Biometrika, *pp*. 1–17

Robust confidence intervals for causal effects with possibly invalid instruments

BY H. KANG, T. T. CAI, AND D. S. SMALL

Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

khyuns@wharton.upenn.edu tcai@wharton.upenn.edu dsmall@wharton.upenn.edu

SUMMARY

Instrumental variables have been widely used to estimate the causal effect of a treatment on an outcome. Existing confidence intervals for causal effects based on instrumental variables assume that all of the putative instrumental variables are valid; a valid instrumental variable is a variable that affects the outcome only by affecting the treatment and is not related to unmeasured confounders. However, in practice, some of the putative instrumental variables are likely to be invalid. This paper presents a simple and general approach to construct a confidence interval that is robust to possibly invalid instruments. The robust confidence interval has theoretical guarantees on having the correct coverage. The paper also shows that the robust confidence interval outperforms traditional confidence intervals popular in instrumental variables literature when invalid instruments are present. The new approach is applied to a study of the causal effect of income on food expenditures.

Some key words: Anderson-Rubin test; Confidence interval; Hypothesis testing; Invalid instrument; Instrumental variable; Weak instrument.

1. INTRODUCTION

The instrumental variables method is a popular method to estimate the casual effect of a treatment, exposure, or policy on an outcome when unmeasured confounding is present (Angrist et al., 1996; Tan, 2006; Baiocchi et al., 2014). Informally speaking, the method relies on having instruments that are (A1) related to the exposure, (A2) only affect the outcome by affecting the exposure (no direct effect), and (A3) are not related to unmeasured confounders that affect the exposure and the outcome. Figure 1 depicts the three core assumptions (see Section 2·2 for a formal definition of valid instruments). Unfortunately, in many applications, practitioners are unsure if all of the candidate instruments satisfy these assumptions. For example, in Mendelian randomization, the candidate instruments are genetic variants that have been shown to be associated with the exposure in Genome-wide Association Studies, but are not currently thought to affect the outcome through a pathway other than through the exposure (Davey Smith & Ebrahim, 2003, 2004; Burgess et al., 2012). However, typically, due to the incomplete understanding of the genetic markers and their complex biological pathways, it is possible that the genetic variants affect the outcome through a pathway other than the exposure that has yet to be discovered, violating (A2) or might be associated with unmeasured confounders through population stratification or linkage disequilibrium, violating (A3) (Davey Smith & Ebrahim, 2003, 2004; Lawlor et al., 2008; Solovieff et al., 2013; Kang et al., 2015). For example, in 1986, Katan (1986), in one of the first discussions of MR, used the apolipoprotein E polymorphism (APOE)'s to study the causal effect of serum cholesterol level on cancer. However, subsequent discoveries about

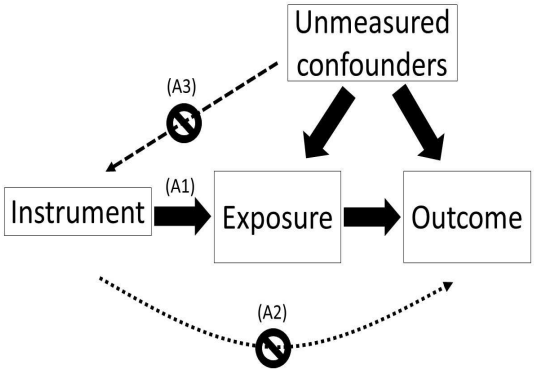

Fig. 1. An illustration of instrumental variables assumptions. Arrows represent associations between variables. Absence of arrows indicate no relationship. Numbers (A1), (A2), and (A3) indicate different instrumental variables assumptions.

the APOE gene and its various biological pathways to the outcome, such as its pathway through cholesterol biomarkers (Wilson et al., 1994), would make the APOE gene an invalid instrument and make an IV analysis based on it biased (Davey Smith & Ebrahim, 2004). A similar problem arises in economics where one is presented with multiple candidate instruments and one is always uncertain whether some of them may violate the core assumptions (Murray, 2006).

Violation of (A1), known as the weak instrument problem, has been studied in great detail and many robust tools exist to handle this case; see Stock et al. (2002) for a survey. In contrast, violations of (A2) and (A3), known as the invalid instrument problem (Murray, 2006), have not been studied in great detail, with recent works by Kolesár et al. (2013) and Kang et al. (2015). Kolesár et al. (2013) considered the case when the instruments violate (A2) and proposed an orthogonality condition where the instruments' effect on the exposure are orthogonal to their effects on the outcome in order to identify the causal effect. Kang et al. (2015) considered more general violations of (A2) and (A3) based on imposing an upper bound on the number of invalid instruments among the candidate instruments, without knowing exactly which instruments are valid or knowing the exact number of invalid instruments, or imposing any structure on the instruments. But, both papers only studied point estimation and not confidence intervals.

This paper focuses on developing robust confidence intervals when candidate instruments may be invalid, specifically the candidate instruments might violate (A2) and (A3). Similar to Kang et al. (2015), we only assume that we know an upper bound on how many of the candidate instruments are invalid and do not know which instruments are valid. For this setting, we propose a simple and general confidence interval procedure that theoretically guarantees the correct coverage rate and is robust to possibly invalid instruments. The confidence interval is based on inverting statistical tests over a range of subsets of instruments that are potentially valid. We also propose various ways to obtain short and informative confidence intervals with our procedure by exploring various tests common in instrumental variables and conducting pretests. The simulation study shows that our method is robust when invalid instruments are present compared to other popular methods in the instrumental variables literature. We also demonstrate that our method can produce valid, short, and informative confidence intervals by analyzing a data set concerning the causal effect of income on food expenditure.

2. ROBUST CONFIDENCE INTERVALS BY INVERTING TESTS

2·1. *Notation*

We use the potential outcomes notation (Rubin, 1974) for instruments laid out in Holland (1988), Small (2007) and Kang et al. (2015). Specifically, let there be L potential candidate instruments and n individuals in the sample. Let $Y_i^{(d,z)}$ be the potential outcome that individual i would have if the individual were to have exposure d, a scalar value, and instruments z, an L dimensional vector. Let $D_i^{(z)}$ be the potential exposure if the individual had instruments z. For each individual, only one possible realization of $Y_i^{(d,z)}$ and $D_i^{(z)}$ is observed, denoted as Y_i and D_i, respectively, based on his/her observed instrument values $Z_{i.}$, an L dimensional vector, and observed exposure D_i. In total, we have n observations of $(Y_i, D_i, Z_{i.})$. We denote $Y = (Y_1, \ldots, Y_n)$, $D = (D_1, \ldots, D_n)$ and Z to be the n by L matrix where row i consists of $Z_{i.}$.

