Weak-Instrument Robust Tests in Two-Sample Summary-Data Mendelian Randomization

Sheng Wang* and Hyunseung Kang†

Department of Statistics, University of Wisconsin-Madison

Abstract

Mendelian randomization (MR) is a popular method in genetic epidemiology to estimate the effect of an exposure on an outcome using genetic variants as instrumental variables (IV), with two-sample summary-data MR being the most popular due to privacy. Unfortunately, many MR methods for two-sample summary data are not robust to weak instruments, a common phenomena with genetic instruments; many of these methods are biased and no existing MR method has Type I error control under weak instruments. In this work, we propose test statistics that are robust to weak instruments by extending Anderson-Rubin, Kleibergen, and conditional likelihood ratio tests in econometrics to the two-sample summary data setting. We conclude with a simulation and an empirical study and show that the proposed tests control size and have better power than current methods.

Key words: Instrumental variables; Mendelian randomization; two-sample summary-data Mendelian randomization; weak instrument asymptotics.

1 Introduction

Recently, Mendelian randomization (MR) is a popular method in genetic epidemiology to study the effect of modifiable exposures on health outcomes by using genetic variants as instrumental variables (IV). In a nutshell, MR finds instruments, typically single nucleotide polymorphisms (SNPs), from publicly available genome-wide association studies (GWAS) and the instruments must be (A1) associated with the exposure; (A2) independent of the unmeasured confounder; and (A3)

*shengw@stat.wisc.edu
†hyunseung@stat.wisc.edu

independent of the outcome variable after conditioning on the exposure (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008) Typically, two non-overlapping GWAS are used to find instruments, one GWAS studying the exposure and another GWAS studying the outcome. Also, due to privacy, when estimating the exposure effect, only summary statistics instead of individual-level data are extracted for analysis. This setup is commonly known as two-sample summary-data MR (Pierce and Burgess, 2013; Burgess et al., 2013, 2015).

The focus of this paper is on testing the exposure effect in two-sample summary-data MR when (A1) is violated, or more precisely when the instruments are weakly associated with the exposure. Many genetic instruments in MR studies only explain a fraction of the variation in the exposure. Using these weak instruments can introduce bias and inflate Type I errors (Burgess and Thompson, 2011). Weak IVs can also amplify bias from minor violations of (A2) and (A3) (Small and Rosenbaum, 2008). Unfortunately, many popular MR methods and software assume instruments are strongly associated with the exposure; many go one step further and assume that the correlation between each instrument and exposure has no sampling error (Bowden et al., 2016b). For example, methods such as the inverse-variance weighted estimator (IVW) (Burgess et al., 2013), MR-Egger regression (Bowden et al., 2015), weighted median estimator (W.Median) (Bowden et al., 2016a) and the modal estimator (Hartwig et al., 2017) typically assume that each instrument's correlation to the exposure of interest is measured without error. In Section 4, we numerically demonstrate the seriousness of making such assumptions in MR by "stress-testing" these methods' performance on a real MR data set, akin to an exercise done by Bound et al. (1995) in econometrics for single-sample, individual-level data.

Many works in econometrics have dealt with weak instruments; see Stock et al. (2002) for an overview. In particular, the Anderson-Rubin (AR) test (Anderson et al., 1949), the Kleibergen (K) test (Kleibergen, 2002), and the conditional likelihood ratio (CLR) test (Moreira, 2003) provide Type I error control regardless of instruments' magnitude of association to the exposure, also called instruments' strength. More formally, the three methods satisfy the necessary requirement for valid $1-\alpha$ confidence intervals with weak instruments, where the confidence interval adapts and becomes infinite in the presence of weak IVs to maintain $1-\alpha$ coverage (Dufour, 1997). However, all these methods assume that individual-level data is available to compute the test statistics and the data comes from the same sample. In GWAS and MR, one rarely has access to individual-level data due to privacy concerns and is forced to work with anonymized summary statistics from multiple GWAS. To the best of our knowledge, no methods in two-sample summary-data MR provide Type I error control when the relationship between the exposure and the instruments is weak, even irrelevant.

