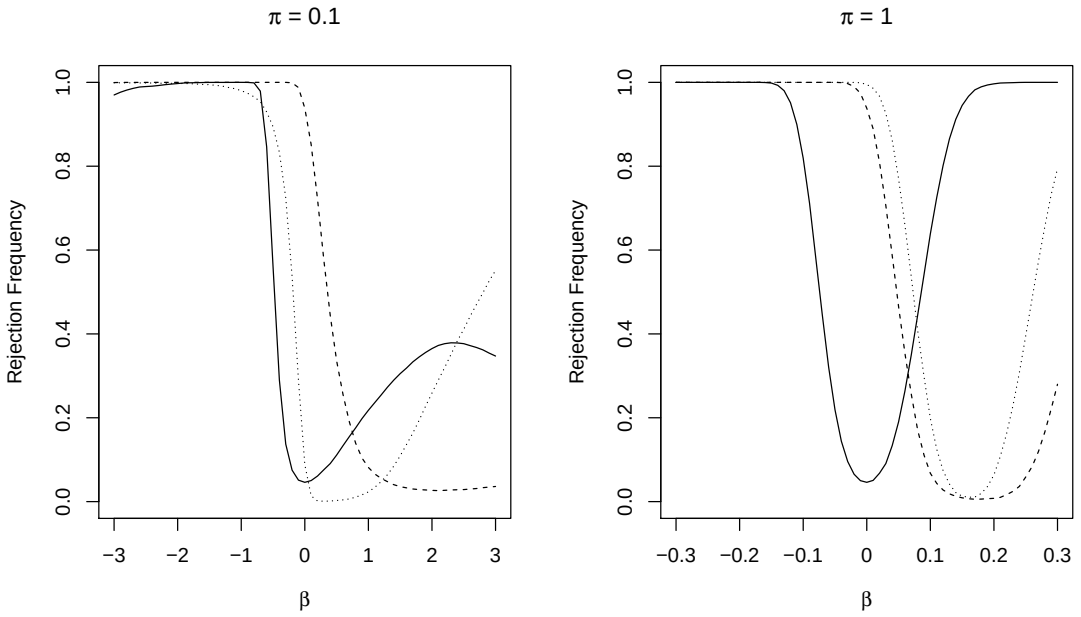

Figure 2: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with five instruments, $\rho = 0.90$ and an outlier in the endogeneous variable.

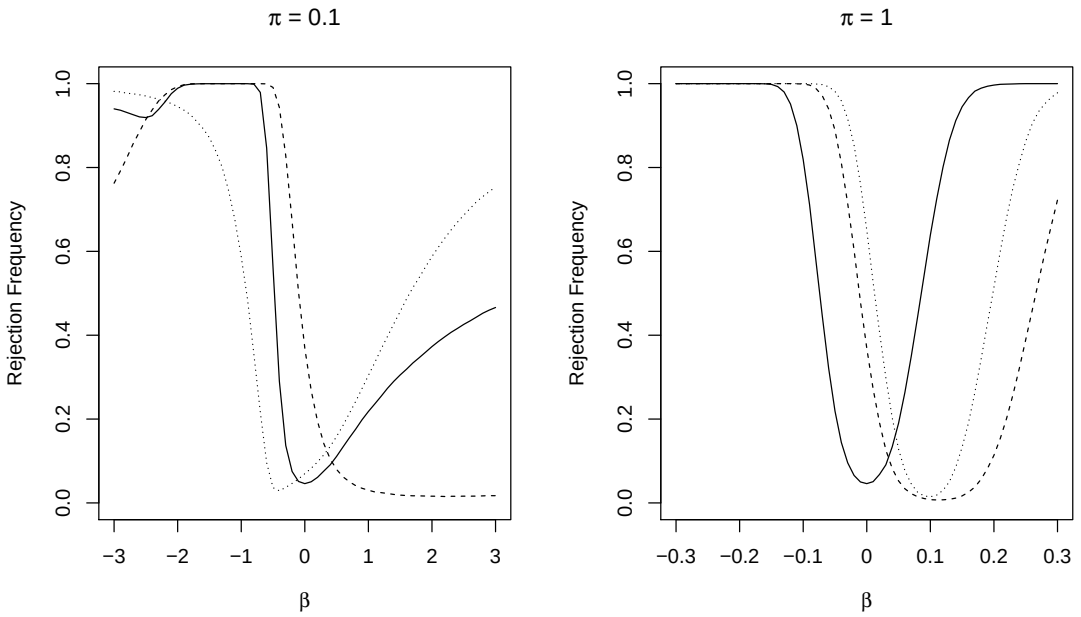

Figure 3: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with five instruments, $\rho = 0.90$ and an outlier in the exogeneous control variable.

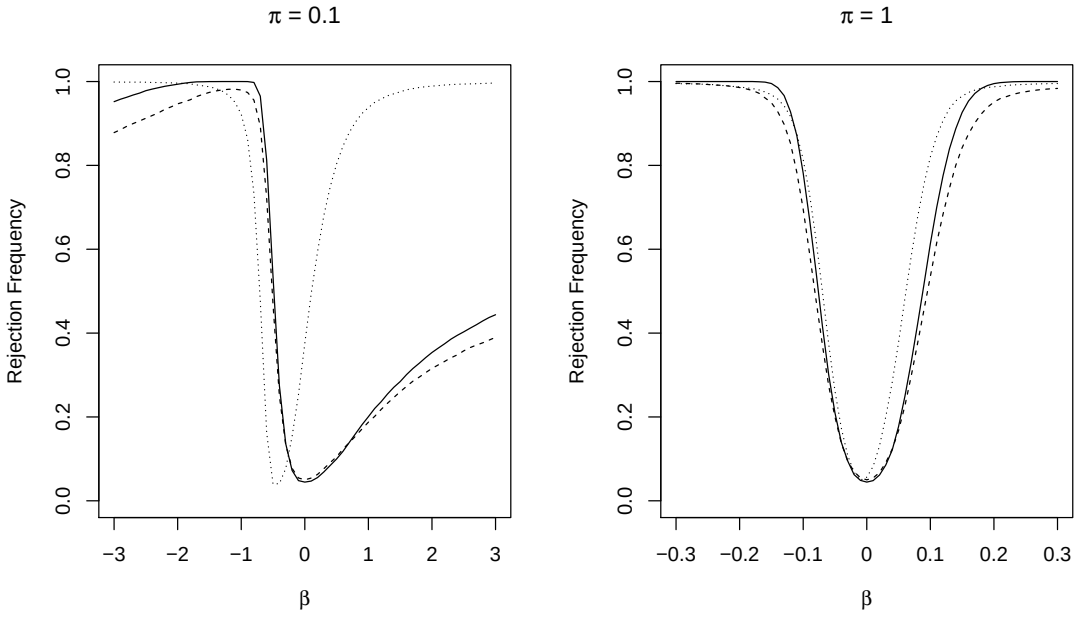

Figure 4: Power curves of robust AR (solid), AR (dashed) and t-test (dotted) that tests $H_0\colon \beta = 0$ for various values of β. Setting with five instruments, $\rho = 0.90$ and contamination in the error term.

References

Maronna, R. A. and Yohai, V. J. (1981), "Asymptotic Behavior of General M-Estimates for Regression and Scale With Random Carriers," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 58, 7–20.