For any subset $A \subseteq \{1, \ldots, L\}$ with cardinality $c(A)$, let Z_A be an n by $c(A)$ matrix of instruments where the columns of Z_A are from the set A, $P_{Z_A} = Z_A(Z_A^T Z_A)^{-1} Z_A^T$ be the orthogonal projection matrix onto the column space of Z_A and R_{Z_A} be the residual projection matrix so that $R_{Z_A} + P_{Z_A}$ is equal to an n by n identity matrix. We assume that Z_A has a proper inverse unless otherwise stated. Also, for any L dimensional vector π, let π_A only consist of elements of the vector π determined by the set $A \subseteq \{1, \ldots, L\}$.

2·2. *Model and definition of valid instruments*

For two possible values of the exposure d', d and instruments z', z, we assume the following potential outcomes model

$$Y_i^{(d',z')} - Y_i^{(d,z)} = (z' - z)^T \phi^* + (d' - d)\beta^*, \quad E\{Y_i^{(0,0)} \mid Z_{i.}\} = Z_{i.}^T \psi^* \tag{1}$$

where ϕ^*, ψ^*, and β^* are unknown parameters. The parameter β^* represents the causal parameter of interest, the causal effect (divided by $d' - d$) of changing the exposure from d' to d on the outcome. The parameter ϕ^* represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\phi^* = 0$. The parameter ψ^* represents violation of (A3), the presence of unmeasured confounding between the instrument and the outcome. If (A3) holds, then $\psi^* = 0$.

Let $\pi^* = \phi^* + \psi^*$ and $\epsilon_i = Y_i^{(0,0)} - E\{Y_i^{(0,0)} \mid Z_{i.}\}$. When we combine equations (1) along with the definition of ϵ_i, the observed data model becomes

$$Y_i = Z_{i.}^T \pi^* + D_i \beta^* + \epsilon_i, \quad E(\epsilon_i \mid Z_{i.}) = 0 \tag{2}$$

The observed model is also known as the under-identified single-equation linear model in econometrics (page 83 of Wooldridge (2010)). Note that (2) is not a usual regression model because D_i might be correlated with ϵ_i. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. Kang et al. (2015) discusses extensions of the model (2) to include heterogeneous causal effects and non-linear effects. Also, the model can incorporate exogenous covariates, say $X_{i.}$ and we can project them out by using Frisch-Waugh-Lovell Theorem to reduce the model to (2) (Davidson & MacKinnon, 1993). The parameter π^* in the observed data model (2) combines both the violation of (A2), represented by ϕ^*, and the violation of (A3), represented by ψ^*. If both (A2) and (A3) are satisfied, then $\phi^* = \psi^* = 0$ and $\pi^* = 0$. Hence, the value of π^* captures whether instruments are valid versus invalid. Definition 1 formalizes this idea.

DEFINITION 1. *Suppose we have L candidate instruments along with the models* (1)–(2). *We say that instrument $j = 1, \ldots, L$ is valid if $\pi_j^* = 0$ and invalid if $\pi_j^* \neq 0$.*

When there is only one instrument, $L = 1$, Definition 1 of a valid instrument is identical to the definition of a valid instrument in Holland (1988). Specifically, assumption (A2), the exclusion restriction, which means $Y_i^{(d,z)} = Y_i^{(d,z')}$ for all d, z, z', is equivalent to $\phi^* = 0$ and assumption (A3), no unmeasured confounding, which means $Y_i^{(d,z)}$ and $D_i^{(z)}$ are independent of $Z_{i\cdot}$ for all d and z, is equivalent to $\psi^* = 0$, implying $\pi^* = \phi^* + \psi^* = 0$. Definition 1 is also a special case of the definition of a valid instrument in Angrist et al. (1996) where here we assume the model is additive, linear, and has a constant treatment effect β^*. Hence, when multiple instruments, $L > 1$, are present, our models (1)–(2) and Definition 1 can be viewed as a generalization of the definition of valid instruments in Holland (1988).

Let $s = 0, \ldots, L-1$ to be the number of invalid instruments and U be an upper bounded on s plus 1, i.e. the number of invalid instruments is assumed to be less than U. We assume that there is at least one valid IV, even if we don't know which among the L IV is valid, since if all L IVs are invalid (i.e. $s = L$), identification would not be possible (Kang et al., 2015). This setup was also considered in Kang et al. (2015) as a relaxation to traditional instrumental variables setups where one knows exactly which instruments are valid and invalid. For simplicity, we consider the case where at less than half of the candidate instruments are invalid, $U \leq L/2$, because all the parameters in the model (2) are always identified under this setup (Kang et al., 2015). However, the proposed procedures will work for any upper bound U, exceeding $L/2$.

2·3. *A general procedure for robust confidence intervals*

Let $I = \{1, \ldots, L\}$ be the L candidate instruments and $B^* \subseteq \{1, \ldots, L\}$ be the true set of valid instruments. Given B^*, consider a test statistic $T(\beta_0, B^*)$ of the null hypothesis $H_0 : \beta^* = \beta_0$ versus the alternative $H_a : \beta^* \neq \beta_0$. It is well known that inverting a test based on $T(\beta_0, B^*)$ that has level α provides a $1-\alpha$ confidence interval for β^*, denoted as $C_{1-\alpha}(Y, D, Z, B^*)$.

$$C_{1-\alpha}(Y, D, Z, B^*) = \{\beta_0 \mid T(\beta_0, B^*) \leq \nu_{1-\alpha}\} \tag{3}$$

where $\nu_{1-\alpha}$ is the $1-\alpha$ quantile of the null distribution of $T(\beta_0, B^*)$.

Unfortunately, in our problem, we do not know the true set B^* of valid instruments, so we cannot directly use (3). However, in our model description in Section 2·2, we have an upper bound on the number of invalid instruments, s, by U where $s < U$ and consequently, a lower bound for the number of valid instruments, $L - s > L - U$ and thus a lower bound on the cardinality of the set B^*, $c(B^*)$, $c(B^*) > L - U$. Using this lower bound, we can take unions of $C_{1-\alpha}(Y, D, Z, B)$ over possible sets of valid instruments $B \subseteq I$ where $c(B) > L - U$; the confidence interval using the true set of instruments $C(Y, D, Z, B^*)$ will be in this union since $c(B^*) > L - U$. Our proposal is exactly this, except that we restrict the subsets B to be of size $c(B) = L - U + 1$.

$$C_{1-\alpha}(Y, D, Z) = \cup_B \{C_{1-\alpha}(Y, D, Z, B) \mid B \subseteq I, c(B) = L - U + 1\} \tag{4}$$

The proposed confidence interval $C_{1-\alpha}(Y, D, Z)$ is simple and general; for any test statistic $T(\beta_0, B)$ with a valid size for $B = B^*$, one simply takes unions of confidence intervals of $T(\beta_0, B)$ over subsets of instruments B where $c(B) = L - U + 1$. In addition, a key feature of our procedure is that it is not necessary to go through all the subsets of possible valid instruments larger than $c(B) > L - U$; simply looking at the smallest possible subsets of valid instruments, i.e. those subsets that are at the lower boundary of $L - U$, $c(B) = L - U + 1$, is sufficient to provide the $1 - \alpha$ coverage.