Our contribution is to propose weak instrument robust tests for two-sample summary-data MR. We extend the three aforementioned weak-instrument robust tests in econometrics, AR test, the

K test , and the CLR test, to work with two-sample summary data by leveraging recent work by Choi et al. (2018) who worked with two-sample, but individual data. We show that under the two-sample summary-data model and weak-IV asymptotics of Staiger and Stock (1997), these modified tests, which we call mrAR, mrK, and mrCLR, asymptotically control Type I error for testing the exposure effect. In the supplementary materials, we also introduce point estimators based on these tests, most notably the MR limited information maximum likelihood estimator (mrLIML) based on taking the minimum of the mrAR test. mrLIML is similar to the original limited information maximum likelihood (LIML) estimator (Anderson et al., 1949) and we show an equivalence relationship between mrLIML and the recent profile-likelihood estimator proposed by Zhao et al. (2018). We conclude with a simulation and a prototypical MR data example concerning the effect of body mass index (BMI) on systolic blood pressure (SBP).

2 Setup and Method

2.1 Review: Two-Sample Summary Data in MR

We review the data generating model underlying MR. Suppose we have two independent groups of people, with n_1 and n_2 participants each of the two groups. For each individual i in sample $l = 1, 2$, let $Y_{li} \in \mathbb{R}$ denote his/her outcome, $D_{li} \in \mathbb{R}$ denote his/her exposure, and $Z_{li} \in \mathbb{R}^L$ denote his/her L instruments. Single-sample individual data MR assumes that for one sample l, Y_{li}, D_{li}, Z_{li} follows a linear structural model common in classical econometrics (Lawlor et al., 2008).

$$Y_{li} = \beta_{\text{int}} + D_{li}\beta + \epsilon_{li}, \quad D_{li} = \gamma_{\text{int}} + Z_{li}^{\mathsf{T}}\gamma + \delta_{li}, \quad E[\epsilon_{li}, \delta_{li} \mid Z_{li}] = 0 \tag{1}$$

The parameter of interest is β and has a causal interpretation under some assumptions (Holland, 1988; Kang et al., 2016; Zhao et al., 2019). Two-sample individual-data MR assumes the same underlying structural model (1) for both samples. But, for sample $l = 1$, the investigator only sees (Y_{1i}, Z_{1i}) and for sample $l = 2$, the investigator only sees (D_{2i}, Z_{2i}) (Pierce and Burgess, 2013; Burgess et al., 2013); this is identical to the setup in Angrist and Krueger (1992). Finally, in two-sample summary-data MR, only summarized statistics of (Y_{1i}, Z_{1i}) and (D_{2i}, Z_{2i}) are available. Specifically, from n_1 samples of (Y_{1i}, Z_{1i}), we obtain (i) $\widehat{\Gamma} \in \mathbb{R}^L$ where $\widehat{\Gamma}_j$ is the estimated association between IV Z_{1ij} and Y_{1i} and (ii) $\widehat{\Sigma}_\Gamma \in \mathbb{R}^{L\times L}$, the estimated covariance of $\widehat{\Gamma}$. Similarly, from n_2 samples of (D_{2i}, Z_{2i}), we obtain (i) $\widehat{\gamma} \in \mathbb{R}^L$ where $\widehat{\gamma}_j$ is the estimated association between IV Z_{2ij} and D_{2i} and (ii) $\widehat{\Sigma}_\gamma \in \mathbb{R}^{L\times L}$, the estimated covariance of $\widehat{\gamma}$. We assume that the summary statistics $(\widehat{\Gamma}, \widehat{\Sigma}_\Gamma, \widehat{\gamma}, \widehat{\Sigma}_\gamma)$ used in the analysis satisfy the following assumptions:

Assumption 1. The IV-exposure effect $\widehat{\gamma}$ and the IV-outcome effect $\widehat{\Gamma}$ are independent, $\widehat{\gamma} \perp \widehat{\Gamma}$.

