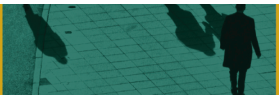
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 4, *pp.* 1231–1254

Inference for instrumental variables: a randomization inference approach

Hyunseung Kang,
University of Wisconsin—Madison, USA

Laura Peck
ABT Associates, Bethesda, USA

and Luke Keele
Georgetown University, Washington DC, USA

[Received April 2017. Revised December 2017]

Summary. The method of instrumental variables provides a framework to study causal effects in both randomized experiments with non-compliance and in observational studies where natural circumstances produce as if random nudges to accept treatment. Traditionally, inference for instrumental variables relied on asymptotic approximations of the distribution of the Wald estimator or two-stage least squares, often with structural modelling assumptions and/or moment conditions. We utilize the randomization inference approach to instrumental variables inference. First, we outline the exact method, which uses the randomized assignment of treatment in experiments as a basis for inference but lacks a closed form solution and may be computationally infeasible in many applications. We then provide an alternative to the exact method, the almost exact method, which is computationally feasible but retains the advantages of the exact method. We also review asymptotic methods of inference, including those associated with two-stage least squares, and analytically compare them with randomization inference methods. We also perform additional comparisons by using a set of simulations. We conclude with three different applications from the social sciences.

Keywords: Effect ratio; Exclusion restriction; Instrumental variables; Randomization inference; Weak instrument

1. Introduction

1.1. *Instrumental variables: a general causal method*

Many randomized trials suffer from non-compliance where subjects fail to comply with his or her assigned treatment status. Although analysts can focus on the causal effect of the treatment assignment on the outcome in an intention-to-treat (ITT) analysis, there is also substantial interest in the causal effect of the treatment that is actually received, especially in the social sciences and public policy or programme evaluations. For example, as part of a comprehensive economic stimulus package funded under the 2009 American Recovery and Reinvestment Act, the US Department of Labor awarded a series of grants across the USA to promote training for employment in energy efficiency, renewable energy and healthcare, and participants were randomized to either participation in the new training programmes, i.e. treatment, or to the

Address for correspondence: Luke Keele, McCourt School of Public Policy, Georgetown University, 304 Old North, Washington DC 20057, USA.
E-mail: lk681@georgetown.edu

0964–1998/18/1811231

existing available training programmes—control; see Section 7.1 for details. However, some trainees who were assigned to the new training initiatives selected not to participate, creating issues of non-compliance, and the 'Green jobs and health care (GJHC) impact evaluation' (Copson *et al.*, 2015; Martinson *et al.*, 2015) wanted to evaluate the actual efficacy of the training programmes, i.e. whether those who actually received the training programmes had any influence on labour outcomes like earnings and employment outcomes.

When non-compliance is present, substantial progress can be made by using the treatment assignment as an instrumental variable (IV), which is a variable that affects exposure to treatment but does not directly affect the outcome (Angrist *et al.*, 1996; Hernán and Robins, 2006); see Section 2.2 for details about IV and related assumptions. These identifying assumptions enable us to estimate the complier average causal effect, which is the average causal effect among subpopulations of individuals who comply with the treatment assignment, or the average treatment effect among the treated. However, IV analysis is not confined to randomized trials with non-compliance. In some applications, the instrument is some naturally occurring nudge to accept a treatment and is characterized as a type of natural experiment. In observational studies, IV analysis may provide additional insights about the causal effects in the presence of unobserved confounding (Kang, 2016). As such, IV analysis is utilized in many disciplines including economics (Angrist and Krueger, 2001; Imbens, 2014), epidemiology (Hernán and Robins, 2006; Baiocchi *et al.*, 2014), political science (Hansford and Gomez, 2010; Keele and Morgan, 2016) and Mendelian randomization (MR) studies (Davey Smith and Ebrahim, 2003, 2004; Lawlor *et al.*, 2008) where the instruments are genetic variants.

1.2. Traditional approaches to inference with instrumental variables

Once identifying assumptions have been accepted, point estimation of the casual effect is typically based on the Wald estimator (Wald, 1940). In the case of randomized experiments with a binary treatment assigned indicator or IV, the Wald estimator is equivalent to the popular two-stage least squares (TSLS) estimator in the IV literature (Angrist and Pischke, 2008); see Section 2.3 for details. For inference, such as confidence intervals, traditional methods rely on econometric theory by using structural models, moment assumptions and/or asymptotic approximations (Angrist and Pischke, 2008; Wooldridge, 2010). For example, under the three moment restrictions on the error terms in a structural model, specifically models 2SLS.1, 2SLS.2 and 2SLS.3 in chapter 5 of Wooldridge (2010), the TSLS estimator is asymptotically normal. The asymptotic normality can be used to derive inferential quantities like p-values and confidence intervals and is the default mode of inference in most statistical software.

Despite the simplicity and familiarity of the Wald approach to inference, especially given its ties to standard asymptotically normal-based testing procedures, inference for IV methods is complicated by the fact that statistical uncertainty depends not only on the sample size, but also on the magnitude of the effect of the instrument on the treatment, which is commonly known as the strength of an instrument. Generally speaking, when the instrument has a large effect on the treatment, it is known as a strong instrument, whereas a small effect is known as a weak instrument. It is well known in the literature that, even in large 'asymptotic' samples, confidence intervals based on TSLS will have incorrect coverage when the instrument is weak (Nelson and Startz, 1990; Bound *et al.*, 1995; Staiger and Stock, 1997; Dufour, 1997) and some proposals have been suggested to remedy this problem (Anderson and Rubin, 1949; Zivot *et al.*, 1998; Wang and Zivot, 1998; Kleibergen, 2002; Moreira, 2003). Unfortunately, these results are derived from structural models, moment assumptions and/or asymptotic approximations. For example, the conditional likelihood ratio approach that was proposed by

Moreira (2003) relies on a specification of a structural model as well as normal structural errors (for finite sample behaviour) or asymptotic moment conditions on the errors (for asymptotic behaviour). Most recently, in MR where weak instruments are common, several proposals have been suggested (Burgess *et al.*, 2011; Burgess and Thompson, 2011; Pierce *et al.*, 2011) to remedy the bias that is induced by weak instruments, most notably by finding additional instruments, combining them to a single score or instrument, and using the Wald approach to inference. Burgess and Thompson (2012) also suggested using a Bayesian approach to deal with weak instruments.

1.3. Our contributions

Given the complexities of inference with IV methods, this paper reviews and extends an alternative unified framework for inference that is based solely on the assumed instrument assignment mechanism. Specifically, we discuss exact and almost exact inferential methods for IV. The exact method uses randomization as the 'reasoned basis for inference' (Fisher, 1935) and mirrors the original design of a randomized experiment. The exact method produces an honest confidence interval for the target causal parameter even when the causal effect of the instrument on the treatment is weak (Imbens and Rosenbaum, 2005; Keele *et al.*, 2017). Also, unlike the aforementioned standard methods based on large sample normal approximations which assume that the participants are a random sample from the target population, the exact method is finite sample based and makes it explicit that further assumptions are required to generalize the IV estimand to other populations.

The second method, the almost exact method, behaves like the exact method, except that it avoids the computationally intensive nature of the exact method in larger sample sizes. The almost exact method is also motivated by the design of the instrument assignment mechanism and is based on finite sample asymptotics (Hájek, 1960). By doing so, the almost exact method preserves the properties of the exact method but is computationally tractable with a closed form expression. We also discuss extensions to both methods, including adjustment of pretreatment covariates, choice of test statistics, multiple treatment and sensitivity analysis. Importantly, we do not rely on the assumption of a constant treatment effect, which is typical in randomization-based inference; see Sections 3.1 and 3.2.

Next, we analytically and numerically compare these randomization-based methods with traditional methods including the TSLS estimator. We show how popular inferential methods in IVs can be derived by using our framework and are special cases of our randomization inference approach. We also show that the traditional approaches are, in essence, approximations of the exact method with varying degrees of accuracy and complexity. We, then, highlight the strengths of the exact and almost exact approach in three empirical applications. Two of the applications are based on randomized trials with non-compliance, and the third is an observational study based on a natural experiment. Finally, we make R code available on line from `http://wileyonlinelibrary.com/journal/rss-datasets` for others to use.

The comparison of exact and almost exact methods to traditional approaches to inference reveals the strengths of the randomization–inference approach in these settings, which are that

(a) the basic assumptions of inference are transparent and based on the study design, not on structural assumptions and assumptions about moments of structural error terms,
(b) these methods always provide honest inference in the form of correct type I error control and
(c) they make it clear where traditional assumptions like homoscedastic errors and/or large n play a critical role in inference for the causal effect.

2. Framework: notation, assumptions and estimators

2.1. Notation

Suppose that there are n individuals in an experiment indexed by $i=1,\ldots,n$. Let Y_i denote the outcome, D_i denote the treatment (or treatment received) and Z_i denote a binary instrument (or treatment assignment indicator). For each individual i, we observe the triplet (Y_i, D_i, Z_i). Also, for each individual i, let $Y_i^{(z,d)}$ be the potential outcome of the outcome given the instrument value $z \in \{0,1\}$ and treatment value $d \in \{0,1\}$ and let $D_i^{(z)}$ be the potential outcome of the treatment given the instrument value $z \in \{0,1\}$. The relationship between the potential outcomes $Y_i^{(z,d)}$ and $D_i^{(z)}$ and the observed triplets (Y_i, D_i, Z_i) is

$$D_i = D_i^{(Z_i)} = Z_i D_i^{(1)} + (1-Z_i) D_i^{(0)},$$

$$Y_i = Y_i^{(Z_i,D_i)} = Y_i^{(Z_i,D_i^{(Z_i)})} = Z_i Y_i^{(1,D_i^{(1)})} + (1-Z_i) Y_i^{(0,D_i^{(0)})}.$$

Our notation implicitly assumes the stable unit treatment value assumption (Rubin, 1980). Let $\mathcal{F} = \{(Y_i^{(1,1)}, Y_i^{(1,0)}, Y_i^{(0,1)}, Y_i^{(0,0)}, D_i^{(1)}, D_i^{(0)}), i=1,\ldots,n\}$ denote the collection of potential outcomes for all n individuals. Also, let $0<n_1<n$ represent the number of individuals who are assigned to treatment $Z_i=1$ and $n_0=n-n_1$ represent the number of individuals who are assigned to control $Z_i=0$ where n_1 and n_0 are non-random. Let $\Omega = \{(z_1,\ldots,z_n) \in \{0,1\}^n, \Sigma_{i=1}^n z_i = n_1\}$ be the possible values that $(Z_1,\ldots,Z_n)$ can take so that, among n individuals, exactly n_1 individuals have $Z_i=1$ and the other n_0 individuals have $Z_i=0$.