Theorem 1 states that the procedure in (4) produces a valid confidence interval since $c(B^*) > L - U$, there is some subset of valid instruments with cardinality $L - U + 1$.

THEOREM 1. *Suppose model* (2) *holds and* $s < U$. *Given* α, *consider any test statistic* $T(\beta_0, B)$ *with the property that for any* $B \subseteq B^*$, $T(\beta_0, B)$ *has size at most* α *under the null hypothesis* $H_0 : \beta^* = \beta_0$. *Then,* $C_{1-\alpha}(Y, D, Z)$ *in* (4) *always has at least* $1 - \alpha$ *coverage.*

Proof of Theorem 1. By $s < U$, we have $c(B^*) > L - U$. Consequently, there is a subset $\tilde{B} \subseteq B^*$ where $c(\tilde{B}) = L - U + 1$ and $\tilde{B}$ only contains only valid instruments. Since $\tilde{B}$ only contains valid instruments, $\text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z, \tilde{B})\} \geq 1 - \alpha$ for all π^*, β^*. Hence, we have

$$\text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z)\} \geq \text{pr}\{\beta^* \in C_{1-\alpha}(Y, D, Z, \tilde{B})\} \geq 1 - \alpha$$

for all values of π^*, β^*. □

A potential caveat to our procedure is computational feasibility. Even though we restrict the union to subsets of exactly size $c(B) = L - U + 1$, if the number of candidate instruments, L, grows, $C(Y, D, Z)$ becomes computationally burdensome. However, in many instrumental variables studies, it is difficult to find good candidate instruments and rarely the number of these candidates instruments exceed $L = 20$, which modern computing can handle. Hence, our procedure in (4) is computationally tractable for most practical applications.

2·4. *Choice of test statistics*

In the instrumental variables literature, there are many tests of causal effects $T(\beta_0, B)$ that can be used with Theorem 1 to construct valid $1 - \alpha$ confidence interval $C_{1-\alpha}(Y, D, Z)$ in the presence of invalid instruments. A natural question to ask, then, is among these tests, which test statistic when used with Theorem 1 provides the smallest length confidence interval and thus, from a practical standpoint, provides the most informative confidence interval?

The most popular test is the t-test based on based on the asymptotic normal distribution of the two-stage least squares estimator. The two-stage least squares estimator of β^* for a given B, denoted as $\hat{\beta}_{B,TSLS}$, is the solution to the minimization problem $\|P_{\tilde{Z}_B} R_{Z_A}(Y - D\beta)\|_2^2$ where A is the complement of the set B, $A = I \setminus B$, and $\tilde{Z}_B = R_{Z_A} Z_B$. If $\hat{u}(B)$ is the residuals from the fitted model, $\hat{u}(B) = R_{Z_A}(Y - D\hat{\beta}_{B,tsls})$ and $\hat{D}(B)$ is the projection of D on to the column space of $\tilde{Z}_B$, $\hat{D}(B) = P_{\tilde{Z}_B} D$, then the t-test is defined as

$$\text{TSLS}(\beta_0, B) = \sqrt{n - c(A) - 1} \left\{ \frac{\hat{\beta}_{B,TSLS} - \beta_0^*}{\sqrt{\|\hat{u}(B)\|_2^2 / \|\hat{D}(B)\|_2^2}} \right\} \tag{5}$$

If $B \subseteq B^*$, standard econometrics arguments show that (5) converges to an asymptotic Normal distribution (Wooldridge, 2010). In practice, the test (5) is approximately valid when all the subset of instruments B among the candidate instruments I are strong, or in other words, strongly associated with the exposure. Unfortunately, instruments can be weak in practice and the nominal size of tests based on two-stage least squares can be misleading (Staiger & Stock, 1997).

Stock et al. (2002) presents a survey of tests that are robust to weak instruments. Specifically, for a given B, let $W(B)$ be an n by 2 matrix where the first column contains $R_{Z_A} Y$ and the second column contains $R_{Z_A} D$ where, again, A is the complement of the set B, $A = I \setminus B$. Let $a_0 = (\beta_0, 1)$ and $b_0 = (1, -\beta_0)$ to be two-dimensional vectors and $\hat{\Sigma} = W(B)^T M_{\tilde{Z}_B} W(B)/(n - L)$. Let $\hat{S}(B)$ and $\hat{T}(B)$ be two-dimensional vectors

$$\hat{S}(B) = \frac{(\tilde{Z}_B^T \tilde{Z}_B)^{-1/2} \tilde{Z}_B^T W(B) b_0}{\sqrt{b_0^T \hat{\Sigma} b_0}}, \quad \hat{T}(B) = \frac{(\tilde{Z}_B^T \tilde{Z}_B)^{-1/2} \tilde{Z}_B^T W(B) \hat{\Sigma}^{-1} a_0}{\sqrt{a_0^T \hat{\Sigma}^{-1} a_0}}$$

along with the following scalar values

$$\hat{Q}_{11}(B) = \hat{S}(B)^T\hat{S}(B), \quad \hat{Q}_{12}(B) = \hat{S}(B)^T\hat{T}(B)$$
$$\hat{Q}_{22}(B) = \hat{T}(B)^T\hat{T}(B)$$

Based on $\hat{Q}_{11}(B)$, $\hat{Q}_{12}(B)$, and $\hat{Q}_{22}(B)$, we define the following tests, the Anderson-Rubin test (Anderson & Rubin, 1949), the Lagrangian multiplier test (Kleibergen, 2002), and the conditional likelihood test (Moreira, 2003).

$$\text{AR}(\beta_0, B) = \hat{Q}_{11}(B)/c(B) \tag{6}$$
$$\text{LM}(\beta_0, B) = \hat{Q}_{12}^2(B)/\hat{Q}_{22}(B) \tag{7}$$
$$\text{CLR}(\beta_0, B) = \frac{1}{2}\left\{\hat{Q}_{11}(B) - \hat{Q}_{22}(B)\right\} + \frac{1}{2}\sqrt{\{\hat{Q}_{11}(B) + \hat{Q}_{22}(B)\}^2 - 4\{\hat{Q}_{11}(B)\hat{Q}_{22} - \hat{Q}_{12}^2(B)\}} \tag{8}$$