Assumption 2. The two effect estimates are distributed as $\widehat{\gamma} \sim N(\gamma, \Sigma_\gamma)$ and $\widehat{\Gamma} \sim N(\gamma\beta, \Sigma_\Gamma)$

Assumption 3. We have $n_1(\widehat{\Sigma}_\Gamma - \Sigma_\Gamma) \xrightarrow{p} 0$, $n_2(\widehat{\Sigma}_\gamma - \Sigma_\gamma) \xrightarrow{p} 0$ and $n_1\Sigma_\Gamma \xrightarrow{p} \Sigma_1, n_2\Sigma_\gamma \xrightarrow{p} \Sigma_2$, where Σ_1, Σ_2 are deterministic positive-definite matrices.

Assumption 4. For some constant $C \in \mathbb{R}^L$, we have $\gamma = C/\sqrt{n_1}$

Assumption 1 is typically satisfied in MR by having two GWAS that independently measure SNPs' associations to the exposure and outcome (Pierce and Burgess, 2013). Assumption 2 is reasonable in publicly-available GWAS where the effect estimates are based on running ordinary least square (OLS) between each instrument and exposure/outcome and since the sample size for each GWAS is on the order of hundreds or thousands, the normality of OLS estimates is plausible. Assumption 3 states that the estimated standard errors converge to their asymptotic variances. Assumption 3 is plausible since the covariance matrices are estimated from OLS residual errors and most of MR assumes that the SNPs are independent of each other. Overall, Assumptions 1-3 are common in two-sample summary-data MR (Zhao et al., 2018). Finally, Assumption 4 is also known as weak-IV asymptotics (Staiger and Stock, 1997) and it provides an asymptotic framework to study the behavior of IV estimators when instruments are weak.

We remark that the literature also assume the instruments are independent to each other, which we do not explicitly impose here. Also, to focus on our contributions to weak IVs in two-sample summary-data MR, we assume that the instruments are valid (e.g. the exclusion restriction holds).

2.2 Weak-IV Robust Tests for the Exposure Effect β

Consider the null hypothesis $H_0 : \beta = \beta_0$ and the alternative $H_a : \beta \neq \beta_0$. We define two statistics $S(\beta_0) \in \mathbb{R}^L$ and $R(\beta_0) \in \mathbb{R}^L$ from the summary statistics $(\widehat{\Gamma}, \widehat{\Sigma}_\Gamma, \widehat{\gamma}, \widehat{\Sigma}_\gamma)$.

$$S(\beta_0) = \left(\widehat{\Sigma}_\Gamma + \beta_0^2\widehat{\Sigma}_\gamma\right)^{-1/2}\left(\widehat{\Gamma} - \beta_0\widehat{\gamma}\right), \quad R(\beta_0) = (\beta_0^2\widehat{\Sigma}_\Gamma^{-1} + \widehat{\Sigma}_\gamma^{-1})^{-1/2}\left(\widehat{\Sigma}_\Gamma^{-1}\widehat{\Gamma}\beta_0 + \widehat{\Sigma}_\gamma^{-1}\hat{\gamma}\right) \tag{2}$$

The statistics $S(\beta_0)$ and $R(\beta_0)$ are similar to the independent sufficient statistics of β and π in the traditional econometric setting (i.e. one-sample individual-data setting) (Moreira, 2003) or in the two-sample individual data setting (Choi et al., 2018). A key difference is that (2) are computed with two-sample summary data. While not exactly sufficient for β and π in our setting, $S(\beta_0)$ and $R(\beta_0)$ is asymptotically independent as the following Lemma shows.