Let $\mathcal{Z}$ be the event $(Z_1,\ldots,Z_n) \in \Omega$. Following our discussion about randomization as a basis for inference, the paper will focus on the finite population settings where the target of inference is a functional of $\mathcal{F}$ that is fixed and unknown. Extensions to inference based on infinite population models is possible and is detailed in chapter 6 of Imbens and Rubin (2015). Such approaches typically require additional assumptions such as that the study participants are a random sample from the target population. Here, we wish to be explicit that further assumptions will be required to generalize the causal quantities of interest to other populations.

2.2. Causal estimands and instrumental variables assumptions

Given the potential outcomes in $\mathcal{F}$, we can define the following causal estimands:

$$\tau_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,D_i^{(1)})} - Y_i^{(0,D_i^{(0)})}, \tag{1}$$

$$\tau_D = \frac{1}{n}\sum_{i=1}^{n} D_i^{(1)} - D_i^{(0)}, \tag{2}$$

$$\tau = \frac{\tau_Y}{\tau_D} = \frac{\sum_{i=1}^{n} Y_i^{(1,D_i^{(1)})} - Y_i^{(0,D_i^{(0)})}}{\sum_{i=1}^{n} D_i^{(1)} - D_i^{(0)}}. \tag{3}$$

Equation (1) is the average causal effect of the instrument on the outcome and is often referred to as the ITT effect. Equation (2) is the average causal effect of the instrument on the exposure. If, in a randomized experiment, the compliance to assigned treatment is one sided where individuals who are assigned to control cannot actually receive the treatment, $D_i^{(0)}=0$, τ_D is known as the compliance rate. Equation (3) is the ratio of the two average causal effects τ_Y and τ_D and is the causal estimand of interest where it is implicitly assumed that $\tau_D \neq 0$; see assumptions 2 and 4 below. The estimand τ is also referred to as the IV estimand in the literature.

Identification of τ requires a set of assumptions. For example, if (assumption 1) we assume ignorability of Z where $P\{(Z_1,\ldots,Z_n)=(z_1,\ldots,z_n)|\mathcal{F},\mathcal{Z}\}=P\{(Z_1,\ldots,Z_n)=(z_1,\ldots,z_n)|\mathcal{Z}\}=1/\binom{n}{n_1}$, (assumption 2) a non-zero causal effect of Z on D where $\tau_D \neq 0$, the exclusion restriction (assumption 3) where, for all d and i, $Y_i^{(d,1)}=Y_i^{(d,0)}$, and (assumption 4) monotonicity where, for all i, $D_i^{(1)} \geqslant D_i^{(0)}$, τ is identified as the complier average treatment effect, or the average treatment effect among those individuals in the experiment who complied with their treatment assignment (Angrist *et al.*, 1996). Alternatively, one can make different sets of assumptions to identify a different interpretation of τ. For example, if we make assumptions 1–3 and assume an additive structural mean model for Y, D and Z, with no effect modification assumption on Y via interactions between D and Z, τ would be the average treatment effect among treated individuals (Hernán and Robins, 2006). Alternatively, if we make only assumptions 1 and 2, τ would simply be identified as the ratio of two average treatment effects (Baiocchi *et al.*, 2010; Kang *et al.*, 2016).

Assumptions 1, 2 and 4 are typically satisfied by design of many randomized experiments, especially in public policy programme evaluations where there is often one-sided compliance. In natural experiments or observational studies, these assumptions require careful consideration. For example, assumption 1 holds by design in a randomized experiment but requires justification when the IV approach is applied to observational data. In Section 5.2, we discuss a sensitivity analysis that enables investigators to probe this assumption.

Since the focus of the paper is on inference for τ, we shall not dwell on identifying assumptions. We shall simply assume that assumptions 1–4 hold so that τ is identified as the complier average treatment effect and to allow easier comparison between our randomization inference approach and the traditional approaches to inference. But, we stress that these assumptions are actually more stringent than is necessary for inference on τ under the approach that we use. In particular, in Section 3, we shall show that the exact and the almost exact methods rely on assumption 1 only and are robust to near violations of assumption 2. Inferences from both methods remain valid even when assumption 3 and/or 4 may not hold. For additional discussions on these assumptions, see the on-line supplementary materials, Angrist *et al.* (1996), Hernán and Robins (2006), Deaton (2010), Imbens (2010, 2014), Baiocchi *et al.* (2014) and Swanson and Hernán (2014).

2.3. Point estimator for τ

To discuss inference for τ, especially comparing the randomization inference approach with the traditional approaches based on point estimators (e.g. TSLS) in Section 4, it is instructive to discuss point estimators for τ_Y and τ_D that make up τ. The most popular point estimators for τ_Y and τ_D are the difference-in-means estimators

$$\hat{\tau}_Y=\frac{1}{n_1}\sum_{i=1}^{n} Y_i Z_i-\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-Z_i), \tag{4}$$

$$\hat{\tau}_D=\frac{1}{n_1}\sum_{i=1}^{n} D_i Z_i-\frac{1}{n_0}\sum_{i=1}^{n} D_i(1-Z_i). \tag{5}$$

Standard arguments can show that $\hat{\tau}_Y$ and $\hat{\tau}_D$ are unbiased estimators of τ_Y and τ_D respectively, i.e. $E(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})=\tau_Y$ and $E(\hat{\tau}_D|\mathcal{F},\mathcal{Z})=\tau_D$ (Imbens and Rubin, 2015). Standard estimators for $\mathrm{var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$ and $\mathrm{var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})$ exist depending on the assumptions that we make about $\mathcal{F}$

(Imbens and Rubin, 2015). For now, we shall leave them unspecified and denote the estimated variances of $\hat{\tau}_Y$ and $\hat{\tau}_D$ as $\widehat{\text{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$ and $\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})$ respectively.

Given the estimators in equations (4) and (5) for τ_Y and τ_D respectively, the most natural estimator for τ would be the ratio of the estimators. Indeed, this is the most frequently used estimator for τ and is often called the 'usual' IV estimator, 'the' IV estimator or the 'Wald' estimator (Wald, 1940; Hernán and Robins, 2006; Wooldridge, 2010; Baiocchi *et al.*, 2014):

$$\hat{\tau} = \frac{\hat{\tau}_Y}{\hat{\tau}_D} = \frac{(1/n_1)\sum_{i=1}^{n} Y_i Z_i - (1/n_0)\sum_{i=1}^{n} Y_i(1-Z_i)}{(1/n_1)\sum_{i=1}^{n} D_i Z_i - (1/n_0)\sum_{i=1}^{n} D_i(1-Z_i)}. \tag{6}$$

We can also arrive at $\hat{\tau}$ by using TSLS, which is another popular point estimator for τ (Wooldridge, 2010; Baiocchi *et al.*, 2014). Specifically, if we

(a) fit a linear regression between Z_i and D_i and save the predicted D_is and
(b) fit a second linear regression between the predicted D_i and Y_i,

the coefficient that is associated with the predicted D_i from the second linear regression is $\hat{\tau}$.

3. Inference for τ by using the randomization-based approach

In this section, we summarize two methods of inference for τ, both motivated by randomization-based inference. The first method, which we call the exact approach, is guaranteed to have correct coverage. Unfortunately, it is computationally intensive for even modest sample sizes and lacks a closed form expression. We then outline an approximation to the exact method which we call the almost exact method. In both cases, we use only the design assumptions, primarily assumption 1, to derive inferential quantities like confidence intervals and p-values.

3.1. *The exact method*

One method of inference for τ is based on the randomization-based inference approach to IVs as described in Rosenbaum (1996, 2002), Imbens and Rosenbaum (2005), Baiocchi *et al.* (2010), Nolen and Hudgens (2011) and Kang *et al.* (2016a). It is also called the exact method because inference relies on exact calculations based on the distribution of Z. As we outline below, the randomization inference test of no effect can be inverted to provide distribution-free confidence intervals.

Typically, under the exact approach, inverting exact tests to derive confidence intervals is associated with the assumption that the treatment effect is constant from unit to unit out of convenience. This assumption of constant treatment effects is often critiqued as unrealistic given the likely presence of 'essential heterogeneity', where individuals who will benefit more from a given treatment are more likely to seek treatment (Heckman *et al.*, 2006). An expanding literature demonstrates how exact inference is possible without constant effect assumptions; see Rosenbaum (2001, 2003), Keele *et al.* (2017) and Ding *et al.* (2016) for examples. In what follows, we do not assume constant treatment effects.

Formally, given $\mathcal{F}$ and $\mathcal{Z}$, consider the null hypothesis $H_0:\tau=\tau_0$ which imposes structure on $\mathcal{F}$. This is a composite null hypothesis because there are several values of $\mathcal{F}$ for which the null can be true. Also, H_0 is not a sharp null hypothesis (Fisher, 1935) whereby a sharp null would allow us to infer other values of the unobserved potential outcomes. In fact, the sharp null of no ITT effect, $Y_i^{(1,D_i^{(1)})} = Y_i^{(0,D_i^{(0)})}$ for all i, implies $H_0 : \tau = 0$, but the converse is not necessarily true; there can be other values of $\mathcal{F}$ that satisfy the null hypothesis $H_0 : \tau = 0$.