Each of the three tests have their unique robustness characteristics and properties, but all of them have been shown to be robust to weak instruments(Staiger & Stock, 1997; Stock et al., 2002; Kleibergen, 2002; Moreira, 2003; Dufour, 2003; Andrews et al., 2006). There is no uniformly most powerful test among the three tests, but Andrews et al. (2006) and Mikusheva (2010) suggest using (8) due to its generally favorable power compared to (6) and (7) in most cases when weak instruments are present. However, the Lagrangian multiplier test (7) and the Anderson-Rubin test (6) have the unique feature where both tests (or derivatives of) can be used as a pretest to check whether the candidate subset of instruments B contain only valid instruments. This feature is particularly useful for our problem where we have possibly invalid instruments (see Section 2·6 and the Supplementary Material). Also, among the three tests, the Anderson-Rubin test is the simplest in that it can be written as a standard F-test in regression where the outcome is $R_{Z_A}(Y - D\beta_0)$, the regressors are $\tilde{Z}_B$, and we are testing whether the coefficients associated with $\tilde{Z}_B$ are zero or not with the standard F-test. Finally, the Lagrangian multiplier test and the conditional likelihood ratio test require an assumption that the exposure, D, is linearly related to the exposure Z by $D_i = Z_{i\cdot}^T\gamma^* + \xi_i$ where γ^* is an L dimensional vector and ξ_i is a random error term with mean zero, homoscedastic variance, and is independent of Z; the Anderson-Rubin test does not require this linearity assumption (Dufour, 2003).

2·5. *Empty confidence intervals and Anderson-Rubin test*

Our procedure $C_{1-\alpha}(Y, D, Z)$ involves taking the union of confidence intervals, at least one of which is based on valid instruments, but some of which may be based on invalid instruments. For instance, if we have a subset B with $c(B) = L - U + 1$ as in equation (4), but B is not a the subset of B^*, B contains at least one true invalid instruments from the set $A^* = I \setminus B^*$ and we may end up with confidence intervals $C_{1-\alpha}(Y, D, Z, B)$ that are biased. Such a potentially biased interval is included in the interval for $C_{1-\alpha}(Y, D, Z)$. Even though $C_{1-\alpha}(Y, D, Z)$ will have correct coverage regardless of this inclusion, the unnecessary inclusion of it may elongate the interval $C_{1-\alpha}(Y, D, Z)$ and produce an uninformative interval.

One method to deal with this problem is to choose a test statistic where for $B \not\subseteq B^*$, $C_{1-\alpha}(Y, D, Z, B)$ will usually produce an empty interval. For example, the Anderson-Rubin test statistic in (6) has this feature. To illustrate, suppose, for simplicity, we assume $D_i = Z_i^T\gamma^* + \xi_i$ where ϵ_i, ξ_i are independent and bivariate Normal with mean 0 and covariance Σ. If we subtract $D\beta_0$ and multiply by R_{Z_A} from both sides of (2) and substitute D_i with $D_i = Z_i^T\gamma^* + \xi_i$, we

obtain

$$R_{Z_A}(Y - D\beta_0) = R_{Z_A}Z_B\kappa^* + R_{Z_A}\epsilon, \quad \kappa = \pi_B^* + \gamma_B^*(\beta^* - \beta_0) \tag{9}$$

where γ_B^* and π_B^* are the components of γ^* and α^* vectors for the indices that belong to the subset B. As explained in Section 2·4, the Anderson–Rubin test can also be written as an F test where the null is $H_0 : \kappa^* = 0$. This null corresponds to testing both whether the instruments are valid, $\pi_B^* = 0$, and whether the treatment effect is β_0, $\beta^* = \beta_0$. Rejecting $H_0 : \kappa^* = 0$ in favor of the alternative would imply that one is rejecting the null because the treatment effect is not β_0 or because the instruments in set B are not valid. Thus, when a candidate set of instruments B contains an invalid instrument, the Anderson–Rubin test will likely reject when $\beta^* = \beta_0$, and so the inversion of the Anderson–Rubin test will produce an empty confidence interval or a short confidence interval (see Kadane & Anderson (1977) and Small (2007) for the exact circumstance under which the Anderson–Rubin test will have this property).

2·6. *Pretest for invalid instruments*

Another method to avoid taking unions of unnecessary intervals is by conducting a preliminary test that checks whether each of the subsets B where $c(B) = L - U + 1$ contains any invalid instruments before proceeding to construct a confidence interval with B. Specifically, for a desired confidence interval $1 - \alpha$, consider the null hypothesis that B contains only valid instruments, $\pi_B^* = 0$, and the corresponding test statistic $S(B)$, which serve as a pretest for the validity of the instruments in B. For any $\alpha_1 < \alpha$, suppose the test based on $S(B)$ has level α_1 under the null hypothesis that B only contains valid instruments with $q_{1-\alpha_1}$ as the $1 - \alpha_1$ quantile of $S(B)$ under the null hypothesis. Then, a $1 - \alpha$ confidence interval can be constructed based on $S(B)$ as follows.

$$P_{1-\alpha}(Y, D, Z) = \cup_B\{C_{1-\alpha_2}(Y, D, Z, B) \mid c(B) = L - U + 1, S(B) \leq q_{1-\alpha_1}\} \tag{10}$$

where $\alpha = \alpha_1 + \alpha_2$. For example, if the desired confidence level is 95% where $\alpha = 0 \cdot 05$, we can set $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$.

Given α and $\alpha = \alpha_1 + \alpha_2$, Theorem 2 shows that $\tilde{C}_{1-\alpha}$ achieves the desired $1 - \alpha$ coverage in the presence of possibly invalid instruments.

THEOREM 2. *Suppose we have the same assumptions about the model and the test statistic* $T(\beta_0, B)$ *as in Theorem* 1. *For any pretest* $S(B)$ *that has the correct size under the null hypothesis that* B *contains only valid instruments,* $P_{1-\alpha}(Y, D, Z)$ *always has at least* $1 - \alpha$ *coverage even with invalid instruments.*

Proof of Theorem 2. Consider $\tilde{B} \subseteq B^*$ where $c(\tilde{B}) = L - U + 1$. Since $S(B)$ has the correct size, under the null hypothesis, $\text{pr}\{S(\tilde{B}) \geq q_{1-\alpha_1}\} \leq \alpha_1$. Then, for $S(B)$ and $T(\beta_0, B)$, we can use Bonferroni's inequality to obtain

$$\begin{aligned}\text{pr}\{\beta^* \in P_{1-\alpha}(Y, D, Z)\} &\geq \text{pr}\{\beta^* \in C_{1-\alpha_2}(Y, D, Z, \tilde{B}) \cap S(\tilde{B}) \leq q_{1-\alpha_1}\} \\ &\geq 1 - \text{pr}\{\beta^* \notin C_{1-\alpha_2}(Y, D, Z, \tilde{B})\} - \text{pr}\{S(\tilde{B}) \geq q_{1-\alpha_1}\} \\ &= 1 - \alpha_1 - \alpha_2 = 1 - \alpha\end{aligned}$$

thereby guaranteeing the correct coverage. □

Similar to Theorem 1, the procedure in (10) is general in the sense that any pretest $S(B)$ with the correct size under the null hypothesis that B contains only valid instruments will guarantee that the pretest confidence interval $P_{1-\alpha}(Y, D, Z)$ will have the desired level of coverage.