Lemma 1. If Assumptions 1-4 hold and $n_1/n_2 \to c \in (0, \infty)$, $(S(\beta_0), R(\beta_0)) \xrightarrow{p} (S_\infty(\beta_0), R_\infty(\beta_0))$ where

$$S_\infty \sim N[(\Sigma_1 + c\beta_0^2\Sigma_2)^{-1/2}(\beta - \beta_0)C, I_L]$$
$$R_\infty \sim N[(\beta_0^2\Sigma_1^{-1} + c^{-1}\Sigma_2^{-1})^{-1/2}(\beta_0\beta\Sigma_1^{-1}C + c^{-1}\Sigma_2^{-1}C), I_L]$$

and S_∞ and R_∞ are independent.

The asymptotic independence is crucial as it allows us to follow Moreira (2003) and Andrews et al. (2006) and use $S(\beta_0)$ and $R(\beta_0)$ to construct AR, K, and CLR tests for two-sample summary-data MR.

$$T_{\text{mrAR}}(\beta_0) = Q_S(\beta_0) \tag{3}$$

$$T_{\text{mrK}}(\beta_0) = Q_{SR}^2(\beta_0)/Q_R(\beta_0) \tag{4}$$

$$T_{\text{mrCLR}}(\beta_0) = \frac{1}{2}\left(Q_S(\beta_0) - Q_R(\beta_0) + \left[\{Q_S(\beta_0) + Q_R(\beta_0)\}^2 - 4\left\{Q_S(\beta_0)Q_R(\beta_0) - Q_{SR}^2(\beta_0)\right\}\right]^{\frac{1}{2}}\right) \tag{5}$$

Here, $Q_S(\beta_0) = S^{\text{T}}(\beta_0)S(\beta_0)$, $Q_{SR} = S^{\text{T}}(\beta_0)R(\beta_0)$, and $Q_R = R^{\text{T}}(\beta_0)R(\beta_0)$. Suppose $\chi_k^2(1-\alpha)$ is the $1-\alpha$ quantile of a Chi-square distribution with k degrees of freedom and $\text{CDF}_{\chi_k^2}(x)$ is the cumulative distribution function of a Chi-square distribution with k degrees of freedom. The following theorem shows that $T_{\text{mrAR}}(\beta_0), T_{\text{mrK}}(\beta_0)$, and $T_{\text{mrCLR}}(\beta_0)$ have asymptotically pivotal distributions under the null $H_0 : \beta = \beta_0$.

Theorem 1. *Suppose Assumptions 1- 4 and $H_0 : \beta = \beta_0$ hold. For any $\alpha \in (0,1)$, as $n_1, n_2 \to \infty, n_1/n_2 \to c \in (0,\infty)$, we have*

$$P\left(T_{\text{mrAR}} > \chi_L^2(1-\alpha)\right) \to \alpha,\ P\left(T_{\text{mrK}} > \chi_1^2(1-\alpha)\right) \to \alpha,\ P\left(w(T_{\text{mrCLR}}; Q_R) < \alpha\right) \to \alpha,$$

where

$$w(x;y) = 1 - \frac{2G\left(\frac{L}{2}\right)}{\sqrt{\pi}G\left(\frac{L-1}{2}\right)} \int_0^1 \text{CDF}_{\chi_L^2}\left(\frac{x+y}{1+y\frac{z^2}{x}}\right)(1-z^2)^{\frac{L-3}{2}}dz$$

and $G(\cdot)$ is the gamma function.

Theorem 1 shows that under $H_0 : \beta = \beta_0$, the two-sample summary data versions of the AR, K, and CLR tests converge to the classical null distributions for the three tests under the single-sample individual data setting. In particular, like the original CLR test, mrCLR test requires solving the integral $w(x;y)$ to obtain critical values; this integral can be computed by using off-the-shelf numerical integral solvers. We can also use the duality between tests and confidence intervals to derive asymptotically valid $1-\alpha$ confidence intervals for each test.

In the supplementary materials, we extend these results and construct a point estimator based on minimizing the mrAR test statistic. We show that when the estimated covariance matrices $\widehat{\Sigma}_\Gamma$ and $\widehat{\Sigma}_\gamma$ are diagonal, $\widehat{\beta}_{\text{mrLIML}}$ is equivalent to the estimator proposed by Zhao et al. (2018).

3 Simulations

We conduct simulation studies to study the performance of our test statistics. The data is generated from the structural model in (1) with $n_1 = n_2 = 100,000$, on the same order as the sample size in the data analysis section 4. The random error $(\delta_{1i}, \delta_{2i})$ is generated from a bivariate standard Normal and the random error ϵ_{li} is equal to $\epsilon_{li} = \rho\delta_{li} + (1-\rho^2)^{1/2}e_{li}$; the term e_{li} is from an independent standard Normal and $\rho = 0.1$. We remark that ρ signifies the endogeneity between the outcome and the exposure. The $L = 10$ instruments take on values $0, 1, 2$, similar to how SNPs are recorded in GWAS, and are generated independently from a Binomial distribution $\text{Binom}(2, p_j), j = 1, \cdots, L$ with p_i drawn from a uniform distribution $\text{Unif}(0.1, 0.9)$. After generating individual-level data, we compute the summary statistics for sample $l = 1$, $\widehat{\Gamma}$ and $\widehat{\Sigma}_\Gamma$, by running an OLS regression between Y_{1i} and Z_{1ij} for each instrument j and extracting the estimated coefficient and standard error. Similarly, we compute the summary statistics for sample $l = 2$, $\widehat{\gamma}$ and $\widehat{\Sigma}_\gamma$, by running an OLS regression between D_{2i} and Z_{2ij} for each instrument. The simulation varies the exposure effect β and the IV-exposure relationship γ. The simulation is repeated $1,000$ times.

We examine the power of our proposed tests, $T_{\text{mrAR}}, T_{\text{mrK}}$, and T_{mrCLR} and the power of existing tests in MR, specifically tests based on the IVW estimator, MR-Egger regression, the W.Median estimator, all implemented in the software `Mendelianrandomization` (Yavorska and Burgess, 2017), and MR-RAPs without the robust loss function (Zhao et al., 2018). Figure 1 shows the power curves when the null hypothesis is $H_0 : \beta = 0$ (left panel) or $H_0 : \beta = 1$ (right panel); significance level is set at $\alpha = 0.05$. The top panels shows the case when γ ranges from $\{(r-0.5)/n_1\}^{1/2}$ to $\{(r+0.5)/n_1\}^{1/2}$ and $r = 1$ the bottom panel shows the case when $r = 4$; the value r approximately corresponds to the first-stage F-statistic test typically used to test instrument strength. Under $H_0 : \beta = 0$, all tests correctly control Type I error under $r = 1$ and $r = 4$. But, our three tests, IVW, and MR-RAPs have power under $r = 1$ and $r = 4$ cases, with T_{mrCLR} having the best power among them; this is in agreement with Andrews et al. (2006) who showed that the CLR test in the single-sample individual data setting is nearly optimal. Under $H_0 : \beta = 1$, none of the pre-existing methods except MR-RAPs have Type I error control when instruments are weak. In contrast, our tests always maintain Type I error control. Also, our tests have power under the alternative, with T_{mrCLR} having the best power among them. In fact, in the supplementary materials, we show that tests based on the IVW estimator, weighted median estimator, and MR Egger regression only locally control Type I error at the null $H_0 : \beta = 0$ when the instruments are weak.