Given H_0, consider the test statistic $T(\tau_0)$ of the form

$$T(\tau_0) = \frac{1}{n_1} \sum_{i=1}^{n} Z_i (Y_i - D_i \tau_0) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i)(Y_i - D_i \tau_0). \tag{7}$$

Let $Q_i(\tau_0) = (Y_i - D_i \tau_0)$, $\bar{Q}^{(1)}(\tau_0) = (1/n_1) \Sigma_{i=1}^n Z_i (Y_i - D_i \tau_0)$ and $\bar{Q}^{(0)}(\tau_0) = (1/n_1) \Sigma_{i=1}^n (1 - Z_i)(Y_i - D_i \tau_0)$. Rosenbaum (2002) called $Q_i(\tau_0)$ an adjusted response where the outcome Y_i is adjusted by the treatment actually received, D_i, based on the value of the null τ_0, i.e. $Y_i - D_i \tau_0$. Then, $\bar{Q}^{(1)}(\tau_0)$ represents the sample average of the adjusted responses $Q_i(\tau_0)$ for individuals who were assigned treatment $Z_i = 1$ and $\bar{Q}^{(0)}(\tau_0)$ represents the sample average of the adjusted responses for individuals who were assigned control $Z_i = 0$. We can also rewrite the test statistic (7) as the difference between the sample averages of the adjusted responses, i.e. $T(\tau_0) = \bar{Q}^{(1)}(\tau_0) - \bar{Q}^{(0)}(\tau_0)$.

Also consider an estimator for the variance of the test statistic $T(\tau_0)$, denoted as $S^2(\tau_0)$:

$$S^2(\tau_0) = \frac{1}{n_1(n_1 - 1)} \sum_{i=1}^{n} Z_i \{Q_i(\tau_0) - \bar{Q}^{(1)}(\tau_0)\}^2 + \frac{1}{n_0(n_0 - 1)} \sum_{i=1}^{n} (1 - Z_i) \{Q_i(\tau_0) - \bar{Q}^{(0)}(\tau_0)\}^2. \tag{8}$$

In the on-line supplementary materials, we show that, under the null hypothesis, the test statistic $T(\tau_0)$ is 0 and, hence, any deviation of $T(\tau_0)$ away from 0, positive or negative, suggests that H_0 is not true. This observation leads us to reject the null if

$$P_{H_0} \left\{ \left| \frac{T(\tau_0)}{S(\tau_0)} \right| \geqslant t | \mathcal{F}, \mathcal{Z} \right\} \tag{9}$$

is less than some prespecified threshold α; for simplicity, we assume that the rejection region is symmetric around zero under H_0. Here, t in equation (9) is the observed value of the standardized deviate $T(\tau_0)/S(\tau_0)$ and the probability distribution is under the null hypothesis. Also, we can use the duality between testing and confidence intervals to obtain confidence intervals for τ (Lehmann, 2006; Lehmann and Romano, 2008). Specifically, the exact $1 - \alpha$ confidence interval for τ would be the set of values τ_0 where

$$\left\{ \tau_0 : P_{H_0} \left\{ \left| \frac{T(\tau_0)}{S(\tau_0)} \right| \leqslant q_{1-\alpha/2} | \mathcal{F}, \mathcal{Z} \right\} \right\} \tag{10}$$

where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the null distribution of $T(\tau_0)/S(\tau_0)$. The confidence interval in equation (10) is exact, as it uses only the distribution of $\mathbf{Z}$ as outlined in assumption 1 and makes no additional distributional assumptions. Importantly, it does not assume the exclusion restriction (assumption 3) although the interpretation of τ would change if we make assumption 3; see Section 2.2 for details. The interval is also honest, as equation (10) will cover the true τ with at least $1 - \alpha$ probability in finite sample.

There are additional advantages of the inference method outlined above. First, the confidence interval for τ may be empty in length (Rosenbaum (2002), chapter 4). The confidence interval may be empty if the adjustment for outcomes is far wrong. This could happen if the instrument strongly predicts the outcome but the treatment dosage does not. Second, the confidence interval may be infinite in length if the instrument is weak. Formally, a weak instrument is an instrument Z that is weakly related to D so that τ_D is close to 0 and is a near violation of assumption 2 (Imbens and Rosenbaum, 2005). In other words, an instrument is weak when most units ignore the encouragement to take the treatment. Under randomization inference, if the instrument is weak, the interval becomes longer and perhaps even infinite in length and a long confidence interval is a warning that the instrument provides little information about the treatment. In sum, the confidence intervals under this approach provide clear warnings about the adherence to IV assumptions 2 and 3.

Despite the attractive properties of the exact confidence interval, one drawback is the lack of a closed form expression and, consequently, the computation that is required to compute the confidence interval. In particular, equation (10) requires computing $q_{1-\alpha/2}$, the quantile of the null distribution, and doing a very large grid search to find the set of values $\mathcal{F}$ under the null hypothesis $H_0 : \tau = \tau_0$. The next section describes a simpler alternative that, unlike the exact method, provides a closed form expression of the confidence interval.

3.2. The almost exact method

The almost exact approach builds on the exact method above by addressing its biggest limitation of computational infeasibility and provides a closed form expression for the confidence interval. Specifically, we can use 'finite sample asymptotics' (Hájek, 1960; Lehmann, 2004) that approximate the exact null distribution in equation (9) by considering an asymptotically stable sequence of finite populations $\mathcal{F}$. In the end, we have an asymptotic approximation to the exact confidence interval in equation (10), which leads to a closed form expression for the confidence interval based on a quadratic inequality. Hansen and Bowers (2009) considered a similar approximation by using sample theoretic arguments.

Formally, suppose that we have the following estimators for the variances and covariance of $\hat{\tau}_D$ and $\hat{\tau}_Y$:

$$\begin{aligned}
\widehat{\mathrm{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n} Z_i \left(D_i - \frac{1}{n_1} \sum_{i=1}^{n} Z_i D_i \right)^2 \\
&\quad + \frac{1}{n_0(n_0-1)} \sum_{i=1}^{n} (1-Z_i) \left\{ D_i - \frac{1}{n_0} \sum_{i=1}^{n} (1-Z_i) D_i \right\}^2 \\
\widehat{\mathrm{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n} Z_i \left(Y_i - \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i \right)^2 \\
&\quad + \frac{1}{n_0(n_0-1)} \sum_{i=1}^{n} (1-Z_i) \left\{ Y_i - \frac{1}{n_1} \sum_{i=1}^{n} (1-Z_i) Y_i \right\}^2 \\
\widehat{\mathrm{cov}}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) &= \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n} Z_i \left(Y_i - \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i \right) \left(D_i - \frac{1}{n_1} \sum_{i=1}^{n} Z_i D_i \right) \\
&\quad + \frac{1}{n_0(n_0-1)} \sum_{i=1}^{n} (1-Z_i) \left\{ Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1-Z_i) Y_i \right\} \left\{ D_i - \frac{1}{n_0} \sum_{i=1}^{n} (1-Z_i) D_i \right\}.
\end{aligned}$$

These are the usual variance estimators for the two-sample mean problems and their properties have been extensively studied under the finite sample framework; see Imbens and Rubin (2015). In particular, Imbens and Rubin (2015) recommended these variance estimators because of their simplicity and attractive properties when extended to infinite population settings.

Let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of the standard normal distribution. Let a, b and c be

$$\begin{aligned}
a &= \hat{\tau}_D^2 - z_{1-\alpha/2}^2 \widehat{\mathrm{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}), \\
b &= -2\{\hat{\tau}_D \hat{\tau}_Y - z_{1-\alpha/2}^2 \widehat{\mathrm{cov}}(\hat{\tau}_D, \hat{\tau}_Y|\mathcal{F},\mathcal{Z})\}, \\
c &= \hat{\tau}_Y^2 - z_{1-\alpha/2}^2 \widehat{\mathrm{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}).
\end{aligned}$$

In the on-line supplementary materials, we show that the exact confidence interval in equation (10) is approximately equal to solving the following quadratic inequality based on a, b and c:

$$\left\{\tau_0 : P_{H_0}\left\{\left|\frac{T(\tau_0)}{S(\tau_0)}\right| \leqslant q_{1-\alpha/2}|\mathcal{F},\mathcal{Z}\right\}\right\} \approx \{\tau_0 : a\tau_0^2 + b\tau_0 + c \leqslant 0\}. \tag{11}$$

We stress that this equivalence relation does not rely on assumptions about constant treatment effects nor the exclusion restriction 3, similarly to the exact method. The equivalence relation in (11), which we call the almost exact method, allows us easily to compute an approximation to the exact confidence intervals by using any standard quadratic inequality solver. In particular, depending on the value of a and the determinant b^2-4ac, the quadratic inequality can lead to different types of confidence intervals. As an example, if $a>0$ and $b^2-4ac>0$, which is the only case where the interval is non-empty and finite, a closed form formula for the confidence interval for τ by using the almost exact method is

$$\frac{\hat{\tau}_D\hat{\tau}_Y - z_{1-\alpha/2}^2\widehat{\text{cov}}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2 - z_{1-\alpha/2}^2\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})} \pm z_{1-\alpha/2}\frac{\surd[\hat{\Delta} + z_{1-\alpha/2}^2\{\widehat{\text{cov}}^2(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) - \widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})\widehat{\text{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})\}]}{\hat{\tau}_D^2 - z_{1-\alpha/2}^2\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})} \tag{12}$$

where

$$\hat{\Delta} = \hat{\tau}_Y^2\,\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) + \hat{\tau}_D^2\,\widehat{\text{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) - \hat{\tau}_D\hat{\tau}_Y\,\widehat{\text{cov}}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z}). \tag{13}$$

The on-line supplementary materials detail the different solutions to a quadratic equation and corresponding confidence intervals that arise from equation (11).