The most well-known pretest is the Sargan test for overidentification (Sargan, 1958), which tests, among other things, whether the instruments B contain only valid instruments (Dufour, 2003). The Sargan test is

$$\mathrm{SAR}(B) = \frac{\|P_{\tilde{Z}_B}\hat{u}(B)\|_2^2}{\|\hat{u}(B)\|_2^2/n} \tag{11}$$

where the $\hat{u}(B)$ corresponds to the residual from the two-stage least squares estimator in (5) and $\tilde{Z}_B = R_{Z_A} Z_B$. Under model (2) and the null hypothesis that Z_B is independent of ϵ_i, $\mathrm{SAR}(B)$ converges to a $\chi^2_{c(B)-1}$ distribution. In other words, as long as B contains a set of valid instruments, $S(B)$ converges to a $\chi^2_{c(B)-1}$. Thus, if we use the Sargan test as a pretest for $P_{1-\alpha}(Y, D, Z)$, then $q_{1-\alpha_1}$ in (10) would be the $1-\alpha_1$ quantile of a $\chi^2_{c(B)-1}$ distribution and we would only proceed to construct a confidence interval with the test statistic $T(\beta_0, B)$ at $1-\alpha_2$ if the null hypothesis is retained.

2·7. *Prior information about s and U*

Throughout our discussion, we used the $U = L/2$ upper bound, that is given L candidates, less than 50% are invalid, out of simplicity along with the fact that at $U \leq L/2$, the parameters in our model (2) are always identifiable (Kang et al., 2015). However, in practice, practitioners may be able to use their subject matter knowledge to assume a smaller upper bound on the number of invalid instruments and we want to be able to incorporate this information into our confidence interval procedures. By having a tighter upper bound on s by U than $U = L/2$, our methods in (4) and (10) are only left with smaller number of subsets of possibly valid instruments to go through. Specifically, in (4), we take less unions over possibly unnecessary intervals and this provides more informative intervals. In (10), having a tighter bound on s translates to doing fewer pretests and having less subsets to take unions of, again providing more informative intervals. In Section 3·2, we examine the effect of having more prior information about s via U on our methods producing more informative intervals through a simulation study.

3. SIMULATION STUDY

3·1. *Robustness with invalid instruments*

We first compare in the simulation study the robustness of our method compared to popular methods for confidence intervals in the instrumental variables literature when there are concerns for invalid instruments.

The simulation setup is similar to the traditional single-equation linear models. We have $n = 5000$ individuals with $L = 10$ candidate instruments where each pair of instruments are correlated with correlation 0.6. For the data generating model, we assume the model in (2) and a linear model between D_i and $Z_{i\cdot}$, specifically $D_i = Z_i^T\gamma^* + \xi_i$ where ϵ_i, ξ_i are either (i) independent and bivariate Normal with mean 0, marginal variance 1, and correlation 0.99, (ii) bivariate t with 3 degrees of freedom and the same moments as (i) and (iii) where the log of the error terms are bivariate Normal with the same moments as (i) so that the error distributions are skewed; note that the individuals $i = 1, \ldots, n$ are independent. We vary the number of invalid instruments s from 0 to 5. We consider the setting where less than 50% of the instruments are invalid since β^* is always identified under this case (Kang et al., 2015). We set γ^* based on the concentration parameter, which is the expected value of the F statistic for the coefficients Z_{B^*} in the regression of D and Z and is a measure of instrument strength (Stock et al., 2002). Specifically, γ^* is set

so that either (i) the instruments are strong with a concentration parameter above 1000 or (ii) the instruments are weak with a concentration parameter below 10.

We compare our methods in (4) and (10) to "naive" methods and "oracles." Naive methods are methods that assume all candidate instruments are valid, which is typically done in practice; we use the four tests described in Section 2·4, specifically the two-stage least squares test in (5), the Anderson-Rubin test in (6), the Lagrange multiplier test in (7), and the conditional likelihood ratio test in (8), all with $B = \{1, \ldots, L\}$ (Murray, 2006). Oracles correspond to knowing exactly which instruments are valid and invalid and using the four procedures with $B = B^*$; these methods typically cannot be used in practice because of the incomplete knowledge about exactly which instruments are invalid versus valid. Also, for our methods involving pretests in (10), we use the Sargan test as the pretests for the two-stage least squares test and the conditional likelihood ratio test, both at level $\alpha_1 = 0.01$ for the pretest, and $\alpha_2 = 0.04$ for the subsequent tests. We do not use the pretesting method for the Anderson-Rubin test since the test produces informative intervals by encouraging empty intervals for subsets B that contain invalid instruments (see Section 2·5). We repeat the simulation 1000 times for each setting. For interpretability, among all our methods, we take the convex hull of the union of confidence intervals to obtain non-disjoint intervals.

Tables 1, 2, and 3 show the coverage proportion when we vary s and assume that at most 50% of the instruments are invalid, $U = L/2 = 5$, for the bivariate Normal, the bivariate t, and the skewed errors, respectively. When there are no invalid instruments, $s = 0$, and the instruments are strong, the naive procedures have the desired 95% coverage. Our methods have higher than 95% coverage because they need to overcompensate to allow for the possibly that not all candidate instruments are valid. When the instruments are weak and there are no invalid instruments, $s = 0$, any procedure using two-stage least squares undercovers, which is to be expected from the literature on two stage least squares' poor performance in the presence of weak instruments (see references in Section 2·4). As the number of invalid instruments, s, increases, regardless of the strength of the instruments, the naive methods fail to have any coverage. The oracle methods have proper coverage, except two-stage least squares when the instruments are weak. Our methods have the desired level of coverage, with the coverage level reaching nominal levels when s is at the boundary of $s < U$, i.e. $s = 4$. The only notable exceptions to our methods having correct coverage are in the presence of weak instruments when the two-stage least squares t-test are used as test statistics or the Sargan test is used as a pretest. This is not surprising because the two-stage least squares t-test and Sargan's tea are known to have actual Type I error rate that can differ greatly from the nominal Type I error rate in the presence of weak instruments Staiger & Stock (1997). The simulations suggest that methods with pretests are only useful when the instruments are sufficiently strong. By contrast, our method using Anderson-Rubin's test, which doesn't use a pretest, is valid regardless of the strength of the instruments.