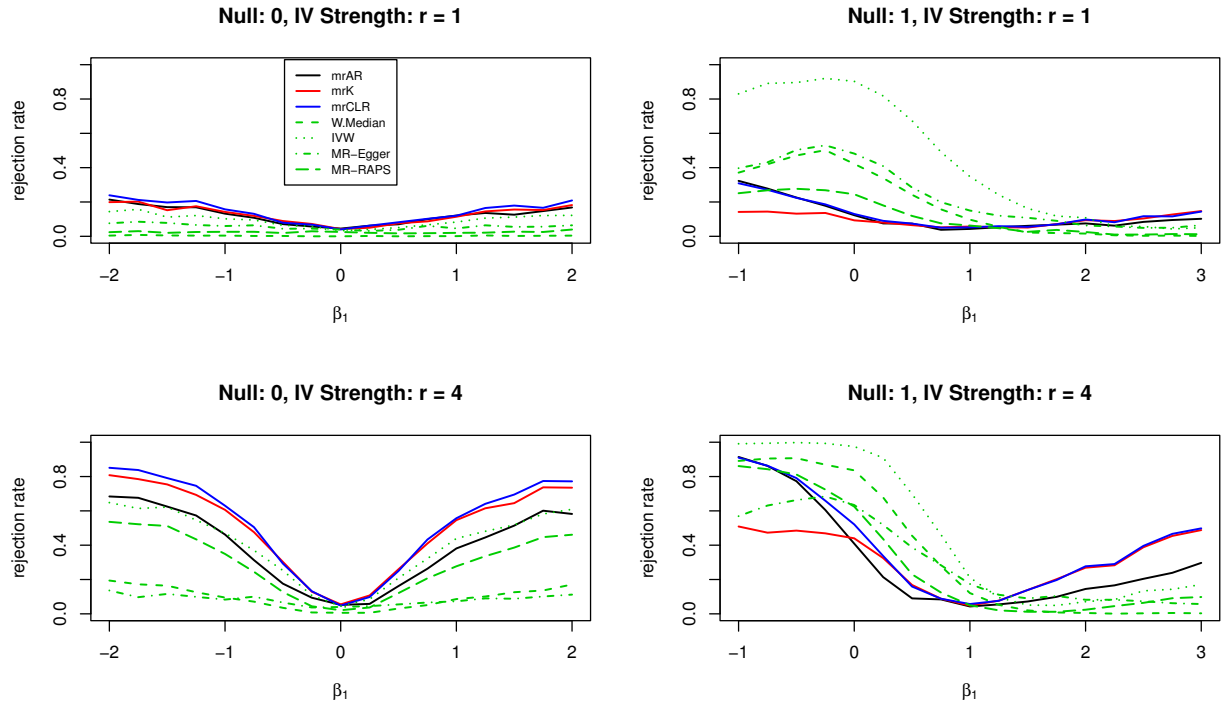

Figure 1: Power curves under different null and IV strength. The left panel is under $H_0 : \beta_0 = 0$ and the right panel is under $H_0 : \beta = 1$. The top panel sets instrument strength to $r = 1$ and the bottom panel sets instrument strength to $r = 4$; r approximately corresponds to the first-stage F-statistic test for IV strength.

4 Data Analysis

We validate our proposed tests by considering a prototypical MR study on the relationship between BMI and systolic blood pressure where the exposure effect is known to be positive. We use the data set prepared by Zhao et al. (2018) where the authors used three independent GWAS, one from the UK Biobank GWAS (SBP-UKBB) and the other two from GWAS by the Genetic Investigation of ANthropometric Traits (GIANT) consortium (Locke et al., 2015). The BMI-MAL and SBP-UKBB datasets provide summary statistics of the IV-exposure and IV-outcome statistics, respectively. The BMI-FEM dataset is a selection GWAS containing summary statistics of the IV-exposure relationship and is used to pre-screen for strong and uncorrelated IVs.

We conduct two types of analysis with the data. First, we use the data as provided and examine differences between the IVW, weighted median, MR-Egger estimator, MR-RAPS with a robust loss function, and our methods when either $L = 25$ or $L = 160$ instruments are used. The results are in Table 1.

We see that almost all methods provide similar 95% confidence intervals. Even though the weighted median, MR-RAPS, and MR-Egger are robust to invalid instruments, their confidence intervals are similar to confidence intervals from T_{mrK} and T_{mrCLR}, which are not robust to invalid

Table 1: MR study concerning the effect of body mass index on systolic blood pressure. Parentheses represent 95% confidence intervals.

	25 Instruments	160 Instruments
mrK	(0.205, 0.530)	(0.377, 0.771)
mrCLR	(0.211, 0.524)	(0.415, 0.731)
mrAR	($\emptyset$)	($\emptyset$)
IVW	0.332 (0.063, 0.600)	0.317 (0.101, 0.534)
W.Median	0.520 (0.280, 0.759)	0.522 (0.318, 0.726)
MR-Egger	0.622 (0.101, 1.143)	0.452 (0.112, 0.792)
MR-RAPS	0.354 (0.097, 0.610)	0.378 (0.141, 0.615)

instruments. This suggests either that invalid instruments play a minimal role in this data or, as Small and Rosenbaum (2008) suggests, in the presence of invalid IVs, the first-order bias comes not from invalid IVs, but from weak IVs. This is also based on the observation that mrK produced two split intervals, one in the negative region (-14.375, -10.905) (when $L = 25$) or (-10.376, -6.447) (when $L = 160$) and the other in the positive region. We only report the positive region in Table 1 since we know a priori that the effect is positive; when the exposure effect direction is unknown, we recommend taking the union of the intervals. Surprisingly, T_{mrAR} produces an empty interval. This behavior may be an indication that the model is incorrect or that the test lacks power; we plan on investigating this property of T_{mrAR} in future work.

Second, inspired by Bound et al. (1995), we "stress-test" the pre-existing methods and replace each of the original IV-exposure effect $\widehat{\gamma}_j$ in BMI-MAL and the IV-outcome effect $\widehat{\Gamma}_j$ in SBP-UKBB by the values generated below.

$$\widehat{\gamma}_j^{\text{new}} \sim N(K\widehat{\gamma}_j, \widehat{\Sigma}_{\gamma,j}), \quad \widehat{\Gamma}_j^{\text{new}} \sim N(K\widehat{\gamma}_j\beta, \widehat{\Sigma}_{\Gamma,j})$$

Here, β is the true exposure effect and is set to be 0.5 and 1.5. The parameter K controls IV strength and ranges from 0 to 1. Under $K = 1$, the new IV-exposure and IV-outcome effects are essentially the original effects, but with a known exposure effect value β. But, as K decreases to 0, the IV becomes weaker than the original ones. In the extreme case when $K = 0$, there is no way to consistently estimate β; the new IV-exposure and IV-outcome effects look statistically indistinguishable if the true exposure effect is $\beta = 1$ or $\beta = 1000$.

An ideal confidence interval should be able to (i) simultaneously and automatically detect the lack of identification of the exposure effect β by producing an infinite confidence interval when $K = 0$ and a bounded confidence intervals as K moves away from zero and (ii) for all values of K, provide 95% coverage. As Figure 2 shows, when we run the existing MR methods, none of them achieve

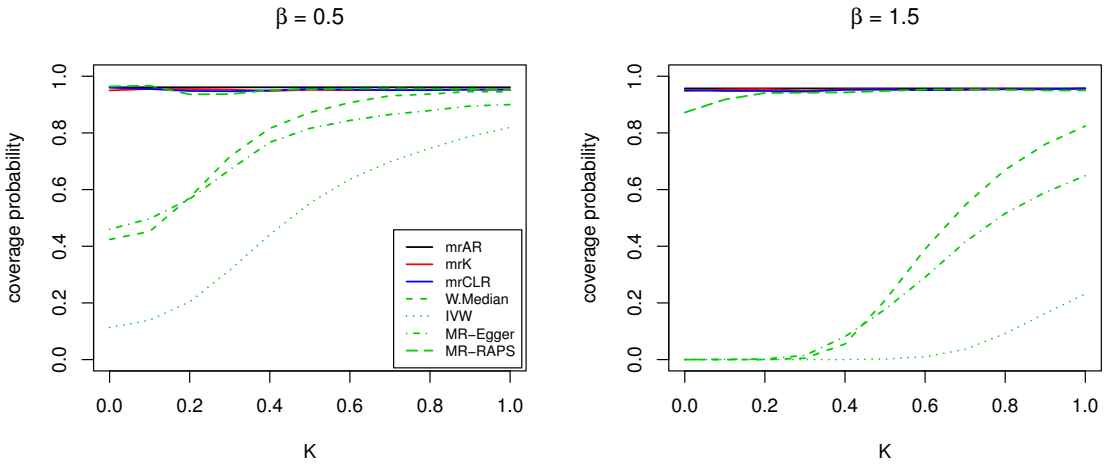

Figure 2: Coverage probability under different IV strength. The left panel sets the true causal effect β to be 0.5 and the right panel sets β to be 1.5.

these two goals. At $K = 0$, they always produce bounded confidence intervals, even though there is no way to identify the exposure effect from data. Specifically, when $K = 0, \beta = 0.5$, the confidence interval given by Weighted median, IVW, MR-Egger and MR-RAPS are (-0.624, 0.321), (-0.411, 0.265), (-0.527, 0.624), and (-0.460, 0.020), respectively. They are bounded and only the confidence interval generated from MR-Egger covers the true effect $\beta = 0.5$. But $T_{\text{mrAR}}, T_{\text{mrK}}, T_{\text{mrCLR}}$ produce unbounded confidence intervals. We observe a similar phenomena when $K = 0$ and $\beta = 1.5$: the confidence intervals generated from Weighted median, IVW, MR-Egger and MR-RAPS are (-0.385, 0.547), (-0.312, 0.330), (-0.826, 0.317), and (-0.575, 0.942), respectively. All of them are bounded and fail to cover the true causal effect, but our tests produce unbounded confidence intervals. Also, when K is near zero so that the IV-exposure is sufficiently weak, all methods except MR-RAPS fail to achieve 95% coverage. In contrast, our tests always satisfy the two criterions (i) and (ii). They automatically produce infinite confidence intervals when $K = 0$ to alert the researcher about lack of identification and produce bounded intervals as K moves away from zero. They also always maintain 95% coverage for any value of K. In short, our proposed tests adapt to the data and always produce honest intervals regardless of the underlying instrument strength.

5 Conclusion

In this paper, we propose weak-IV robust test statistics for two-sample summary data in MR. We extend the existing AR, Kleibergen, and CLR tests in econometrics and show that it has the same Type I error control under weak instrument asymptotics. The numerical results show that the proposed tests, especially the mrCLR test, have better size control and power compared to current methods when the instruments are weak. Additionally, when we stress-test different methods, our

methods, especially the CLR test, adapts to the underlying instrument strength and provides valid 95% coverage. In practice, we recommend MR investigators use the mrCLR test to test exposure effects as it provides valid confidence intervals regardless of IV strength. The code to implement our tests is in the supplementary materials.

References

Anderson, T. W., Rubin, H., et al. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Andrews, D. W., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752.

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336.

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

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016a). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016b). Assessing the suitability of summary data for two-sample mendelian randomization analyses using mr-egger regression: the role of the i 2 statistic. *International journal of epidemiology*, 45(6):1961–1974.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665.

Burgess, S., Scott, R. A., Timpson, N. J., Smith, G. D., Thompson, S. G., Consortium, E.-I., et al. (2015). Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology*, 30(7):543–552.

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from mendelian randomization studies with weak instruments. *Statistics in medicine*, 30(11):1312–1323.

Choi, J., Gu, J., and Shen, S. (2018). Weak-instrument robust inference for two-sample instrumental variables regression. *Journal of Applied Econometrics*, 33(1):109–125.

Davey Smith, G. and Ebrahim, S. (2003). "mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica: Journal of the Econometric Society*, pages 1365–1387.

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46(6):1985–1998.

Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

Pierce, B. L. and Burgess, S. (2013). Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933.

Staiger, D. and Stock, J. H. (1997). Instrumental variables with weak instruments. *Econometrica*, 65(3):557–586.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Yavorska, O. O. and Burgess, S. (2017). Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology*, 46(6):1734–1739.

Zhao, Q., Wang, J., Bowden, J., and Small, D. S. (2019). Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34(2):317–333.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652*.