As we shall see in Sections 6 and 7, the approximation in equation (11) works very well, even in situations when the instrument is very weak so that $\tau_D \approx 0$ and assumption 2 is almost violated. Indeed, the almost exact method, like the exact method, produces infinite confidence intervals and this occurs if $a<0$ or, equivalently,

$$\left|\frac{\hat{\tau}_D}{\surd\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}\right| \leqslant z_{1-\alpha/2}. \tag{14}$$

Equation (14) is the t-test for testing the strength of the instrument Z's association with D under the null hypothesis $H_0:\tau_D=0$ and we retain the null of the t-test at $z_{1-\alpha/2}$; (see the supplementary materials for the technical derivations), i.e. the almost exact method produces an infinite confidence interval if we cannot reject the null that the instrument is weak, i.e. $H_0:\tau_D=0$, at the α-level. In short, the almost exact method retains the advantages of the exact method, especially with regard to a weak instrument, but has a closed form expression that can be easily computed. The only cost to this approach is that it is not exact in finite samples. However, our empirical applications in Section 7.1 show that this cost is minimal, especially in sample sizes where the exact method becomes computationally infeasible, while retaining many of the advantages of the exact method.

In sum, both the exact and the almost exact approaches to IV inference will produce infinite confidence intervals when an instrument is weak. How should analysts interpret infinite confidence intervals? Is infinite confidence some type of mistake? An infinite confidence interval is a designed feature of the exact and almost exact approaches. An infinite confidence interval is a warning that the data contain little information (Rosenbaum (2002), page 185). Specifically, it is a warning that the instrument has little effect on the exposure. In fact, the possibility of an infinite confidence interval is a necessary condition for proper inference when an instrument is weak (Dufour, 1997). Investigators may choose to narrow an infinite confidence interval by using prior information. This can be done informally by using IV methods that rely on asymptotic approximations, altering α *post hoc*, or using Markov chain Monte Carlo methods (Kleibergen and Zivot, 2003). We have no objections to this approach, so long as the role of prior information is communicated clearly to the audience. Ideally, the infinite confidence interval would be reported along with any interval by using prior information. This will communicate clearly that prior information is needed to produce a non-infinite confidence interval.

4. Comparison with traditional methods of inference

In this section, we discuss more popular modes of inference for IVs and provide a comparison with randomization-based inferential methods that were discussed previously. First, the most widely used method of inference depends on an approximation to the normal distribution and is the basis for inference using TSLS. Specifically, we simply add or subtract the standard error to the point estimate of τ, $\hat{\tau}$, which is also the TSLS estimator, to obtain a $1-\alpha$ confidence interval:

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\tau}|\mathcal{F}, \mathcal{Z})}. \tag{15}$$

The validity of expression (15) relies on $\hat{\tau}$ being approximately normally distributed with mean τ and standard moment arguments. Specifically, the following results must be true:

(a) the moment of the product of the structural error term and the instrument is 0,
(b) there is at least one instrument and
(c) the instrument is sufficiently strong

(see chapter 5 of Wooldridge (2010) for details). If these conditions hold, the asymptotic approximation in expression (15) should be accurate. This approach is the default method of inference for IVs in many econometrics textbooks (Angrist and Pischke, 2008; Wooldridge, 2010) and software (e.g. the AER R package (Kleiber and Zeileis, 2008)).

Estimating the variance of $\hat{\tau}$ in expression (15) can be done by using a variety of ways. For example, following Imbens and Rubin (2015), suppose that we assume that

$$\begin{pmatrix} \hat{\tau}_Y \\ \hat{\tau}_D \end{pmatrix} \sim N\left\{ \begin{pmatrix} \tau_Y \\ \tau_D \end{pmatrix}, \begin{pmatrix} \text{var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z}) & \text{cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) \\ \text{cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F},\mathcal{Z}) & \text{var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z}) \end{pmatrix} \right\}.$$

Then, the delta method can be used to derive an approximation of the variance of $\hat{\tau}$:

$$\text{var}(\hat{\tau}|\mathcal{F},\mathcal{Z}) \approx \frac{\text{var}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\tau_D^2} + \frac{\tau_Y^2 \,\text{var}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\tau_D^4} - \frac{2\tau_Y \,\text{cov}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F}_n,\mathcal{Z})}{\tau_D^3}. \tag{16}$$

And plugging an estimate of this variance into expression (15) results in the following $1-\alpha$ confidence interval for τ:

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\left\{ \frac{\widehat{\text{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^2} + \frac{\hat{\tau}_Y^2 \,\widehat{\text{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})}{\hat{\tau}_D^4} - \frac{2\hat{\tau}_Y \,\widehat{\text{cov}}(\hat{\tau}_Y,\hat{\tau}_D|\mathcal{F}_n,\mathcal{Z})}{\hat{\tau}_D^3} \right\}}. \tag{17}$$

With some algebra, one can show that the confidence interval in expression (17) is equivalent to the confidence interval that is used for the TSLS estimator; see the on-line supplementary materials for details. For simplicity, we shall refer to this approach as the TSLS or the delta method.

Another approach to estimating the variance in expression (15) is by treating $\hat{\tau}_D$ as fixed so that the only random component of $\hat{\tau}$ is $\hat{\tau}_Y$, i.e. assume that the effect of Z_i on D_i is known without error. This approach leads us to a variance of $\hat{\tau}$ which is the variance of $\hat{\tau}_Y$ divided by $\hat{\tau}_D$ and the resulting $1-\alpha$ confidence interval formula (15) is

$$\hat{\tau} \pm z_{1-\alpha/2} \sqrt{\left\{ \frac{\widehat{\mathrm{var}}(\hat{\tau}_Y | \mathcal{F}, \mathcal{Z})}{\hat{\tau}_D^2} \right\}}. \tag{18}$$

This simple approximation was proposed by Bloom (1984) who suggested it in the context of programme evaluation under non-compliance, and it is widely used in the programme evaluation literature judging from recent citation patterns. Econometrics, statistics and MR also utilize this approximation (Heckman *et al.*, 1998; Yang *et al.*, 2014; Bowden *et al.*, 2016); in particular, MR studies with summary data refer to this approximation as the no-measurement-error assumption (Bowden *et al.*, 2016). Hereafter, we refer to this method of inference as the Bloom method.

The two asymptotic variance estimates based on the delta method and the Bloom method are related as follows. In simple terms, the variance from the Bloom method is exactly the first term of the variance from the delta method in equation (16). More specifically, if we denote $\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Bloom}}$ as the variance estimate that is used in expression (18), i.e. $\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Bloom}} = \widehat{\mathrm{var}}(\hat{\tau}_Y | \mathcal{F}, \mathcal{Z}) / \hat{\tau}_D^2$, and $\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Delta}}$ as the variance estimate that is used in expression (17), the two variance estimates are related by a factor $C > 0$:

$$\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Delta}} = \widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Bloom}} C,$$

and C is defined as

$$C = 1 + \frac{\hat{\tau}_Y^2 \, \widehat{\mathrm{var}}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z})}{\hat{\tau}_D^2 \, \widehat{\mathrm{var}}(\hat{\tau}_Y | \mathcal{F}, \mathcal{Z})} - \frac{2 \hat{\tau}_Y \, \widehat{\mathrm{cov}}(\hat{\tau}_Y, \hat{\tau}_D | \mathcal{F}, \mathcal{Z})}{\hat{\tau}_D \, \widehat{\mathrm{var}}(\hat{\tau}_Y | \mathcal{F}, \mathcal{Z})}. \tag{19}$$

When $C > 1$, $\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Delta}}$ is larger than $\widehat{\mathrm{var}}(\hat{\tau} | \mathcal{F}, \mathcal{Z})_{\mathrm{Bloom}}$ and the delta confidence interval is larger than the Bloom confidence interval. In contrast, $C < 1$ would imply the opposite and the delta confidence interval would be smaller than the Bloom confidence interval. We can show that $C < 1$ occurs if and only if

$$|\hat{\tau}| < \left| \frac{2 \, \widehat{\mathrm{cov}}(\hat{\tau}_Y, \hat{\tau}_D | \mathcal{F}, \mathcal{Z})}{\widehat{\mathrm{var}}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z})} \right|.$$

Also, given any $\hat{\tau}_Y$, $\widehat{\mathrm{var}}(\hat{\tau}_Y | \mathcal{F}, \mathcal{Z})$ and $\widehat{\mathrm{cov}}(\hat{\tau}_Y, \hat{\tau}_D | \mathcal{F}, \mathcal{Z})$, as $|\hat{\tau}_D / \surd \widehat{\mathrm{var}}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z})|$ increases, i.e. as the instrument becomes stronger on the basis of the t-test measure of instrument strength in equation (14), or if $\hat{\tau}_D$ has little variability so that $\surd \widehat{\mathrm{var}}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z})$ decreases to 0 and the t-test increases, C will grow closer to 1. This is because the denominator in equation (19) grows larger, effectively making $C \approx 1$. Ultimately, this suggests that, for strong instruments, the difference between the two variance estimates, and consequently their respective confidence intervals in expressions (18) and (17), will be negligible.

However, we emphasize an important *caveat* for both approaches: both depend on the assumption of asymptotic normality of $\hat{\tau}$. In fact, if the instrument is weak so that $\hat{\tau}_D \approx 0$, $\hat{\tau}$ will be far from normal and the asymptotic confidence interval via expression (15) will be highly misleading: see Staiger and Stock (1997) and Stock *et al.* (2002) for details. In contrast, the exact and the almost exact confidence intervals in equations (10) and (11) respectively can provide honest coverage for τ when the instrument is weak. These methods do not rely on the normality assumption and, instead, use the distribution of $\mathbf{Z}$ from the experimental design as the starting point for inference on τ. Furthermore, the derivation of inference for methods based on expression (15) typically requires additional identifying assumption 3, which is akin to the fact that the moment of the product of the structural errors and the instrument is zero.

We can also compare the delta method confidence interval (i.e. the TSLS confidence interval) and the almost exact interval as follows. We can rewrite the delta method confidence interval in expression (17) as

$$\hat{\tau} \pm z_{1-\alpha/2} \frac{\sqrt{\hat{\Delta}}}{\hat{\tau}_D^2} \tag{20}$$

where $\hat{\Delta}$ was defined in equation (13) of the almost exact interval. Thus, if the almost exact interval produces finite intervals, i.e. if the instrument is sufficiently strong at the α-level (see Section 3.2 for details), the delta method confidence intervals and the almost exact interval look similar, but with notable differences in the centre and scaling between the two intervals.