In short, in the presence of possibly invalid instruments, the naive, popular approach of simply assuming all the instruments are valid would lead to misleading inference. In contrast, our methods, especially the method in (4), provide honest coverage regardless of whether instruments are invalid or valid (as long as the number of invalid instruments is less than the assumed upper bound U) and should be used whenever there is concern for possibly invalid instruments. In particular, (4) works regardless of the strength of the instruments while our method in (10) provides a desired level of coverage so long as the instruments are strong.

3·2. *Informative intervals and median length*

While our methods provide the desired level of coverage, both theoretically and in simulation, it is unclear whether the resulting robust intervals would be informative in terms of not being too

Table 1. *Comparison of coverage between 95% confidence intervals under Normal errors*

Strength	Case	Test	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
Strong	Naive	TSLS	94	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	96
		AR	100	100	100	100	95
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	100	100	100	100	94
		SAR + CLR	100	100	100	100	95
	Oracle	TSLS	94	95	94	95	94
		AR	95	96	95	95	95
		LM	98	98	97	97	97
		CLR	95	95	94	95	94
Weak	Naive	TSLS	5	0	0	0	0
		AR	96	0	0	0	0
		LM	98	0	0	0	0
		CLR	98	0	0	0	0
	Our method	TSLS	30	43	39	30	17
		AR	100	100	100	100	96
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	31	44	41	32	18
		SAR + CLR	100	100	98	91	56
	Oracle	TSLS	5	7	10	13	17
		AR	96	96	96	96	96
		LM	98	97	97	97	97
		CLR	98	97	97	97	97

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 2%.

long. It is expected that our methods will produce longer confidence intervals than oracles since the oracles know more about instrument validity than our methods assumes. In this section, we quantify this difference through a simulation study.

The first simulation setup is identical to Section 3·1 and we look at the median length of the confidence intervals in Table 1. We exclude the naive methods since they do not provide the desired level of coverage. Also, for weak instruments, we exclude two-stage least squares since it is not robust to weak instruments and does not provide correct coverage.

In Tables 4 and 5, for both bivariate Normal errors and bivariate t errors, respective, we see that the discrepancy between our method and the oracles shrink as s grows for strong instruments, especially when $s = 3$ and $s = 4$. The one notable exception is our method using two stage least squares, which still have wide intervals as s increases. We also find that our method using pretests tends to provide the shortest intervals among the various versions of our method under the strong instrument case. This is to be expected since the motivation for the pretesting was to remove taking unnecessary unions of intervals in (10). For weak instruments, our method and the oracles

Table 2. *Comparison of coverage between 95% confidence intervals under bivariate t errors*

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Naive	TSLS	95	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	97
		AR	100	100	100	100	96
		LM	100	100	100	100	98
		CLR	100	100	100	100	98
		SAR + TSLS	100	100	100	100	96
		SAR + CLR	100	100	100	100	96
	Oracle	TSLS	95	95	95	96	96
		AR	95	96	95	96	96
		LM	98	98	97	98	98
		CLR	95	95	96	96	95
Weak	Naive	TSLS	5	0	0	0	0
		AR	96	0	0	0	0
		LM	98	0	0	0	0
		CLR	98	0	0	0	0
	Our method	TSLS	30	45	41	32	16
		AR	100	100	100	100	97
		LM	100	100	100	100	98
		CLR	100	100	100	100	98
		SAR + TSLS	31	47	45	34	17
		SAR + CLR	100	100	98	90	58
	Oracle	TSLS	5	6	8	12	15
		AR	96	96	96	96	96
		LM	98	97	98	97	98
		CLR	98	97	98	97	98

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 1%.

are generally in agreement by providing infinite length intervals, with our method almost always producing infinite length intervals. This agreement is to be expected since using tests that are robust to weak instruments must produce infinite intervals (Dufour, 1997).

Table 6 presents the same simulation results as Tables 4 and 5, except the errors are skewed. While the patterns of simulations are mostly the same as the two preceding tables, one notable exception is when the instruments are strong and $s = 0$. In this case, two-stage least squares dominates our pretesting method as well as the Anderson and Rubin confidence intervals. Otherwise, the patterns of the simulations are similar across the three Tables.

The second simulation study examines the strategy in Section 2·7 where prior information on s and U are available and whether the prior information provides informative intervals. The simulation setup is, again, identical as above, except we fix $s = 2$ and vary U from $3, 4$ and 5; if U were to be less than s where $U \leq s$, our methods cannot produce the right coverage since the U was mis-specified. Here, the distribution of the error term is Normal. We compare our methods to the oracle intervals in Table 4, specifically the column corresponding to $s = 2$.

Table 3. *Comparison of coverage between 95% confidence intervals under skewed errors*

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Naive	TSLS	94	0	0	0	0
		AR	95	0	0	0	0
		LM	98	0	0	0	0
		CLR	95	0	0	0	0
	Our method	TSLS	100	100	100	100	95
		AR	100	100	100	100	95
		LM	100	100	100	100	97
		CLR	100	100	100	100	97
		SAR + TSLS	100	100	100	100	94
		SAR + CLR	100	100	100	100	94
	Oracle	TSLS	94	94	94	93	94
		AR	95	95	94	94	95
		LM	98	97	97	97	97
		CLR	95	94	94	94	94
Weak	Naive	TSLS	0	0	0	0	0
		AR	96	45	1	0	0
		LM	98	15	0	0	0
		CLR	97	15	0	0	0
	Our method	TSLS	17	60	60	48	26
		AR	100	100	100	100	99
		LM	100	100	100	100	100
		CLR	100	100	100	100	100
		SAR + TSLS	18	55	56	48	25
		SAR + CLR	100	100	100	100	89
	Oracle	TSLS	0	0	0	1	2
		AR	96	96	96	95	96
		LM	98	97	96	96	96
		CLR	97	97	96	96	96

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L = 10$ candidate instruments and U is set to $L/2 = 5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The standard error for all the coverage proportions do not exceed 2%.

