*Biometrika* (2017), **104**, 1, *pp.* 229–236
*Printed in Great Britain*
doi: 10.1093/biomet/asw064
Advance Access publication 23 January 2017

# On falsification of the binary instrumental variable model

By LINBO WANG

*Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*
linbowang@g.harvard.edu

JAMES M. ROBINS

*Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*
robins@hsph.harvard.edu

and THOMAS S. RICHARDSON

*Department of Statistics, University of Washington, Box 354322, Washington 98195, U.S.A.*
thomasr@u.washington.edu

## Summary

Instrumental variables are widely used for estimating causal effects in the presence of unmeasured confounding. The discrete instrumental variable model has testable implications for the law of the observed data. However, current assessments of instrumental validity are typically based solely on subject-matter arguments rather than these testable implications, partly due to a lack of formal statistical tests with known properties. In this paper, we develop simple procedures for testing the binary instrumental variable model. Our methods are based on existing techniques for comparing two treatments, such as the $t$-test and the Gail–Simon test. We illustrate the importance of testing the instrumental variable model by evaluating the exogeneity of college proximity using the National Longitudinal Survey of Young Men.

*Some key words*: Binary response; Gail–Simon test; Instrumental variable; Qualitative interaction; $t$-test; Two-by-two table.

## 1. Introduction

The instrumental variable method has been widely used for estimating causal effects in the presence of unmeasured confounders. A variable $Z$ is called an instrumental variable if: (a) it is independent of unmeasured confounders $U$; (b) it does not have a direct effect on the outcome $Y$; (c) it has a nonzero average causal effect on the treatment $D$ (Angrist et al., 1996). In many applications, assumption (a) is reasonable only after controlling for observed covariates $V$ (Baiocchi et al., 2014). The resulting model is called the conditional instrumental variable model. Figure 1 gives a directed acyclic graphical model representation (Pearl, 2009) of the conditional instrumental variable model, in which the faithfulness (Spirtes et al., 2000) of the edge $Z \rightarrow D$ is assumed.

Unlike the assumption of no unmeasured confounders between $D$ and $Y$, the instrumental variable model with discrete observables $(Z, D, Y)$ imposes nontrivial constraints on the observed-data distribution.

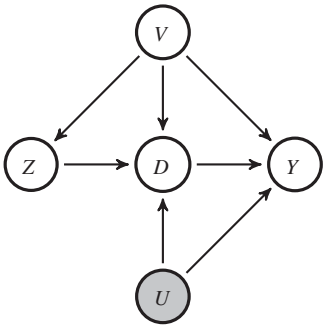


Fig. 1. Directed acyclic graph representing an instrumental variable model. The variables $V$, $Z$, $D$ and $Y$ are observed; $U$ is unobserved.

In particular, Balke & Pearl (1997) and Bonet (2001) give the following necessary and sufficient condition for an observed-data distribution $p(d, y \mid z)$ to be compatible with an unconditional binary instrumental variable model where $Z$, $D$ and $Y$ take values 0 and 1:

$$\text{pr}(D = d, Y = y \mid Z = 1) + \text{pr}(D = d, Y = 1 - y \mid Z = 0) \leqslant 1 \quad (d = 0, 1;\ y = 0, 1). \tag{1}$$

Here the unconditional instrumental variable model refers to the model with an empty control variable set $V$. In particular, if the potential instrument $Z$ is randomized so that assumption (a) holds, then violation of each inequality in (1) corresponds to a nonzero average controlled direct effect of $Z$ on $Y$, which violates assumption (b) (Cai et al., 2008; Richardson et al., 2011). Although assumption (c) imposes on the observables the constraint

$$\text{pr}(D = 1 \mid Z = 1) \neq \text{pr}(D = 1 \mid Z = 0), \tag{2}$$

it is in general not possible to reject (2) with a statistical test. Hence hereafter we do not discuss constraint (2). Similarly, the testable implications of a conditional binary instrumental variable model are given by

$$\begin{aligned}\text{pr}(D = d, Y = y \mid Z = 1, V = v) + \text{pr}(D = d, Y = 1 - y \mid Z = 0, V = v) \leqslant 1 \\ (d = 0, 1;\ y = 0, 1;\ v \in \mathcal{V}),\end{aligned} \tag{3}$$

where $\mathcal{V}$ contains all possible values for $V$. In practice, the inequalities (1) can be used to partially test the binary unconditional instrumental variable model. Likewise, (3) can be used to test the binary conditional instrumental variable model. In contrast, it is impossible to empirically falsify the assumption of no unmeasured confounders between $D$ and $Y$ as in an observational study without an instrument.

Although there have been many discussions of estimation of causal effects under the binary instrumental variable model (Vansteelandt et al., 2011; Clarke & Windmeijer, 2012), less attention has been paid to testing its validity. Prior to our work, Ramsahai & Lauritzen (2011) considered testing an unconditional binary instrumental variable model using a likelihood ratio test. Their approach involves solving a constrained optimization problem and cannot be used to test the conditional binary instrumental variable model as described by Fig. 1. Furthermore, their approach tests the four inequalities in (1) jointly. Hence, without modification, it can only be used to falsify the binary instrumental variable model, but cannot identify which specific average controlled direct effect of $Z$ on $Y$ must be positive or negative. In related work, Kang et al. (2013) provided a falsification test for the instrumental variable assumptions given knowledge of a subpopulation where the edge $Z \to D$ is absent. In this paper we develop a novel perspective on falsification of the binary instrumental variable model. Specifically, we show that

testing (1) or (3) is equivalent to testing for a nonpositive effect of the instrument $Z$ on a constructed variable.

## 2. Tests for the unconditional binary instrumental variable model

To fix ideas, we first consider testing the instrumental variable inequality

$$\mathrm{pr}(D = 0, Y = 1 \mid Z = 1) + \mathrm{pr}(D = 0, Y = 0 \mid Z = 0) \leqslant 1. \tag{4}$$

Equation (4) can be rewritten as

$$\mathrm{pr}(D = 0, Y = 1 \mid Z = 1) - 1 + \mathrm{pr}(D = 0, Y = 0 \mid Z = 0) \leqslant 0.$$

Define a new variable

$$Q^{01} = \begin{cases} I(D = 0, Y = 1), & Z = 1, \\ 1 - I(D = 0, Y = 0), & Z = 0, \end{cases}$$

where $I(\cdot)$ is the indicator function. It then follows that

$$\begin{aligned} &\mathrm{pr}(D = 0, Y = 1 \mid Z = 1) - \{1 - \mathrm{pr}(D = 0, Y = 0 \mid Z = 0)\} \\ &= \mathrm{pr}(Q^{01} = 1 \mid Z = 1) - \mathrm{pr}(Q^{01} = 1 \mid Z = 0) \equiv \Delta^{01}. \end{aligned}$$

Testing (4) is therefore equivalent to the testing problem

$$\mathcal{H}_0^{01} : \Delta^{01} \leqslant 0 \quad \text{versus} \quad \mathcal{H}_{\mathrm{a}}^{01} : \Delta^{01} > 0, \tag{5}$$

which is simply one-sided testing for a $2 \times 2$ table.

In general, we have four inequalities of the form (4) with a binary instrumental variable model, so multiplicity adjustment is needed. Suppose for now that we have one-sided tests $\phi^{00}, \phi^{01}, \phi^{10}, \phi^{11}$ such that the size of $\phi^{dy}$ goes to zero asymptotically in the interior of the null space defined by $\mathcal{H}_0^{dy}$. Furthermore, assume that the rejection region of $\phi^{dy}$ has no intersection with the null space defined by $\mathcal{H}_0^{dy}$ (Perlman & Wu, 1999). To get a level-$\alpha$ test for (1), a naive Bonferroni correction would require that each $\phi^{dy}$ have size less than or equal to $\alpha/4$ for testing $\mathcal{H}_0^{dy}$. However, the left-hand sides of the four inequalities in (1) sum to 2, and hence at most two of them can simultaneously hold with equality. Based on this, we now show that it suffices to control the level of each test $\phi^{dy}$ at $\alpha/2$.

Specifically, let $u^{dy} = \mathrm{pr}(D = d, Y = y \mid Z = 1) + \mathrm{pr}(D = d, Y = 1 - y \mid Z = 0)$, and let $\zeta = (u^{00}, u^{01}, u^{10})$. The null space defined by (1) can be represented by an octahedron $\mathcal{Z}_0$ in the simplex $\mathcal{Z}$, where $\mathcal{Z}$ is defined as

$$\mathcal{Z} = \left\{\zeta : u^{00} + u^{01} + u^{10} \leqslant 2, \quad u^{00}, u^{01}, u^{10} \geqslant 0\right\}.$$

Figure 2 gives a graphical depiction of $\mathcal{Z}$ and $\mathcal{Z}_0$. Each of the four blue-shaded facets corresponds to one inequality in (1) holding with equality. Six points, shown in red, have two inequalities in (1) holding with equality. The interior of the null space $\mathcal{Z}_0$ corresponds to cases where none of the four inequalities in (1) holds with equality.

We are now ready to present our multiplicity adjustment procedure. The proof of Theorem 1 is given in the Appendix.

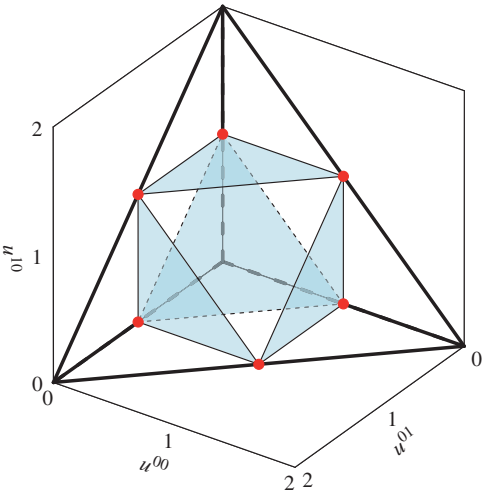


Fig. 2. Representation of the simplex $\mathcal{Z}$ and the null space $\mathcal{Z}_0$. The edges of the simplex $\mathcal{Z}$ are represented by thick black lines, and the null space $\mathcal{Z}_0$ is the octahedron whose vertices are shown in red; four of the eight surfaces are shaded blue.

THEOREM 1. *Propose a testing procedure as follows: reject* (1) *if for* $d = 0, 1$ *and* $y = 0, 1$ *at least one of* $\mathcal{H}_0^{dy}$ *is rejected by* $\phi^{dy}$ *at level* $\alpha/2$. *Under the null hypotheses* (1)*:*

(i) *if two inequalities in* (1) *hold with equality at the true value* $\dot{\zeta}$*, then the proposed test has size* $\alpha$*;*
(ii) *if only one of the inequalities in* (1) *holds with equality at the true value* $\dot{\zeta}$*, then asymptotically the proposed test has size* $\alpha/2$*;*
(iii) *if none of the inequalities in* (1) *holds with equality at the true value* $\dot{\zeta}$*, then asymptotically the proposed test has size* 0.

*In particular, the proposed test always has asymptotic size no greater than* $\alpha$.

We now turn to the choice of $\phi^{dy}$. Over the past century, there has been much discussion on testing association in $2 \times 2$ tables, including size and power comparisons for different test statistics and methods of computing the $p$-value; see Lydersen et al. (2009) for a review. When the sample size is large, asymptotic tests such as those based on the $t$-statistic are popular among researchers. However, under independent and identically distributed sampling they may not preserve the test size with small samples, in which case unconditional exact tests such as the Fisher–Boschloo test are recommended.

*Remark* 1. The computation time for unconditional tests can be excessive when the sample size is moderate or large, in which case it may be desirable to use the procedure of Berger & Boos (1994) to reduce computation time. The proposed test still has asymptotic size no greater than $\alpha$ provided $\gamma < \alpha/2$, where $100(1-\gamma)\%$ is the confidence level for the nuisance parameter.

*Remark* 2. The Wald test for the $2 \times 2$ table corresponding to (5) coincides with the Wald test for (4), where $\text{pr}(D = 0, Y = 1 \mid Z = 1)$ and $\text{pr}(D = 0, Y = 0 \mid Z = 0)$ are estimated via maximum likelihood. However, our introduction of $Q^{dy}$ builds the connection between testing unconditional instrumental inequalities and testing $2 \times 2$ tables, and hence motivates many more approaches to testing the unconditional instrumental inequalities.

We now discuss the interpretation of results from our testing procedure. As noted by Cai et al. (2008) and Richardson et al. (2011), under the randomization assumption, the average controlled direct effect of

$Z$ on $Y$, $\text{ACDE}(d) = E\{Y(z=1,d=d)\} - E\{Y(z=0,d=d)\}$, satisfies

$$\begin{aligned}\text{pr}(D=d,Y=1 \mid Z=1) + \text{pr}(D=d,Y=0 \mid Z=0) - 1 \leqslant \text{ACDE}(d)\\ \leqslant 1 - \text{pr}(D=d,Y=0 \mid Z=1) - \text{pr}(D=d,Y=1 \mid Z=0).\end{aligned} \tag{6}$$

It follows that violation of each inequality in (1) corresponds to a nonzero average controlled direct effect of $Z$ on $Y$. Our testing procedure is therefore interpretable in the sense that if we reject the binary instrumental variable model, we would also know which average controlled direct effect is positive or negative. For example, suppose we reject the null that $\text{pr}(D=0,Y=1 \mid Z=1) + \text{pr}(D=0,Y=0 \mid Z=0) \leqslant 1$, then from (6) we would also conclude that $\text{ACDE}(0)$ is positive.

## 3. Tests for the conditional binary instrumental variable model

Suppose now we wish to test the instrumental variable inequality that

$$\text{pr}(D=0,Y=1 \mid Z=1,V=v) + \text{pr}(D=0,Y=0 \mid Z=0,V=v) \leqslant 1, \quad v \in \mathcal{V}. \tag{7}$$

Using the same arguments as in § 2, we can rewrite the testing problem of (7) as

$$\mathcal{H}_{0,\text{c}}^{01} : \text{for all } v \in \mathcal{V},\ \Delta^{01}(v) \leqslant 0 \quad \text{versus} \quad \mathcal{H}_{\text{a,c}}^{01} : \text{there exists } v \in \mathcal{V} \text{ such that } \Delta^{01}(v) > 0, \tag{8}$$

where $\Delta^{01}(v) = \text{pr}(Q^{01}=1 \mid Z=1,V=v) - \text{pr}(Q^{01}=1 \mid Z=0,V=v)$ and a subscript c denotes conditional.

The testing problem (8) concerns the null hypothesis that a particular treatment is at least as good as the other treatment in all subsets of units, which has been studied extensively. For example, with $V$ discrete, the Gail–Simon test for qualitative interaction can be used to test hypotheses of the form (8) with a slight modification (Gail & Simon, 1985, p. 364). Chang et al. (2015) considered the problem with a general $V$ based on $\ell_1$-type functionals of uniformly consistent nonparametric kernel estimators of $\Delta^{01}(v)$. These tests make no assumptions on the functional form of $\text{pr}(Q^{01}=1 \mid Z=z,V=v)$, which is particularly appealing as $Q^{01}$ is not directly interpretable.

As we have four hypotheses of the form (7), a multiplicity adjustment is warranted. However, unlike the case with unconditional instrumental variable models, the four inequalities in (3) can be violated simultaneously, as each of them concerns multiple covariate values. In other words, no result analogous to Theorem 1 holds unless $V$ takes only one value. Instead, a naive Bonferroni correction may be used to account for multiple comparisons so that to get an overall level-$\alpha$ test, hypotheses of the form (7) are tested at level $\alpha/4$.

*Remark* 3. When $V$ is discrete, one can alternatively apply Theorem 1 to test, for each $v$, the hypotheses

$$\begin{aligned}\text{pr}(D=d,Y=y \mid Z=1,V=v) + \text{pr}(D=d,Y=1-y \mid Z=0,V=v) \leqslant 1\\ (d=0,1;\ y=0,1),\end{aligned} \tag{9}$$

and then use a Bonferroni correction to account for multiple testing due to levels of $V$. In this way, each hypothesis in (9) is tested at level $\alpha/(2K)$, where $K$ is the number of possible levels for $V$. Since tests of the forms (7) and (9) are different, neither approach generally dominates the other.

*Remark* 4. The Gail–Simon test examines the hypotheses (3) for all possible values of $V$ simultaneously. Alternatively, it may be tempting to test (3) for different levels of $V$ and claim that $Z$ is a valid instrument within the subset of the population for which the hypotheses (3) are not rejected. Failure to violate an instrumental variable inequality, however, does not prove that $Z$ is an instrument. This will ultimately rest on whether, based on subject-matter knowledge, we believe that we have measured enough

Table 1. *The p-values and numbers of subgroups obtained from partial tests for the binary instrumental variable models using college proximity as an instrument for education after high school*

| Covariate set $V$ | $\mathcal{H}_{0,c}^{00}$ | $\mathcal{H}_{0,c}^{01}$ | $\mathcal{H}_{0,c}^{10}$ | $\mathcal{H}_{0,c}^{11}$ | Number of subgroups |
|---|---|---|---|---|---|
| (I) | 1·000 | 0·010 | 1·000 | 0·034 | 24 |
| (II) | 1·000 | 0·132 | 1·000 | 0·143 | 47 |
| (III) | 1·000 | 1·000 | 1·000 | 1·000 | 819 |

covariates $V$ to control confounding, as well as subject-matter arguments for the absence of direct effects of $Z$ on $Y$.

Consequently, one should avoid using the test (3) as a way to restrict the range of $V$, unless a substantive argument could be made that $Z$ is an instrument for one range of $V$ but not for another.

## 4. THE CAUSAL EFFECT OF EDUCATION ON EARNINGS

We illustrate the use of the proposed tests by examining the instrumental variable model assumed by Okui et al. (2012). The goal of their analysis was to estimate the causal effect of education on earnings. To account for unobserved preferences for education levels, Okui et al. (2012) followed Card (1995) and used presence of a nearby four-year college as an instrument. The validity of this approach relies on the assumptions that college proximity affects earnings only through education and, conditional on adjusted potential confounders, college proximity is independent of underlying factors that also affect earnings. These assumptions, however, are hardly watertight. In fact, as pointed out by Card (1995), living near a college may influence earnings through higher elementary and secondary school quality, and it may also be associated with higher motivation to achieve labour market success.

To investigate the possible exogeneity of college proximity, we use the dataset provided by Okui et al. (2012), which contains 3010 observations from the National Longitudinal Survey of Young Men. Following Tan (2006), we consider education after high school as the treatment $D$. The outcome wage is dichotomized at its median. For illustrative purposes, we consider three instrumental variable models with nested sets of covariates: (I) experience only; (II) experience and race; (III) experience, race and region of residence. The third set was also considered previously by Okui et al. (2012). We use the Gail–Simon test with Bonferroni correction to examine the testable implications of these instrumental variable models.

Table 1 summarizes the test results. The model conditional only on experience is rejected by the proposed test. The $p$-value from the test on $\mathcal{H}_{0,c}^{01}$ is significant at the 0·05 level, and the $p$-value from the test on $\mathcal{H}_{0,c}^{11}$ is also borderline significant. These show that either college proximity has positive direct effects on earnings in some subgroups, or after adjusting for experience college proximity is still correlated with unmeasured confounders such as underlying motivation for labour market success. The proposed test fails to reject the instrumental variable model of Okui et al. (2012). However, as we discussed in Remark 4, with large sample sizes, failure to violate the instrumental variable inequalities shows that an instrumental variable model is compatible with the observed data, but does not validate such a model. Specifically, if one believes that the sample size is sufficiently large, then the results in Table 1 show that the instrumental variable model of Okui et al. (2012) is compatible with the observed data. One should use their model if one also believes that college proximity affects earnings only through education, and that there is no unmeasured confounding after adjusting for experience, race and region of residence. In contrast, one should not trust the instrumental variable model conditional only on experience, regardless of one's prior substantive belief.

## 5. DISCUSSION

Although instrumental variable methods are widely used to identify causal effects in the presence of unmeasured confounding, their assumptions have mainly been assessed based on subject-matter arguments

rather than statistical evidence. However, there are controversies about the validity of many instruments, especially if they are not randomized; for example, see Rosenzweig & Wolpin (2000) for a discussion on using natural experiments as instruments. Therefore, it should be routine to check the instrumental variable model against the observed data; see also Didelez et al. (2010). In this paper, we introduce a simple method for testing the binary instrumental variable model.

Our approach can be extended to test discrete instrumental variable models with binary outcomes. According to Pearl (1995), testable implications in this case include

$$\max\{p(0,d \mid 0),\ldots,p(0,d \mid z_{\max})\} + \max\{p(1,d \mid 0),\ldots,p(1,d \mid z_{\max})\} \leqslant 1 \quad (d=0,\ldots,d_{\max}), \quad (10)$$

where $Z$ takes values in $0,\ldots,z_{\max}$ and $D$ takes values in $0,\ldots,d_{\max}$. With slight modifications of the multiplicity adjustments, the techniques introduced in this paper can be used to test the inequalities (10); see the Appendix for details. In general, there are other observed-data constraints implied by the discrete instrumental variable model (Bonet, 2001), the testing of which is an interesting topic for future research.

Monotonicity is also often assumed in instrumental variable analysis. See Huber & Mellace (2015) for a joint test of the unconditional instrumental variable model and the monotonicity assumption. A future research problem would be to extend the proposed method to test the binary instrumental variable model under monotonicity.

Although we have focused primarily on testing the binary instrumental variable model, as we explain in § 2, with randomized experiments our proposed tests can be directly applied to identify the sign of the average controlled direct effects $\text{ACDE}(d) = E\{Y(z=1,d=d)\} - E\{Y(z=0,d=d)\}$ $(d=0,1)$. These average controlled direct effects quantify the extent to which the randomized treatment $Z$ affects the outcome $Y$ not through the mediator $D$, and are important for explaining causal mechanisms.

## Acknowledgement

We thank Chengchun Shi for helpful comments. This research was supported by the U.S. National Institutes of Health and Office of Naval Research. This work was initiated when the first author was a graduate student at the University of Washington.

## Supplementary material

Supplementary material available at *Biometrika* online includes code for the data application.

## Appendix

### *Proof of Theorem 1*

The second and third claims in Theorem 1 follow directly from the assumption that the size of $\phi^{dy}$ goes to zero asymptotically in the interior of the null space defined by $\mathcal{H}_0^{dy}$. We now consider the case where two inequalities in (1) hold with equality at the true value $\dot{\zeta}$. Without loss of generality, we assume $\dot{\Delta}^{00} = \dot{\Delta}^{01} = 0$, where the dot denotes the true value. As $\sum_{d,y} \dot{\Delta}^{dy} = -2$ and $-1 \leqslant \dot{\Delta}^{dy} \leqslant 0$ $(d=0,1;\ y=0,1)$, we immediately get that for $d=1$ and $y=0,1$, $\dot{\Delta}^{dy} = -1$ and hence $\text{pr}(Q^{dy}=1 \mid Z=1)=0$ and $\text{pr}(Q^{dy}=1 \mid Z=0)=1$. As a result, for $d=1$ and $y=0,1$ one cannot reject $\mathcal{H}_0^{dy}$, with probability 1. On the other hand, as at most one of $\mathcal{H}_0^{00}$ and $\mathcal{H}_0^{01}$ can be violated empirically, they cannot be rejected simultaneously given our assumptions on $\phi^{dy}$. The probability of rejecting at least one of $\phi^{00}$ and $\phi^{01}$ therefore equals $\alpha$ in this case.

### *Multiplicity adjustment with the discrete instrumental variable model*

The constraints in (10) can be written as

$$p(0,d \mid z_1) + p(1,d \mid z_2) \leqslant 1 \quad (z_1,z_2=0,\ldots,z_{\max},\ z_1 \neq z_2;\ d=0,\ldots,d_{\max}). \quad (A1)$$

There are $(d_{\max}+1)z_{\max}(z_{\max}+1)$ inequalities in (A1), the left-hand sides of which sum to $z_{\max}(z_{\max}+1)$. Hence at most $z_{\max}(z_{\max}+1)$ of them can hold with equality simultaneously. Similar to Theorem 1, the proposed testing procedure for the unconditional discrete instrumental variable model proceeds as follows: reject (A1) if for $z_1, z_2 = 0, \ldots, z_{\max}, z_1 \neq z_2$ and $d = 0, \ldots, d_{\max}$, at least one of the hypotheses in (A1), denoted as $\mathcal{H}_0^{dy}(z_1, z_2)$, is rejected by the corresponding $\phi^{dy}(z_1, z_2)$ at level $\alpha/\{z_{\max}(z_{\max}+1)\}$. For the conditional discrete instrumental variable model, the Bonferroni correction is appropriate; see also Remark 3.

## REFERENCES


ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–55.

BAIOCCHI, M., CHENG, J. & SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Statist. Med.* **33**, 2297–340.

BALKE, A. & PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Assoc.* **92**, 1171–6.

BERGER, R. L. & BOOS, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Am. Statist. Assoc.* **89**, 1012–6.

BONET, B. (2001). Instrumentality tests revisited. In *Proc. 17th Conf. Uncert. Artif. Intel.*, J. Breese & D. Koller, eds. San Francisco, California: Morgan Kaufmann Publishers, pp. 48–55.

CAI, Z., KUROKI, M., PEARL, J. & TIAN, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64**, 695–701.

CARD, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, L. N. Christofides, R. Swidinsky & E. K. Grant, eds. Toronto: University of Toronto Press, pp. 201–22.

CHANG, M., LEE, S. & WHANG, Y.-J. (2015). Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *Economet. J.* **18**, 307–46.

CLARKE, P. S. & WINDMEIJER, F. (2012). Instrumental variable estimators for binary outcomes. *J. Am. Statist. Assoc.* **107**, 1638–52.

DIDELEZ, V., MENG, S. & SHEEHAN, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25**, 22–40.

GAIL, M. & SIMON, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361–72.

HUBER, M. & MELLACE, G. (2015). Testing instrument validity for LATE identification based on inequality moment constraints. *Rev. Econ. Statist.* **97**, 398–411.

KANG, H., KREUELS, B., ADJEI, O., KRUMKAMP, R., MAY, J. & SMALL, D. S. (2013). The causal effect of malaria on stunting: A Mendelian randomization and matching approach. *Int. J. Epidemiol.* **42**, 1390–8.

LYDERSEN, S., FAGERLAND, M. W. & LAAKE, P. (2009). Recommended tests for association in $2 \times 2$ tables. *Statist. Med.* **28**, 1159–75.

OKUI, R., SMALL, D. S., TAN, Z. & ROBINS, J. M. (2012). Doubly robust instrumental variable regression. *Statist. Sinica* **22**, 173–205.

PEARL, J. (1995). Causal inference from indirect experiments. *Artif. Intel. Med.* **7**, 561–82.

PEARL, J. (2009). *Causality*. Cambridge: Cambridge University Press.

PERLMAN, M. D. & WU, L. (1999). The emperor's new tests. *Statist. Sci.* **14**, 355–69.

RAMSAHAI, R. & LAURITZEN, S. (2011). Likelihood analysis of the binary instrumental variable model. *Biometrika* **98**, 987–94.

RICHARDSON, T. S., EVANS, R. J. & ROBINS, J. M. (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statist.* **9**, 569–610.

ROSENZWEIG, M. R. & WOLPIN, K. I. (2000). Natural "natural experiments" in economics. *J. Econ. Lit.* **38**, 827–74.

SPIRTES, P., GLYMOUR, C. N. & SCHEINES, R. (2000). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press.

TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *J. Am. Statist. Assoc.* **101**, 1607–18.

VANSTEELANDT, S., BOWDEN, J., BABANEZHAD, M. & GOETGHEBEUR, E. (2011). On instrumental variables estimation of causal odds ratios. *Statist. Sci.* **26**, 403–22.


[*Received on* 12 *May* 2016. *Editorial decision on* 14 *November* 2016]