Further comparisons of the expressions for different confidence intervals under the same notation provide some insight into the relationship between the traditional confidence intervals based on a point estimator plus and minus the variance in expression (15) *versus* confidence intervals based on randomization inference. The Bloom confidence interval in equation (18) is the crudest, yet simplest, approximation of inference for τ. The delta confidence interval in expression (17) offers a better approximation than the Bloom confidence interval by incorporating the variability of $\hat{\tau}_D$ and this is reflected by additional scaling terms to the right of $\text{var}(\hat{\tau}_Y|\mathcal{F}, \mathcal{Z})/\hat{\tau}_D^2$ in equation (16). The almost exact interval in equation (11) improves on the delta confidence interval by considering the case when $\hat{\tau}_D \approx 0$. Consequently, as seen in expression (20), we have slight differences in centring and scaling between the almost exact method and the delta confidence interval. Finally, the exact interval in equation (10) provides the exact confidence interval for τ, but without a closed form solution.

Next, we highlight some relationships between the almost exact interval, the Anderson–Rubin confidence interval (Anderson and Rubin, 1949) and the interval that was suggested by Fieller (Fieller, 1954; Burgess *et al.*, 2017). First, the almost exact interval is similar to the Anderson–Rubin confidence interval which is popular in the weak instrument literature in econometrics, with asymptotic equivalence when

(a) n is large and
(b) homoscedastic variance is assumed;

see the on-line supplementary materials for details. However, the motivation for the Anderson–Rubin confidence interval, like the TSLS-type confidence intervals, typically relies on structural modelling assumptions or moment conditions; see Stock *et al.* (2002) for one example. Instead, we motivated the almost exact method by using the randomization–inference framework where we made assumptions only on the randomization distribution of the instrument Z_i that was part of the design of the experiment. Also, the asymptotic equivalence between the Anderson–Rubin confidence interval and the almost exact interval suggests that the Anderson–Rubin confidence interval is very similar to the exact interval (10). Second, the interval by Fieller is equivalent to the almost exact interval; specifically, equation (5) of Fieller (1954) is identical to our quadratic equation of the almost exact interval in equation (11). However, similarly to the Anderson–Rubin confidence interval whose derivation typically required moment assumptions, Fieller derived his interval under stronger assumptions of

(a) bivariate normality of $\hat{\tau}_Y$ and $\hat{\tau}_D$ and
(b) independent estimates of the covariances of the bivariate normal and the mean of the bivariate normal distribution.

In contrast, our almost exact interval is derived on the basis of the distribution of the instrument Z_i, which is inherent in the experimental design. Indeed, the fact that our randomization-based intervals achieve these equivalence relationships with weaker assumptions demonstrates the strength of the randomization–inference framework to conduct inference in IV settings.

5. Extensions

5.1. Alternative test statistics

As we noted above, we have thus far considered exact approaches to inference that do not assume constant treatment effects at the unit level through the use of averages. However, results based on averages may be misleading when the tails of the outcome distribution are not well behaved. One additional advantage of the exact approach is that one can easily use other statistics that may be more robust. For example, one can use rank-based test statistics, which are robust in the presence of outliers and heavy-tailed distributions. Next, we briefly demonstrate how, under the exact approach, investigators may avoid the use of averages. Here, we must adopt a model of effects or a model for how units respond to treatment. Rosenbaum (1999) outlined a model of effects where the effect of encouragement on response is proportional to its effect on the treatment dose received:

$$Y_i^{(1,D_i^{(1)})} - Y_i^{(0,D_i^{(0)})} = \beta(D_i^{(1)} - D_i^{(0)}). \tag{21}$$

If this model holds then observed responses are related to observed doses through the equation

$$Y_i - \beta D_i = Y_i^{(1,D_i^{(1)})} - \beta D_i^{(1)} = Y_i^{(0,D_i^{(0)})} - \beta D_i^{(0)}.$$

Under this model of effects, the response will take the same value regardless of the value of Z_i, which makes this model of effects consistent with the exclusion restriction. Informally, the exclusion restriction implies that instrument assignment Z_i is related to the observed response

$$Y_i = Z_i Y_i^{(1,D_i^{(1)})} + (1 - Z_i) Y_i^{(0,D_i^{(0)})}$$

only through the realized dose of the treatment D_i. That is true here since $Y_i - \beta D_i$ is a constant that does not vary with Z_i. Under this model of effects, the treatment effect varies from unit to unit on the basis of the level of D_i as measured by $D_i^{(1)} - D_i^{(0)}$. If the unit received no dose, then

$$Y_i^{(1,D_i^{(1)})} - Y_i^{(0,D_i^{(0)})} = \beta(D_i^{(1)} - D_i^{(0)}) = 0.$$

Exact inference about β by using rank-based methods involves no new principles except that we must invoke a model of effects. Note that we do not assume that the effect is constant; here, the treatment effect varies from unit to unit on the basis of the level of D_i. However, we must invoke a model for treatment response. Here, we assume that the effect of encouragement on response is proportional to its effect on the treatment dose received. Thus, alternative tests statistics may require an additional assumption about the response to treatment; see Keele *et al.* (2017) for an exception when the outcome is binary. To test the sharp null hypothesis, $H_0: \beta = \beta_0$, by using rank-based methods, we use the observed quantity $Y_i - \beta_0 D_i = W_i$ as a set of adjusted responses. For example, to use the Wilcoxon rank sum statistic, we rank $|W_i|$ from 1 to N and take the sum of the ranks for which $Z_i = 1$ to form the test statistic W_{β_0}. Comparing W_{β_0} with the randomization distribution for the Wilcoxon rank sum statistic provides an exact p-value for a test of the sharp null hypothesis that $H_0: \beta = \beta_0$. A 95% confidence interval for the treatment effect is formed by inverting the test, or testing a series of hypotheses $H_0: \beta = \beta_0$ and retaining the set of values of β_0 not rejected at the 5% level (Rosenbaum, 2002). Rank-based methods are just one alternative. Using the adjusted responses, one could use a variety of test statistics. For example, one could test for differences in higher order moments of the adjusted responses.

5.2. Sensitivity analysis

When the IV method is used outside randomized experiments with non-compliance, assumption 1 no longer holds by design. Typically, analysts assume that assignment to encouragement (the instrument) is as if random or as if random conditional on observed covariates. In applications of this type, exact and almost exact methods allow a sensitivity analysis that assesses how departures from random assignment of the instrument might alter our conclusions. Rosenbaum (2002) demonstrated that, with exact methods, the analyst can place sharp bounds on the p-value from the test of the sharp null given the effect of a hypothetical confounder on the probability of being assigned to encouragement.

Rosenbaum used Γ to represent the odds of being assigned to treat $(Z_i = 1)$ for a unit as a function of a possible binary unobserved confounder. When $\Gamma = 1$, then the units do not differ in their odds of treatment assignment as a function of the unobserved confounder. This is true by design when Z_i is randomly allocated, but it may not be true in a natural experiment or observational study. To conduct a sensitivity analysis, the analyst uses values of Γ that are larger than 1 to place bounds on the p-value from the test of the sharp null. For example, if we find that the bounds on the p-value exceed 0.05 when $\Gamma = 1.05$ this suggests that a very slight departure from randomization of the instrument might overturn the conclusion from the study. If, in contrast, we find that the p-value from our study exceeds 0.05 when Γ is greater than say 4, this suggests that unless the departure from randomization of the instrument is fairly substantial our conclusions would still hold. These methods can be easily extended to the almost exact approach. See Baiocchi *et al.* (2010), Keele and Morgan (2016) and Kang *et al.* (2016) for examples of this form of sensitivity analysis applied to studies with IVs.

5.3. Covariate adjustment

Next, we consider covariate adjustment. Analysts often include baseline covariates in IV analyses. In randomized trials, covariates may increase precision, and in observational studies adjusting for covariates may be used to remove biases due to non-random assignment of instrument status. Under the asymptotic approaches in Section 4, such as TSLS, analysts can simply include baseline covariates in the linear regression models for the outcome and the treatment. The exact and almost exact approaches, however, do not preclude the use of baseline covariates. Rosenbaum (2002) outlined a general method for covariate adjustment with exact methods. Under his approach, exact methods are applied to the residuals from a model where the outcome has been regressed on baseline covariates. This method directly extends to the almost exact approach. Alternatively, as noted above covariate adjustment may also be applied via matching and then either exact and almost exact methods are used. Baiocchi *et al.* (2010), Keele and Morgan (2016) and Kang *et al.* (2016) have detailed the use of both exact and almost exact methods in IV analyses where matching is used for covariate adjustment.

5.4. Multivalued instruments and treatments

In many IV applications, either the instrument Z_i or actual treatment exposure D_i may not be binary. For example, D_i may record self-selected treatment dosage or Z_i may be some continuous level of encouragement to take the treatment. When D_i is multivalued, the randomization inference approach that is laid out in this paper remains the same and still provides correct confidence intervals without any modification. Note, however, that a multivalued treatment does change the interpretation of the estimand τ; see Angrist and Imbens (1995) and Kang *et al.* (2016) for examples. When instruments are multivalued, randomization inference can remain valid by using two strategies. First, investigators can simply create a binary IV from the continuous version. Angrist and Imbens (1994) showed that an IV estimate with a binary instrument and a continuous instrument converge to weighted averages of treatment effects, where the more 'compliant' subjects obtain greater weight. Second, Baiocchi *et al.* (2012) demonstrated how dichotomizing multivalued instruments may be avoided through the use of non-bipartite matching.

6. Simulation

We now present a simulation study to compare the properties of the various inferential methods. The study evaluates the confidence interval coverage of three methods of IV inference: the almost exact method, the TSLS method and the Bloom method. We do not include the exact method in the simulation since it is guaranteed always to have nominal coverage in finite samples whereas the other three methods are approximations of the finite sample behaviour. The simulation considers one-sided compliance under a finite sample and we evaluate the coverage rate for each method as the proportion of the compliers, τ_D, varies. First, we sample Z_i from a Bernoulli(0.5) distribution. To simulate the proportion of compliers, τ_D, we set the compliance rate to π and sample units from a uniform distribution. Let P_i denote the compliance class, which includes only compliers and never-takers in the one-sided compliance setting. We denote a unit as complier co if the draw from the uniform distribution is less than π. The outcomes Y_i follow a normal distribution with mean $\kappa + \gamma I(P_i = \text{co})$ and variance σ^2 where co indicates that a unit is a complier and $I(\cdot)$ is the indicator function. Under this model, γ is the effect on Y_i for $D_i = 1$ *versus* $D_i = 0$ for compliers. In the simulation, we set $\kappa = \gamma = \sigma^2 = 1$. This type of simulation set-up is not new and follows closely the simulation design in Guo *et al.* (2014).

In the simulations, we varied the compliance rate by using an interval of 5%, 10%, 25%, 50%, 75% and 90%. This implies that assignment to $Z_i = 1$ results in 5–90% of units being exposed to $D_i = 1$. Also, to study the behaviour at low compliance rates, we also add one additional compliance rate based on our discussion in Section 3.2. Specifically, on the basis of equation (14), the almost exact method of inference will return an infinite confidence interval if

$$\hat{\tau}_D \leqslant \frac{z_{1-\alpha/2}^2}{n + z_{1-\alpha/2}^2} \tag{22}$$

where we set $\widehat{\text{var}}(\hat{\tau}_D | \mathcal{F}, \mathcal{Z}) = \hat{\tau}_D(1 - \hat{\tau}_D)/n$ for one-sided compliance. In the simulations, we set the sample size to 100 and $\alpha = 0.05$ so, under equation (22), infinite confidence intervals will occur when the compliance rate is approximately 1.9%. Therefore, for one set of simulations, we set the compliance rate to 1.9%. We expect that the coverage rates for confidence intervals based on normal distribution approximations will tend to have incorrect coverage when the compliance rate falls around 1.9%. We repeat our simulation 5000 times for each method in each scenario. For each simulation, we record the 95% coverage rate for each method.

The results from the simulations are in Tables 1 and 2. First, we observe that, for the almost exact method, the coverages are at the nominal rate for any level of compliance i.e., even when less than 2% of the units are compliers, the almost exact method maintains 95% coverage. As such, the almost exact approximation appears to be quite accurate, even when the sample size is 100. Later, in an empirical example, we compare the almost exact method with the exact method.

For the asymptotic methods, both perform very poorly at the lowest levels of compliance. When the compliance rate is less than 2%, the coverage rates for both methods fail to reach

Table 1. Coverage rates of confidence intervals from the almost exact, Bloom and the TSLS method

Method	*Results for the following compliance rates:*						
	1.9%	*5%*	*10%*	*25%*	*50%*	*75%*	*90%*
Almost exact	0.950	0.944	0.945	0.947	0.955	0.941	0.948
Bloom	0.477	0.885	0.947	0.956	0.967	0.953	0.956
TSLS	0.503	0.934	0.996	0.978	0.964	0.945	0.949

Table 2. Median length of confidence intervals from almost exact, Bloom and the TSLS method

Method	*Results for the following compliance rates:*						
	1.9%	*5%*	*10%*	*25%*	*50%*	*75%*	*90%*
Almost exact	∞	∞	28.962	4.934	1.766	1.067	0.942
Bloom	28.510	18.561	8.956	4.060	1.768	1.097	0.965
TSLS	30.697	20.695	9.493	4.044	1.683	1.053	0.935

50%. When the compliance rate is 5%, the Bloom method fails to have a 90% coverage rate, whereas the TSLS method is close to the nominal coverage rate at 94%. However, for compliance rates above 75% both the Bloom and the TSLS method have the correct coverage. Moreover, the Bloom method appears to be an accurate approximation to the TSLS method. Once the compliance rate is 25% or higher the two methods have nearly identical coverage rates.

Table 2 looks at the median length of the confidence intervals. At low compliance rates, we observe that the confidence interval length for the almost exact method is either infinite or very large, reflecting the uncertainty that is inherent with low compliance and theoretically achieving the infinite length requirement that was laid out in Dufour (1997) for a weak instrument. In contrast, the Bloom and the TSLS methods tend to have large intervals as the compliance rate decreases but fails to achieve coverage. In fact, by design, the Bloom and the TSLS methods can never have infinite confidence intervals whereas the almost exact method can create infinite confidence intervals. Specifically, in our simulations, we noted that 97.6% of the 5000 simulated confidence intervals from the almost exact method were infinite when the compliance rate was 1.9%, 93.5% when the compliance rate was 5%, 26.2% when the compliance rate was 10% and 0.1% when the compliance rate was 25%. However, the length of the almost exact method confidence intervals conveys important information even when they are not infinite. For example, when compliance is 10% the asymptotic confidence intervals are very similar, whereas the almost exact intervals are nearly three times longer. Thus these intervals better convey the true level of statistical uncertainty. Finally, we also recorded the average point estimate for the almost exact method. For any of the compliance rates of 5% or higher, the bias that is associated with the almost exact method was 1% or less. When the compliance rate was 1.9%, the almost exact point estimate was too small by 32%.

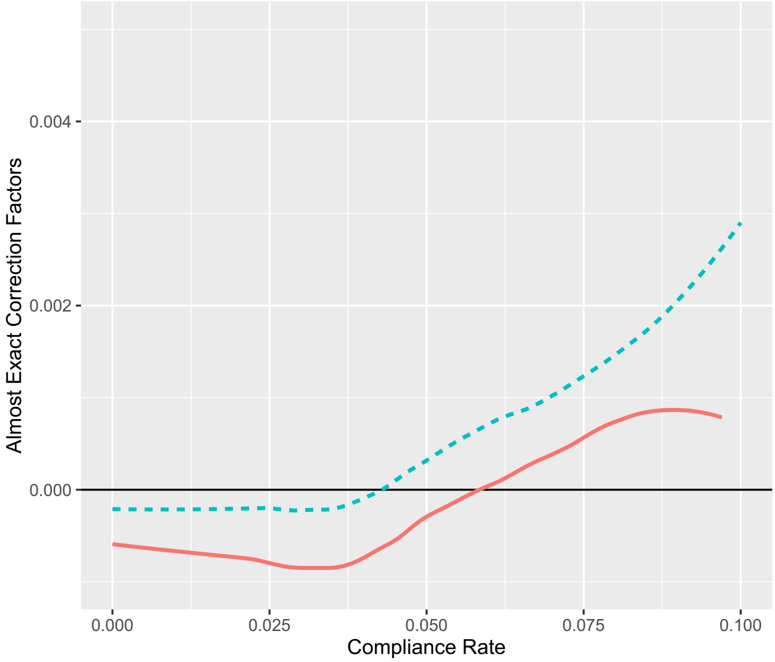


Fig. 1. Almost exact correction factors a (———) and b^2-4ac (– – –) plotted against compliance rates in simulated data: the almost exact interval will be finite when both values are greater than 0; in general, the width of the almost exact interval shrinks as these factors increase ($a=\hat{\tau}_D^2-z_{1-\alpha/2}^2\widehat{\mathrm{var}}(\hat{\tau}_D|\mathcal{F},\mathcal{Z})$, $b=-2\{\hat{\tau}_D\hat{\tau}_Y-z_{1-\alpha/2}^2\widehat{\mathrm{cov}}(\hat{\tau}_D,\hat{\tau}_Y|\mathcal{F},\mathcal{Z})\}$ and $c=\hat{\tau}_Y^2-z_{1-\alpha/2}^2\widehat{\mathrm{var}}(\hat{\tau}_Y|\mathcal{F},\mathcal{Z})$)

Next we include one additional simulation to convey how the almost exact method approximates the exact interval. Recall that the almost exact method depends on the three terms a, b and c that are defined in equation (11). These terms effectively act as correction factors that produce a confidence interval that approximates the exact interval. Specifically, the almost exact confidence interval depends on a and $b^2-4ac>0$: as the value of these two terms increases, the width of the almost exact interval decreases, and it will be finite when $a>0$ and $b^2-4ac>0$. Note that c plays a relatively small role in the width of the interval, since it depends only on $\hat{\tau}_Y$ whereas both a and b depend on $\hat{\tau}_D$.

In the second simulation, we used the same data-generating process and simulation parameters as in the first simulation, except that we varied the compliance rate from 1% to 10% in increments of 0.001. For each compliance rate, we ran 1000 simulations and recorded the average values of a, b and c. In Fig. 1, we plot smoothed values of a and b^2-4ac against the compliance rate. When the compliance rate is less than 2.5%, both terms are essentially flat. As the compliance rate increases, the correction terms increase, which narrows the almost exact interval. We observe that, whereas the term b^2-4ac exceeds 0 once compliance rates are around 4%, neither term exceeds 0 until the compliance rate is approximately 6%. Next, we use three different empirical applications to highlight different aspects of these inferential methods.

7. Applications

7.1. Application 1: the ‘Green jobs and health care’ intervention

As part of a comprehensive economic stimulus package funded under the 2009 American Recovery and Reinvestment Act, the US Department of Labor awarded a series of grants to promote training for employment in energy efficiency, renewable energy and healthcare. Grants were awarded to four sites across the USA. At two sites, additional training was offered on topics such as the installation of solar and wind power systems. At the other two sites, additional training was offered in the healthcare sector. These new training initiatives were subject to evaluation in the GJHC impact evaluation (Copson *et al.*, 2015; Martinson *et al.*, 2015).

At each site, participants were randomized to either participation in the new training programme, i.e. treatment, or to the existing available training programme—control. At all four of the sites, some trainees who were assigned to the new training initiatives selected not to participate. However, in the study design, those who were randomized to the standard training condition could not access the treatment. Thus non-compliance was one sided in this application. The primary outcomes were earnings and employment status. Here, we focus on the employment status outcome which was measured through a survey after trainees completed their course of study. We use a binary outcome measure which asked whether the participants had been employed at any time since ending the training programme.

We conducted a separate analysis for each site given that the content of the training programmes varied significantly across the four sites. Table 3 contains the total number of participants in the randomized intervention at each site along with the compliance rate. At all four sites, compliance with assigned treatment status was high. The lowest level of compliance was at site 1, where 62.1% of those who were randomized to treatment participated. At the other three sites, participation among those assigned to treatment exceeded 75%.

In the GJHC application, the sample sizes are sufficiently small that exact methods are feasible. As such, we compare results from an exact method that was developed in Keele *et al.* (2017) with the almost exact method. Table 3 contains point estimates and 95% confidence intervals by using both exact methods and the approximation to the exact method outlined in Section 3.2. First, the approximation clearly improves as the sample size increases. Site 4 has the largest sample size with 719 participants, and the exact and almost exact methods provide essentially identical results. For site 4, the exact 95% confidence interval is [−0.05, 0.06], and the almost exact 95% confidence interval is [−0.05, 0.05]. However, the computation time for the exact method was over 8 min on a desktop computer with a 4.0-GHz processor and 32.0 Gbytes of random-access memory. The almost exact routine is essentially instantaneous, since it is based on a closed form solution. As such, the almost exact method provides an accurate approximation to the exact results but requires very little computing power. Site 2 has the smallest sample size, so we might expect the discrepancy between the exact and almost exact confidence intervals to be largest for the analysis of this training site. Here, the exact 95% confidence interval is [−0.06, 0.29] and the almost exact 95% confidence interval is [−0.07, 0.19]. The exact confidence interval, then,

Table 3. Point estimates and confidence intervals for exact and almost exact methods in the GJHC data

	Results for the following sites:			
	Site 1	*Site 2*	*Site 3*	*Size 4*
Hodges–Lehmann point estimate	0.050	0.060	0.084	−0.003
Almost exact 95% confidence interval	[−0.06, 0.16]	[−0.07, 0.19]	[0.02, 0.15]	[−0.05, 0.05]
Exact 95% confidence interval	[−0.05, 0.19]	[−0.06, 0.29]	[0.02, 0.17]	[−0.05, 0.06]
Computation time (min)	0.04	0.01	0.63	8.9
N	318	169	546	719
Compliance rate (%)	62.1	79.3	79.9	83.9

is longer as it exactly reflects finite sample uncertainty, and in this case exact methods require very little computation time. These results suggest that analysts should use exact methods when sample sizes are smaller.

7.2. Application 2: a get out the vote intervention

One literature in political science studies methods for increasing voter turnout through the use of randomized field experiments. This research both focuses on the effectiveness of various get out the vote methods and tests social psychological theories about voters (Green *et al.*, 2013). One entry in this literature focused on the effectiveness of door-to-door canvassing where volunteers knock on doors urging people to vote in an upcoming election (Green *et al.*, 2003). In this study, the researchers conducted six separate field experiments in the following cities: Bridgeport, Columbus, Detroit, Minneapolis, Raleigh and St Paul in November 2001. In each city, households were randomized either to receive face-to-face contact from local staffers encouraging them to vote, i.e. treatment, or were not contacted, i.e. control. Many of the households that were randomized to the treatment were not available for the face-to-face message encouraging them to vote. Although the ITT effects are easily estimable, in this context, one might argue that IV estimates are of greater interest, since these reveal the causal effect of actually receiving the get out the vote message. In the original analysis, the analysts estimated complier effects by using asymptotic approximations for the variance estimates (Green *et al.*, 2003).

The sample sizes for these experiments, however, make using exact methods computationally intensive. For example, the experiment in St Paul had 2146 participants. When we attempted to obtain exact results on a desktop computer with a 4.0-GHz processor and 32.0 Gbytes random-access memory, computation stopped after 345 min because the computer had run out of memory. Thus, even in fairly modest sample sizes, exact methods may be infeasible. Here, we compare the almost exact results with results based on the delta or TSLS method and the Bloom approximation. Table 4 contains the point estimates and confidence intervals for these three methods. All three methods produce essentially identical results. Despite the fact that

Table 4. Point estimates and confidence intervals for exact and almost exact methods in the voting intervention data

	Results for the following cities:					
	Bridgeport	*Columbus*	*Detroit*	*Minneapolis*	*Raleigh*	*St Paul*
Point estimate	0.163	0.105	0.083	0.104	−0.020	0.138
Almost exact 95% confidence interval	[0.052, 0.274]	[−0.061, 0.270]	[−0.007, 0.174]	[−0.071, 0.280]	[−0.081, 0.039]	[0.013, 0.263]
TSLS 95% confidence interval	[0.053, 0.273]	[−0.059, 0.269]	[−0.007, 0.173]	[−0.070, 0.279]	[−0.080, 0.039]	[0.014, 0.262]
Bloom 95% confidence interval	[0.051, 0.275]	[−0.060, 0.269]	[−0.007, 0.173]	[−0.070, 0.279]	[−0.080, 0.039]	[0.013, 0.263]
N	1650	2424	4954	2827	4660	2146
Compliance rate (%)	28.9	14.0	30.7	18.5	45.2	33.1

the compliance rates are, at times, less than 15%, the larger sample sizes ensure that all three methods produce identical inferences.

7.3. Application 3: rainfall as an instrument for voter turnout

As we noted earlier, instruments have a long history of use outside randomized trials. Here, instruments are used as a type of natural experiment, where the instrument is a haphazard nudge or 'encouragement' to treatment exposure. The likelihood of weak instruments tends to be higher when instruments are used outside randomized evaluations. As such, the utility of exact or almost exact methods is likely to be greater when instruments are used in observational studies. Here, we conduct a reanalysis of Hansford and Gomez (2010) to explore whether almost exact methods may be useful in an observational study with an instrument.

Hansford and Gomez (2010) used deviations from average rainfall on election day as an instrument for voter turnout to estimate the causal effect of voter turnout on vote share in US elections. Using this instrument, they found that higher turnout tends to help Democratic candidates. The original analysis spanned all presidential elections in non-southern counties from 1948 to 2000. Here, we investigate the possibility that the strength of rainfall as an IV for voter turnout perhaps declined over time, i.e. changes in transportation patterns over time might weaken the effect of rainfall on turnout, since voters may be less affected by rain when they can drive to the polls. In our analysis, we conduct three separate analyses. The first analysis uses all presidential elections from 1976 to 2000. The second uses all presidential elections from 1980 to 2000, and the third uses all presidential elections from 1984 to 2000. For every analysis, we have close to 10000 observations or more. Thus, uncertainty in the IV estimate will be largely driven by the strength of the instrument, instead of the sample size. We used the almost exact method, the Bloom method and TSLS, which was the method that was used in the original analysis, for interval estimation.

Table 5 contains the results from the analysis. When we analyse the elections from 1976 to 2000, all three methods return 95% confidence intervals that are quite similar. The almost exact method results do have wider confidence intervals, but the difference is relatively small. The almost exact 95% confidence interval is $[-0.91, -2.42]$, whereas the 95% confidence interval based on TSLS is $[-0.94, -1.94]$. For presidential elections from 1980 to 2000, the confidence interval for the almost exact method is noticeably wider; in fact the almost exact interval, $[-1.26, -5.19]$, is almost twice as long as the interval from TSLS, $[-1.18, -3.25]$, and the Bloom method, $[-1.59, -3.25]$. Finally, we restrict the data to the period from 1984 to 2000; the differences in confidence intervals are quite stark. Now the almost exact method returns an interval that covers from $-\infty$ to ∞. The intervals from the TSLS and the Bloom method are wider than before, but both are closed intervals. For example the TSLS 95% confidence interval is $[-0.037, -5.99]$, and the 95% confidence interval based on the Bloom method is

Table 5. Point estimates and confidence intervals for analysis of rainfall as an instrument for voter turnout

	Results for elections 1976–2000	*Results for elections from 1980–2000*	*Results for elections from 1984–2000*
N	13687	11729	9770
Point estimate	−1.4	−2.2	−4.6
Almost exact 95% confidence interval	$[-0.91, -2.42]$	$[-1.26, -5.19]$	$[-\infty, \infty]$
TSLS 95% confidence interval	$[-0.94, -1.94]$	$[-1.18, -3.25]$	$[-0.037, -5.99]$
Bloom 95% confidence interval	$[-1.04, -1.85]$	$[-1.59, -3.25]$	$[-3.27, -5.99]$

[−3.27, − 5.99]. The confidence interval from the almost exact approximation provides a clear warning that the instrument in this case is weak. It is worth noting that in this analysis there are nearly 10000 observations, so the sample size is more than adequate. The lack of statistical certainty, here, is driven almost entirely by the weakness of the instrument.

8. Discussion

In this paper, we have reviewed inferential methods for the IVs method, with a focus on exact methods. In particular, we highlighted exact and almost exact methods, which see little use in practice but have important advantages. We used a Monte Carlo study to show that the almost exact method maintains the nominal coverage rate even when the instrument is quite weak. In contrast, methods based on normal distribution approximations had poor coverage in the same setting. However, when the instrument is strong and sample sizes are large, all the methods provide very similar results, both in simulations and empirical applications. In fact, the normal distribution approximation via the Bloom method seems to provide the simplest form of inference for τ under this case. The Bloom method still sees widespread use when analysts attempt to convey results to non-technical audiences. This appears to be a safe practice when applied to a randomized encouragement design with one-sided non-compliance rates that do not fall below 25%. Although we do not provide systematic evidence on this point, we suspect that in most randomized policy interventions compliance rates are typically not this low. However, in observational studies, the likelihood of weak instruments is greater and exact or almost exact methods should see more use by applied analysts.

One additional method of inference that may be applied to IV estimators is the bootstrap. We did not consider the bootstrap in this paper, since we confined ourselves to finite population inference, and the bootstrap typically assumes an infinite population model. Moreover, the smoothness condition that is required for the bootstrap may fail when the instrument is weak. Finally, whereas the bootstrap is generally second-order accurate, that is not always so for IVs (Horowitz, 2001). In our opinion, unless investigators are interested in population inferences, exact or almost exact methods are generally preferred over the bootstrap when applied to IV estimators.

One area of applied study where the problem of weak instruments is common is in the study of genetics. In MR studies, the IV method has become a standard analytic tool. In these studies, the source of the instruments is genetic variations, and the compliance rates, or, in the MR context, the explained genetic variations, can be very low. Whereas there has been work on weak instruments within the MR context (Burgess and Thompson, 2011; Burgess *et al.*, 2011; Pierce *et al.*, 2011), we believe that our work here, specifically the almost exact method with its attractive formula and robustness guarantees, can complement some of the proposals to deal with the problem of weak instruments in MR.

We provide an R function which is available from `http://wileyonlinelibrary.com/journal/rss-datasets` that returns confidence intervals under the almost exact method for use by applied researchers. One additional advantage of almost exact methods is that they can be combined with rank-based test statistics when the outcome distribution is heavy tailed or has unusual observations (Rosenbaum, 1996). When a rank-based test statistic is used, asymptotic approximations to the randomization distribution again provide convenient results when sample sizes are larger.

Acknowledgements

We thank Frank Windemeijer, Guido Imbens, Jacob Klerman and participants of the University of Pennsylvania Causal Inference Group for comments and suggestions. The research of Hyunseung Kang was supported in part by National Science Foundation grant DMS-1502437. The research of Luke Keele was supported in part by US Department of Labor grant GS-10F-0086K.

References

Anderson, T. W. and Rubin, H. (1949) Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.*, **20**, 46–63.

Angrist, J. D. and Imbens, G. W. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.

Angrist, J. D. and Imbens, G. W. (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Statist. Ass.*, **90**, 431–442.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Angrist, J. D. and Krueger, A. B. (2001) Instrumental variables and the search for identification: from supply and demand to natural experiments. *J. Econ. Perspect.*, **15**, 69–85.

Angrist, J. D. and Pischke, J.-S. (2008) *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton: Princeton University Press.

Baiocchi, M., Cheng, J. and Small, D. S. (2014) Instrumental variable methods for causal inference. *Statist. Med.*, **33**, 2297–2340.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010) Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Am. Statist. Ass.*, **105**, 1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D. and Groeneveld, P. W. (2012) Near/far matching: a study design approach to instrumental variables. *Hlth Serv. Outcms Res. Methodol.*, **12**, 237–253.

Bloom, H. S. (1984) Accounting for no-shows in experimental evaluation designs. *Evaln Rev.*, **8**, 225–246.

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. *J. Am. Statist. Ass.*, **90**, 443–450.

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N. A. and Thompson, J. R. (2016) Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I 2 statistic. *Int. J. Epidem.*, **45**, 1961–1974.

Burgess, S., Small, D. S. and Thompson, S. G. (2017) A review of instrumental variable estimators for Mendelian randomization. *Statist. Meth. Med. Res.*, **26**, 2333–2355.

Burgess, S. and Thompson, S. G. (2011) Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statist. Med.*, **30**, 1312–1323.

Burgess, S. and Thompson, S. G. (2012) Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statist. Med.*, **31**, 1582–1600.

Burgess, S., Thompson, S. G. and CRP CHD Genetics Collaboration (2011) Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidem.*, **40**, 755–764.

Copson, E., Martinson, K., Benson, V., DiDomenico, M., Williams, J., Needels, K. and Mastri, A. (2015) The Green Jobs and Health Care Impact Evaluation: Findings from the implementation study of four training programs for unemployed and disadvantaged workers. ABT Associates, Bethesda.

Davey Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidem.*, **32**, 1–22.

Davey Smith, G. and Ebrahim, S. (2004) Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidem.*, **33**, 30–42.

Deaton, A. (2010) Instruments, randomization, and learning about development. *J. Econ. Lit.*, **48**, 424–455.

Ding, P., Feller, A. and Miratrix, L. (2016) Randomization inference for treatment effect variation. *J. R. Statist. Soc.* B, **78**, 655–671.

Dufour, J.-M. (1997) Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, **65**, 1365–1387.

Fieller, E. C. (1954) Some problems in interval estimation. *J. R. Statist. Soc.* B, **16**, 175–185.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Green, D. P., Gerber, A. S. and Nickerson, D. W. (2003) Getting out the vote in local elections: results from six door-to-door canvassing experiments. *J. Polit.*, **65**, 1083–1096.

Green, D. P., McGrath, M. C. and Aronow, P. M. (2013) Field experiments and the study of voter turnout. *J. Elect. Publ. Opin. Parties*, **23**, 27–48.

Guo, Z., Cheng, J., Lorch, S. A. and Small, D. S. (2014) Using an instrumental variable to test for unmeasured confounding. *Statist. Med.*, **33**, 3528–3546.

Hájek, J. (1960) Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 361–374.

Hansen, B. B. and Bowers, J. (2009) Attributing effects to a clustered randomized get-out-the-vote campaign. *J. Am. Statist. Ass.*, **104**, 873–885.

Hansford, T. G. and Gomez, B. T. (2010) Estimating the electoral effects of voter turnout. *Am. Polit. Sci. Rev.*, **104**, 268–288.

Heckman, J., Smith, J. and Taber, C. (1998) Accounting for dropouts in evaluations of social programs. *Rev. Econ. Statist.*, **80**, 1–14.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006) Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Statist.*, **88**, 389–432.

Hernán, M. A. and Robins, J. M. (2006) Instruments for causal inference: an epidemiologists dream. *Epidemiology*, **17**, 360–372.

Horowitz, J. L. (2001) The bootstrap. In *Handbook of Econometrics*, vol. 5 (eds J. J. Heckman and E. Leamer), pp. 3159–3228. Amsterdam: Elsevier.

Imbens, G. W. (2010) Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.*, **48**, 399–423.

Imbens, G. W. (2014) Instrumental variables: an econometrician's perspective. *Statist. Sci.*, **29**, 323–358.

Imbens, G. W. and Rosenbaum, P. R. (2005) Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J. R. Statist. Soc.* A, **168**, 109–126.

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction*. Cambridge: Cambridge University Press.

Kang, H. (2016) Matched instrumental variables: a possible solution to severe confounding in matched observational studies? *Epidemiology*, **27**, 633–636.

Kang, H., Kreuels, B., May, J. and Small, D. S. (2016) Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Ann. Appl. Statist.*, **10**, 335–364.

Keele, L. J. and Morgan, J. (2016) How strong is strong enough?: Strengthening instruments through matching and weak instrument tests. *Ann. Appl. Statist.*, **10**, 1086–1106.

Keele, L., Small, D. and Grieve, R. (2017) Randomization-based instrumental variables methods for binary outcomes with an application to the 'IMPROVE' trial. *J. R. Statist. Soc.* A, **180**, 569–586.

Kleiber, C. and Zeileis, A. (2008) *Applied Econometrics with R*. New York: Springer.

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

Kleibergen, F. and Zivot, E. (2003) Bayesian and classical approaches to instrumental variable regression. *J. Econmetr.*, **114**, 29–72.

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

Lehmann, E. L. (2004) *Elements of Large-sample Theory*. New York: Springer.

Lehmann, E. L. (2006) *Nonparametrics: Statistical Methods based on Ranks*. New York: Springer.

Lehmann, E. L. and Romano, J. P. (2008) *Testing Statistical Hypotheses*. New York: Springer.

Martinson, K., Williams, J., Needels, K., Peck, L., Moulton, S., Paxton, N., Mastri, A., Copson, E., Nisar, H., Comfort, A. and Brown-Lyons, M. (2015) The Green Jobs and Health Care Impact Evaluation: findings from the impact study of four training programs for unemployed and disadvantaged workers. *Report*. ABT Associates, Bethesda.

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

Nelson, C. R. and Startz, R. (1990) Some further results on the exact sample properties of the instrumental variables estimator. *Econometrica*, **58**, 967–976.

Nolen, T. L. and Hudgens, M. G. (2011) Randomization-based inference within principal strata. *J. Am. Statist. Ass.*, **106**, 581–593.

Pierce, B. L., Ahsan, H. and VanderWeele, T. J. (2011) Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidem.*, **40**, 740–752.

Rosenbaum, P. R. (1996) Identification of causal effects using instrumental variables: comment. *J. Am. Statist. Ass.*, **91**, 465–468.

Rosenbaum, P. R. (1999) Using quantile averages in matched observational studies. *Appl. Statist.*, **48**, 63–78.

Rosenbaum, P. R. (2001) Effects attributable to treatment: inference in experiments and observational studies with a discrete pivot. *Biometrika*, **88**, 219–231.

Rosenbaum, P. R. (2002) *Observational Studies*, 2nd edn. New York: Springer.

Rosenbaum, P. R. (2003) Exact confidence intervals for nonconstant effects by inverting the signed rank test. *Am. Statistn*, **57**, 132–138.

Rubin, D. B. (1980) Randomization analysis of experimental data: the Fisher randomization test comment. *J. Am. Statist. Ass.*, **75**, 591–593.

Staiger, D. and Stock, J. H. (1997) Instrumental variables regression with weak instruments. *Econometrica*, **65**, 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. *J. Bus. Econ. Statist.*, **20**, 518–529.

Swanson, S. A. and Hernán, M. A. (2014) Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. *Statist. Sci.*, **29**, 371–374.

Wald, A. (1940) The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, **11**, 284–300.

Wang, J. and Zivot, E. (1998) Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica*, **66**, 1389–1404.
Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge: MIT Press.
Yang, F., Zubizaretta, J., Small, D. S., Lorch, S. and Rosenbaum, P. (2014) Dissonant conclusions when testing the validity of an instrumental variable. *Am. Statistn*, **68**, 253–263.
Zivot, E., Startz, R. and Nelson, C. R. (1998) Valid confidence intervals and inference in the presence of weak instruments. *Int. Econ. Rev.*, 1119–1144.

Supporting information
Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary materials for “Inference for instrumental variables: a randomization inference approach”’.