Table 7 shows the result from the simulation. We see that if U is close to the true $s = 2$, our interval lengths are very close to the oracle intervals in Table 4 for strong instruments. Again, the notable exception is our method using two-stage least squares which produces wide intervals. As U increases, our methods tend to produce longer intervals, which is expected since our prior information about s at $U = 5$ is not as accurate as when $U = 3$. Also, similar to Table 4, our method with pretesting seems to produce the most informative interval compared to our method without the pretest for strong instruments. For weak instruments, our intervals produce the same type of non-informative intervals as the oracle intervals in Table 4. In this case, prior information on U does not help because the instruments are already weak and no extra information can be gained by having more accurate ideas about s.

Table 4. *Comparison of median lengths between different 95% confidence intervals under Normal errors*

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Our method	TSLS	$0\cdot28$	$0\cdot73$	$0\cdot59$	$0\cdot51$	$0\cdot44$
		AR	$0\cdot38$	$0\cdot22$	$0\cdot15$	$0\cdot11$	$0\cdot07$
		LM	$1\cdot18$	$1\cdot13$	$1\cdot09$	$1\cdot07$	$1\cdot05$
		CLR	$0\cdot29$	$0\cdot67$	$0\cdot58$	$0\cdot50$	$0\cdot44$
		SAR + TSLS	$0\cdot29$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
		SAR + CLR	$0\cdot29$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
	Oracle	TSLS	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
		AR	$0\cdot06$	$0\cdot06$	$0\cdot06$	$0\cdot07$	$0\cdot07$
		LM	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot04$
		CLR	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		SAR + CLR	∞	∞	∞	∞	$46\cdot12$
	Oracle	AR	∞	∞	∞	∞	∞
		LM	$10\cdot22$	$18\cdot79$	∞	∞	∞
		CLR	$9\cdot45$	$17\cdot97$	∞	∞	∞

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L=10$ candidate instruments and U is set to $L/2=5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0\cdot05$ and $0\cdot02$, respectively. The interquartile range of all weak intervals are infinite.

4. Data Analysis

We reanalyze the instrumental variables analysis done in Bouis & Haddad (1990), Bouis & Haddad (1992), and Small (2007) to demonstrate our method in a practical setting. The goal is to study the causal effect of income on food expenditures among Philippine farm households from a survey of $n=406$ Philippine farm households. The exposure is the household's log income, D_i and the outcome is the household's food expenditures, Y_i. We have four candidate instruments, cultivated area per capita, Z_{i1}, worth of assets, Z_{i2}, a binary dummy variable on presence of electricity at the household, Z_{i3}, and quality of flooring at the house, Z_{i4}. Page 82 of Bouis & Haddad (1990) states that the reasoning behind proposing these variables as instrumental variables is that "land availability is assumed to be a constraint in the short run, and therefore exogenous to the household decision making process". We also control for the measured covariates, which are mother's education, father's education, mother's age, father's age, mother's nutritional knowledge, price of corn, price of rice, population density of the municipality, and number of household members in adult equivalents; see Bouis & Haddad (1990) and Bouis & Haddad (1992) for further details on the data.

The F-statistic for instrument strength is 103·77, indicating reasonably strong instruments. The Sargan test for overidentification, which tests assumptions (A2) and (A3), produces a p-value of 0·079. Even though the p-value is low, usually practitioners of the instrumental variables method would naively assume (A2) and (A3) are true since the p-value is above 0·0.5, the typical threshold for significance level and use one of the four procedures in (5)–(8) naively to obtain confidence intervals. In contrast, our methods do not take for granted that the four instruments are valid; instead, we assume there may be no more than one invalid instrument (i.e. U is assumed

Table 5. *Comparison of median lengths between different 95% confidence intervals under bivarate t errors*

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Our method	TSLS	$0\cdot28$	$0\cdot73$	$0\cdot58$	$0\cdot50$	$0\cdot44$
		AR	$0\cdot37$	$0\cdot22$	$0\cdot15$	$0\cdot11$	$0\cdot07$
		LM	$1\cdot17$	$1\cdot13$	$1\cdot09$	$1\cdot07$	$1\cdot05$
		CLR	$0\cdot28$	$0\cdot67$	$0\cdot58$	$0\cdot50$	$0\cdot44$
		SAR + TSLS	$0\cdot28$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
		SAR + CLR	$0\cdot29$	$0\cdot17$	$0\cdot12$	$0\cdot08$	$0\cdot05$
	Oracle	TSLS	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
		AR	$0\cdot06$	$0\cdot06$	$0\cdot07$	$0\cdot07$	$0\cdot07$
		LM	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot03$	$1\cdot04$
		CLR	$0\cdot04$	$0\cdot04$	$0\cdot04$	$0\cdot05$	$0\cdot05$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		SAR + CLR	∞	∞	∞	∞	$46\cdot53$
	Oracle	AR	∞	∞	∞	∞	∞
		LM	$9\cdot40$	$15\cdot34$	$130\cdot38$	∞	∞
		CLR	$8\cdot98$	$14\cdot11$	$167\cdot52$	∞	∞

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L=10$ candidate instruments and U is set to $L/2=5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0\cdot05$ and $0\cdot02$, respectively.

to be 50% of the total number of instruments, $U=L/2=2$). The results from both the naive method and our methods are in Table 7. For tests that produced multiple, disjoint intervals, we took the lowermost and uppermost values of all the confidence intervals (i.e. the convex hull) to obtain a non-disjoint confidence interval. Also, for procedures with pretests, we used the same α_1 and α_2 threshold as we did in the simulations in Section 3·1.

As long as the modeling assumption is true and that no more than one instruments are invalid, we have a theoretical guarantee that our methods provide the correct 95% confidence interval, which cannot be said for the four naive intervals in Table 7. Also, even though our confidence interval is longer than the the naive intervals, it is still informative in the sense that most of our intervals do not contain $\beta^*=0$ and therefore, the null hypothesis of no causal effect can be rejected at the usual 5% significance level. The notable exception is the confidence interval based on the Lagrange multiplier test without any pretests. For this test, both the naive method and the method based on (4) contain zero. Among the intervals that are theoretically guaranteed to have $1-\alpha$ coverage, our method in (4) using the Anderson–Rubin provides the shortest interval.

The data example illustrates the usefulness of our procedure whenever there is a concern for invalid instruments in practice. Our procedures yield confidence intervals that are honest with respect to coverage and can be informative.

5. DISCUSSION

This paper proposes a simple and general method to construct robust confidence intervals for causal effects using instrumental variables estimates when the instruments are possibly invalid, with theoretical guarantees with respect to coverage. We propose two methods in (4) and (10),

Table 6. *Comparison of median lengths between different 95% confidence intervals under skewed errors*

Strength	Case	Test	$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
Strong	Our method	TSLS	$0\cdot62$	$0\cdot84$	$0\cdot66$	$0\cdot56$	$0\cdot47$
		AR	$0\cdot94$	$0\cdot50$	$0\cdot34$	$0\cdot24$	$0\cdot16$
		LM	$1\cdot46$	$1\cdot28$	$1\cdot19$	$1\cdot14$	$1\cdot10$
		CLR	$0\cdot67$	$0\cdot81$	$0\cdot66$	$0\cdot56$	$0\cdot48$
		SAR + TSLS	$0\cdot64$	$0\cdot37$	$0\cdot25$	$0\cdot18$	$0\cdot11$
		SAR + CLR	$0\cdot69$	$0\cdot38$	$0\cdot26$	$0\cdot18$	$0\cdot11$
	Oracle	TSLS	$0\cdot08$	$0\cdot09$	$0\cdot09$	$0\cdot10$	$0\cdot11$
		AR	$0\cdot14$	$0\cdot14$	$0\cdot15$	$0\cdot15$	$0\cdot16$
		LM	$1\cdot06$	$1\cdot06$	$1\cdot07$	$1\cdot07$	$1\cdot07$
		CLR	$0\cdot08$	$0\cdot09$	$0\cdot09$	$0\cdot10$	$0\cdot11$
Weak	Our method	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞
		SAR + CLR	∞	∞	∞	∞	∞
	Oracle	AR	∞	∞	∞	∞	∞
		LM	∞	∞	∞	∞	∞
		CLR	∞	∞	∞	∞	∞

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L=10$ candidate instruments and U is set to $L/2=5$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals and strong oracle intervals do not exceed $0\cdot20$ and $0\cdot05$, respectively. The interquartile range of all weak intervals are infinite.

Table 7. *Comparison of median lengths between different 95% confidence intervals with prior information on s and U*

Strength	Case	Test	$U=3$	$U=4$	$U=5$
Strong	Our method	TSLS	$0\cdot51$	$0\cdot55$	$0\cdot59$
		AR	$0\cdot07$	$0\cdot11$	$0\cdot15$
		LM	$1\cdot05$	$1\cdot07$	$1\cdot09$
		CLR	$0\cdot50$	$0\cdot54$	$0\cdot58$
		SAR + TSLS	$0\cdot05$	$0\cdot08$	$0\cdot12$
		SAR + CLR	$0\cdot04$	$0\cdot08$	$0\cdot12$
Weak	Our method	AR	∞	∞	∞
		LM	∞	∞	∞
		CLR	∞	∞	∞
		SAR + CLR	$59\cdot73$	∞	∞

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are $L=10$ candidate instruments and $s=2$. Strong instruments correspond to concentration parameter exceeding 100. Weak instruments correspond to concentration parameter value around 2. The interquartile range of our intervals do not exceed $0\cdot02$. The interquartile range of all weak intervals are infinite except for SAR + CLR ($U=3$), which range from $160\cdot55$ ($U=3$) to $42\cdot75$ ($U=5$).

Table 8. *Comparison of median lengths between different 95% confidence intervals*

Case	Test	95% Confidence Interval
Naive	TSLS	(0 · 043, 0 · 053)
	AR	(0 · 044, 0 · 054)
	LM	(−0 · 031, 0 · 055)
	CLR	(0 · 043, 0 · 055)
Our Method	TSLS	(0 · 031, 0 · 059)
	AR	(0 · 037, 0 · 058)
	LM	(−0 · 037, 0 · 067)
	CLR	(0 · 034, 0 · 066)
	SAR + TSLS	(0 · 031, 0 · 058)
	SAR + CLR	(0 · 034, 0 · 066)

TSLS, two stage least squares; AR, Anderson–Rubin test ; LM, Lagrange multiplier test; CLR, conditional likelihood ratio test; SAR, Sargan test. There are four candidate instruments and we assume that at most one is invalid.

with the latter using pretests tending to produce informative intervals when the instruments are strong. Our data analysis example illustrates that our method can be a robust alternative to confidence interval estimation that has the proper coverage whenever there is concern for possibly invalid instruments.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional test statistics as pretests for the confidence interval.

REFERENCES

ANDERSON, T. W. & RUBIN, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20**, 46–63.

ANDREWS, D. W. K., MOREIRA, M. J. & STOCK, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* **74**, 715–752.

ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**, 444–455.

BAIOCCHI, M., CHENG, J. & SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine* **33**, 2297–2340.

BOUIS, H. E. & HADDAD, L. J. (1990). *Agricultural Commercialization, Nutrition, and the Rural Poor*. Lynne Rienner Publishers.

BOUIS, H. E. & HADDAD, L. J. (1992). Are estimates of calorie-income elasticities too high?: A recalibration of the plausible range. *Journal of Development Economics* **39**, 333–364.

BURGESS, S., BUTTERWORTH, A., MALARSTIG, A. & THOMPSON, S. G. (2012). Use of mendelian randomisation to assess potential benefit of clinical intervention. *British Medical Journal* **345**.

DAVEY SMITH, G. & EBRAHIM, S. (2003). mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1–22.

DAVEY SMITH, G. & EBRAHIM, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30–42.

DAVIDSON, R. & MACKINNON, J. G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

DUFOUR, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* , 1365–1387.

DUFOUR, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *The Canadian Journal of Economics / Revue canadienne d'Economique* **36**, 767–808.

HOLLAND, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology* **18**, 449–484.

KADANE, J. B. & ANDERSON, T. W. (1977). A comment on the test of overidentifying restrictions. *Econometrica* **45**, 1027–1031.

KANG, H., ZHANG, A., CAI, T. T. & SMALL, D. S. (2015). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* , to appear.

KATAN, M. B. (1986). Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet* **327**, 507–508.

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

KOLESÁR, M., CHETTY, R., FRIEDMAN, J. N., GLAESER, E. L. & IMBENS, G. W. (2013). Identification and inference with many invalid instruments. *National Bureau of Economic Research* , No. w17519.

LAWLOR, D. A., HARBORD, R. M., STERNE, J. A. C., TIMPSON, N. & DAVEY SMITH, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163.

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.

MURRAY, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives* **20**, 111–132.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688.

SARGAN, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* , 393–415.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102**, 1049–1058.

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. & SMOLLER, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483–495.

STAIGER, D. & STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65**, 557–586.

STOCK, J. H., WRIGHT, J. H. & YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* **20**.

TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* **101**, 1607–1618.

WILSON, P. W. F., MYERS, R. H., LARSON, M. G., ORDOVAS, J. M., WOLF, P. A. & SCHAEFER, E. J. (1994). Apolipoprotein e alleles, dyslipidemia, and coronary heart disease: the framingham offspring study. *Journal of the American Medical Association* **272**, 1666–1671.

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed.