J. R. Statist. Soc. A (2005)
168, *Part* 1, *pp.* 109–126

Robust, accurate confidence intervals with a weak instrument: quarter of birth and education

Guido W. Imbens

University of California, Berkeley, USA

and Paul R. Rosenbaum

University of Pennsylvania, Philadelphia, USA

[Received November 2001. Final revision October 2003]

Summary. An instrument or instrumental variable manipulates a treatment and affects the outcome only indirectly through its manipulation of the treatment. For instance, encouragement to exercise might increase cardiovascular fitness, but only indirectly to the extent that it increases exercise. If instrument levels are randomly assigned to individuals, then the instrument may permit consistent estimation of the effects caused by the treatment, even though the treatment assignment itself is far from random. For instance, one can conduct a randomized experiment assigning some subjects to 'encouragement to exercise' and others to 'no encouragement' but, for reasons of habit or taste, some subjects will not exercise when encouraged and others will exercise without encouragement; none-the-less, such an instrument aids in estimating the effect of exercise. Instruments that are weak, i.e. instruments that have only a slight effect on the treatment, present inferential problems. We evaluate a recent proposal for permutation inference with an instrumental variable in four ways: using Angrist and Krueger's data on the effects of education on earnings using quarter of birth as an instrument, following Bound, Jaeger and Baker in using simulated independent observations in place of the instrument in Angrist and Krueger's data, using entirely simulated data in which correct answers are known and finally using statistical theory to show that *only* permutation inferences maintain correct coverage rates. The permutation inferences perform well in both easy and hard cases, with weak instruments, as well as with long-tailed responses.

Keywords: Hodges–Lehmann estimate; Instrumental variable; Observational study; Permutation test; Randomization test

1. Introduction: the need for greater realism with weak instruments

1.1. Instrumental variables: definition, goal and finding instruments

Instrumental variable analyses are designed to estimate effects of treatments when the level of the treatment is confounded by unobserved covariates that cannot be controlled by adjustments. An instrument manipulates a treatment without fully controlling it, but the instrument itself does not have effect beyond its manipulation of the treatment. For a recent survey of instrumental variables from several perspectives, see Angrist *et al.* (1996) and the associated discussion by Robins, Greenland, Heckman, Moffitt and Rosenbaum.

In a simple experiment, the experimenter manipulates the treatment, assigning it to subjects at random, and the treatment has effects. Here, the treatment is playing two roles: it is both what the experimenter manipulates and also the effectual aspect of that manipulation. The notion of

Address for correspondence: Paul R. Rosenbaum, Department of Statistics, 473 Jon Huntsman Hall, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA.
E-mail: rosenbaum@stat.wharton.upenn.edu

 0964–1998/05/168109

an instrumental variable separates these two roles: the experimenter manipulates the instrument, perhaps encouragement to exercise, but believes that it is not the instrument, not encouragement, but rather another measured quantity, the treatment, exercise, that affects cardiovascular health. Here, the experimenter is interested in the effects of the treatment itself but has only a partial indirect influence over the level of treatment that is received.

In practice, instruments are most commonly used in observational studies of treatments, not experiments, in settings in which subjects select their own treatments, often with specific goals in mind. Because the treatments are self-inflicted, the individuals who choose one treatment are often very different, before treatment, from the individuals who choose another. The example in this paper concerns additional schooling, which to a large extent is determined by an individual's own preferences and decisions. People who stay in school longer, obtaining advanced degrees, may reap economic benefits arising from additional education, but they may also reap economic benefits from the ambition, perseverance or talent that led them initially to stay in school longer. Measures of ambition, perseverance and talent are at best imperfect and are often unavailable. In this context, an instrument is something of no direct economic consequence, something unrelated to ambition, perseverance and talent, but something that lengthens education for some and shortens it for others. What might that be?

Finding instruments is an art rather than a science, and Angrist and Krueger (2001) have surveyed a variety of interesting and suggestive examples, from lotteries to administrative policy discontinuities to accidents of birth. Even in the best of situations, the assumptions that are required for an instrument are plausible, not certain. For instance, although one might plausibly expect encouragement to exercise to affect cardiovascular health only indirectly through exercise, that might not be true. If encouragement emphasizes future health benefits, pride, self-discipline, physical attractiveness, etc., then encouragement that is not heeded might lead to anxiety about future illness, stress, a sense of failure or defeat, and these might possibly affect cardiovascular health. If the assumptions are merely plausible, not certain, then of what value are efforts to find instrumental variables?

There are two issues, both important. First and more obviously, the use of an appropriate instrument in a difficult setting is intended to replace an implausible assumption by a plausible assumption, albeit not a certain assumption. Second, and arguably more importantly, compare two sequences of studies of the same self-inflicted treatment: one sequence without instruments comparing treated with untreated; the other sequence using a variety of different instruments to manipulate the treatment. Throughout the first sequence, the comparison is likely to be biased in the same way. A repeated finding that people with more education earn more than people with less education does little to isolate the effects that are caused by education from the consequences of unmeasured ambition, perseverance and talent that led to extended education—the same bias appears repeatedly. However, if different instruments are used to manipulate education—a lottery, a temporal discontinuity in educational policy or a regional discontinuity in educational policy—and if each instrument is plausible but not certain, there may be no reason why these different instruments should be biased in the same direction. In replicating observational studies, the goal is to replicate whatever treatment effects may exist without replicating whatever biases may coexist (Rosenbaum, 2001), and this goal is sometimes achievable by using a variety of instruments or adjustment strategies subject to different biases. The literature in labour economics on estimating the returns to education is a nice example of this, where researchers have used instrumental variables strategies (e.g. compulsory schooling laws as in the current paper or distance to educational institutions), as well as strategies that are based on direct adjustment by using cleverly selected samples (e.g. twins); see Card (2001) for an overview.

1.2. Strong instruments: non-compliance in trials and the draft lottery

One of the most compelling instances of an instrument occurs in randomized trials with non-compliance (e.g. Sommer and Zeger (1991), Sheiner and Rubin (1995), Goetghebeur and Molenberghs (1996), Imbens and Rubin (1997) and Heitjan (1999)). Here, the instrument—the assigned treatment—is truly random; however, the assigned treatment influences, but does not completely determine, the treatment that is actually received. Holland (1988) has called this an encouragement design: subjects are randomly selected and encouraged to take the treatment, but it is the effects of the treatment itself, not the effects of encouragement, which are of interest; see also Zelen (1979). Randomization of the instrument is not, by itself, sufficient; to be valid, an instrument must affect the outcome only by manipulating the treatment—for example, encouragement must not have effects of its own. When subjects are assigned to one of two conditions, treatment or control, and receive one of two conditions, treatment or control, it is tempting to estimate the treatment effect as the typical or mean difference in outcomes in two groups, suitably defined; however, this does not generally work. It is easy to show the following.

(a) If the assigned treatment is ignored, then the mean difference in outcomes in groups defined by the treatment received can be severely biased as an estimate of the effect caused by the treatment.
(b) If non-compliers—those who did not follow the assignment that they received—are set aside, and the two groups are defined by acceptance of the treatment to which they were assigned—so-called per protocol analysis—then the mean difference in outcomes in these two groups can again be severely biased for the treatment effect (Sheiner and Rubin, 1995).
(c) If the two groups are defined by the assigned treatment, ignoring the treatment that they received—the so-called intent-to-treat analysis—then a comparison of the two groups provides a valid test of no effect, but it can substantially underestimate the effects of the treatment, because many people who were assigned to the treatment did not receive it.
(d) In contrast, methods that do not compare two groups, methods that use both the assigned treatment and the received treatment in different roles—the instrumental variable analysis—can continue to use randomization at the 'reasoned basis for inference', in Fisher's (1935) phrase, while yielding an undiluted estimate of the treatment effect (Rosenbaum (1996, 1999), Rosenbaum (2002a), chapter 5, and Rosenbaum (2002b)). The approach, which is developed in a slightly more general context in Section 3.2, says that the treatment effect is a function of the treatment that is actually received, and once that effect has been removed from responses the responses are independent of the treatment that was randomly assigned.

A compelling example of an observational or non-randomized study using an instrument is Angrist's (1990) study of the effects of military service during the Vietnam War and its effects on lifetime earnings. Angrist used the draft lottery number as an instrument for the actual treatment, military service. As in the case of assignment in randomized trials with non-compliance, the draft lottery was essentially random, and it encouraged but did not determine military service, because of volunteers and draft evaders.

In observational studies, although we try to find an instrument, such as the draft lottery, that is not biased in its assignment to subjects, the possibility remains that even the instrument is not randomly assigned. This possibility is addressed through sensitivity analyses, which are discussed and illustrated in Rosenbaum (1996, 1999), Rosenbaum (2002a), chapter 5, and Rosenbaum (2002b) but will not be further discussed here.

An instrument is *weak* if manipulation of the instrument has only a slight effect on the treatment (Staiger and Stock, 1997). Weak instruments are common but create inferential problems that we address in the current paper.

1.3. Weak instruments: quarter of birth and education

Angrist and Krueger (1991) found a clever but weak instrument for years of schooling. School years begin in September, but children are born all year long, so children throughout a 1-year period all begin school together in September. Some students may drop out of school as soon as the law allows, typically when they reach a specified age. This means that a child's month or quarter of birth may force one child to attend school for up to a year longer than another. The instrument, quarter of birth, is plausibly haphazard, but the treatment itself, total years of schooling, is subject to severe systematic biases. The instrument is quite weak, adding on average a tenth of a year of schooling. None-the-less, Angrist and Krueger (1991) showed that both years of schooling and earnings do track quarter of birth in the zigzag pattern that is consistent with their argument.

Angrist and Krueger (1991) did a variety of analyses, some of which have held up well over the years, but others have been subjected to sharp yet illuminating criticism. Their simplest analyses incorporated a single instrument, quarter of birth, using Wald's (1940) estimator. In the Wald estimator, the mean difference in log-earnings in two quarters is divided by the mean difference in years of education. Angrist and Krueger found about a 7% increase in earnings for a year of schooling by using the 1970 census data for men born between 1920 and 1929. That analysis has held up fairly well. They then built two simultaneous equations for education and log-earnings, with numerous indicator variables for states and years, fitting these by two-stage least squares (TSLS) with many instruments created by interacting quarter of birth with year and state of birth. These analyses have generated debate. Bound *et al.* (1995) replaced the actual instruments, the quarter-of-birth dummy variables, by useless, randomly generated values in some of the TSLS regressions, but none-the-less obtained qualitatively similar results, with small standard errors and short confidence intervals, suggesting that the TSLS estimates can be highly misleading. In Section 2, we review Angrist and Krueger's study, and in Section 4 we apply our approach to their data.

1.4. Two problems with a weak instrument

Weak instruments present two entirely distinct problems. First, if the instrument is extremely weak, it may provide little or no useful information. An accurate method of analysis will correctly report this. Some commonly used statistical methods for instrumental variables are not accurate in this sense, and this is the second problem. With weak instruments, asymptotic approximations for standard errors and confidence intervals often wrongly suggest that an unstable estimate is very stable.

Commonly used methods assume that the problem is identified and apply asymptotic theory, but this asymptotic theory is inapplicable with an uninformative instrument, and it performs poorly when the instrument is weak. Nelson and Startz (1990) and Maddala and Jeong (1992) demonstrated substantial deviations of finite sample and asymptotic distributions of instrumental variable estimates, and Bound *et al.* (1995) illustrated this with the quarter-of-birth data. Han and Schmidt (2001) showed that, with irrelevant instruments, instrumental variable estimates are asymptotically centred on least squares estimates. Various improvements have been proposed, including the alternative asymptotics of Bekker (1994) and Staiger and Stock (1996), the jackknife methods of Angrist and Krueger (1995) and Angrist *et al.* (1999), the tests based

on the pivotal statistics of Kleibergen (2002) and Moreira (2003) and the hierarchical Bayes methods of Chamberlain and Imbens (1996).

In this paper we examine the performance with weak instruments of an alternative approach, a slight extension of the procedure that was proposed in Rosenbaum (1996, 1999), Rosenbaum (2002a), chapter 5, and Rosenbaum (2002b). The method solves the second problem, the problem with statistical methods, although of course it does not solve the first problem—it cannot make uninformative data informative.

1.5. Outline of the paper: quarter-of-birth example, simulation and theory

After reviewing aspects of Angrist and Krueger's (1991) quarter-of-birth data in Section 2, the current paper evaluates the performance of this permutation approach in four ways. A slightly extended version of the method is defined in Section 3, and then, in Section 4, we apply the method to Angrist and Krueger's quarter-of-birth data. We also follow Bound *et al.* (1995), creating a non-informative variant of the quarter-of-birth data. In both sets of data, the method does well. With the actual data of Angrist and Krueger (1991), the permutation analysis analogous to the Wald estimator yields a confidence interval that is quite short, not unlike the interval that was found by Angrist and Krueger. However, with a useless instrument, the 95% confidence intervals are wide and uninformative, as they should be when identification is lacking. When adjustments are made for state and year with the real quarter-of-birth data, the interval is informative, but longer than Angrist and Krueger's interval, a finding that is consistent with the discussion of Bound *et al.* (1995).

A small simulation is presented in Section 5. The simulation covers various settings, including some where standard inference performs poorly similarly to those studied by Nelson and Startz (1990) and Maddala and Jeong (1992). We find that the permutation method performs very well in a wide range of settings, with the asymptotic approximations very accurate even with weak instruments, and power high compared with that for TSLS methods when error distributions are thick tailed.

Finally, in Section 6, we show that the *only* accurate, nonparametric methods for instrumental variables are permutation methods. The result is a slight extension of a famous theorem due to Lehmann and Stein (1949) and Lehmann (1959), section 5.7, who considered the case in which the treatment, not the instrument, is randomized.

2. Review: Angrist and Krueger's (1991) study of the returns to education

2.1. Data and assumptions: census microdata, identification and the exclusion restriction

Angrist and Krueger (1991) used data from the public use data files, describing samples of individuals, from the US censuses of 1960, 1970 and 1980, with sample sizes that were greater than 200 000 for men born between 1920 and 1929, greater than 300 000 for men born between 1930 and 1939 and greater than 400 000 for men born between 1940 and 1949; however, the relevant sample size varies somewhat depending on the details of the different analyses that they performed. They also obtained data from the 50 states and the District of Columbia concerning laws about compulsory school attendance in 1960, 1970 and 1980.

Angrist and Krueger (1991) carefully built an argument claiming that laws which require students to begin school in September but to attend school until a particular birthday, say the 16th birthday, are the cause of a small amount of variation in the number of years of school attended. The claim is that some students drop out of school as soon as the law allows, with a few months less education for students who are a few months older. First, they showed that

the mean years of schooling follows a saw-tooth pattern, typically dipping down in the first quarter of the year, for the oldest students in a given annual class. Second, they showed that this pattern largely disappears when attention focuses on subsamples who completed high school, college or advanced degrees, suggesting that quarter of birth matters only during high school, not subsequently. Third, they compared states which require different birth dates (16th *versus* 17th or 18th) and students of different ages, demonstrating that a decline in enrolment abruptly occurs at the minimum legal age for leaving school. This first claim—that the quarter of birth causes a small amount of variation in years of education due to the structure of minimum age laws—has been largely undisputed in subsequent literature.

A second claim is required if quarter of birth is to be an instrument, namely the *exclusion restriction*, which asserts that the quarter of birth is related to earnings *only* because it affects years of education. For instance, an obvious concern is that children who are born in the first quarter are a few months older than other children who enter school at the same time, and at very young ages a difference of a few months might be an advantage in performance in school; see Halliwell (1966), Angrist and Krueger (1991) and Bound *et al.* (1995). However, such an effect would predict higher earnings for the children who were born in the first quarter, whereas the effect of compulsory schooling laws would predict lower earnings for the first quarter, and Angrist and Krueger (1991) found lower earnings in the first quarter. As a result, if there is a bias due to age, Angrist and Krueger (1991), page 1007, suggested that the bias would lead to underestimates of the returns to schooling. In addition, Angrist and Krueger (1991), page 1008, showed that quarter of birth predicts earnings for the population, but not in the subpopulation of college graduates; this pattern is not easily explained by the claim that being a little older at the start of schooling has substantial, long lasting effects on school performance.

2.2. Analytical tools: Wald's estimate and two-stage least squares

In TSLS, the so-called 'endogenous' variable, here years of education, is first regressed on the instrument or instruments, here quarter of birth, together with exogenous variables, such as age or state. Then the dependent variable, here log-earnings, is regressed on the fitted values of the endogenous variable, here years of education as predicted by quarter of birth, together with the exogenous variables; see Amemiya (1985) for general discussion. When the instrument is a single binary variable and there are no other exogenous variables, Durbin (1954) showed that TSLS is the same as Wald's (1940) method of fitting a line when the predictor is subject to error.

In Angrist and Krueger (1991), the economic return to an additional year of schooling is estimated several times, always by using TSLS. In their simplest application of the method, they compared men who had been born in the first quarter with other men by using Wald's estimator, producing an estimated 10.2% increase in weekly wages for a year of additional education, with an estimated standard error of about $\pm 2.4\%$ for 327 509 men in the 1980 census who were born between 1930 and 1939. The claim that quarter of birth is an instrument is more plausible within fairly homogeneous cohorts of men, say men born in the same year in the same state, so that

> 'the variability in education used to identify the return to education in the TSLS estimates is solely due to differences by season of birth'

(Angrist and Krueger (1991), page 1004). In one of several similar analyses, Angrist and Krueger included indicators of year of birth and state of birth, as well as age (measured in quarters) and age^2, with 30 instruments formed as the product of year-of-birth indicators and quarter-of-birth indicators, and 150 instruments formed as the product of state-of-birth indicators and quarter-of-birth indicators. Using this approach, Angrist and Krueger (1991), Table VII, column 4, reported an estimated 9.1% increase in weekly wage per year of education with a standard

error of ±1.1%, so the point estimate is changed only slightly, but the standard error is reduced by more than 50%. From this, it appears that the use of many instruments in TSLS has confirmed the magnitude of the simpler Wald estimate but has greatly enhanced precision. Is this appearance accurate?

Bound *et al.* (1995), pages 448–449, raised concerns about TSLS with many weak instruments. Specifically, using exactly the same data, they replaced the quarter-of-birth information by entirely irrelevant random numbers, which should be useless for estimating the return to education, repeating this process 500 times, applying TSLS each time, obtaining a mean estimated return of a 6.0% increase in weekly wages for a year of additional education, with an estimated standard error of about ±1.5%. Moreover, by varying the instruments, they found greater apparent precision with more instruments than with fewer instruments, even though the instruments are all just irrelevant random noise. Clearly, something went wrong.

In the second stage of TSLS, as the number of instruments increases, the fitted years of education increasingly resembles the actual years of education, and the TSLS estimate increasingly resembles the usual least squares regression of wages on years of education. This particular problem with TSLS is purely technical. A confidence interval promised a certain rate of coverage and failed because the associated asymptotic approximations are poor when the instrument is weak. It is this technical problem that the literature on weak instruments (e.g. Staiger and Stock (1997), Kleibergen (2002) and Moreira (2003)) is concerned with, and that we address in the current paper by using randomization inference. We return to the quarter-of-birth data in Section 4 after defining the proposed method in Section 3.

3. Review: randomization inference with an instrument

3.1. Notation: strata, responses and instruments

There are S strata, $s = 1, \ldots, S$, with n_s subjects in stratum s and $N = n_1 + \ldots + n_S$ subjects in total. There is dose of treatment d, with one value labelled '0' signifying the 'control' or reference level. The ith subject in stratum s would exhibit response r_{Csi} if this subject received the control dose, $d = 0$, and would exhibit response r_{dsi} if this subject received dose d. In the current paper, the effect of receiving dose d rather than the control dose $d = 0$ is assumed to be proportional to the dose, $r_{dsi} - r_{Csi} = \beta d$. In Angrist's (1990) study of the draft, d is a binary variable indicating military service. In Angrist and Krueger (1991), d is years of education beyond the minimum that are required by law.

In stratum s there is a preset, sorted, fixed list of n_s instrument settings, h_{sj}, $j = 1, \ldots, n_s$, with $h_{sj} \leqslant h_{s,j+1}$ for each s and j. Write $\mathbf{h} = (h_{11}, h_{12}, \ldots, h_{1,n_1}, h_{21}, \ldots, h_{S,n_S})^{\mathrm{T}}$. In the draft lottery in Angrist (1990), there is only one stratum, $S = 1$, and $0 = h_{11} = \ldots = h_{1k}$ and $1 = h_{1,k+1} = \ldots = h_{1,n_1}$, where the draft lottery would divide the n_1 men who were eligible for the draft into k who were not drafted and $n_1 - k$ who were drafted. Instrument settings in $\mathbf{h}$ are randomly permuted within strata and assigned to subjects, i.e. the lottery picked draftees at random. An assignment of instrument settings, $\mathbf{z}$, is $\mathbf{z} = \mathbf{ph}$ where $\mathbf{p}$ is a stratified permutation matrix, i.e. an $N \times N$ block diagonal matrix with S blocks, $\mathbf{p}_1, \ldots, \mathbf{p}_S$. Block $\mathbf{p}_s$ is an $n_s \times n_s$ permutation matrix, i.e. $\mathbf{p}_s$ is a matrix of 0s and 1s such that each row and each column sum to 1. Let Ω be the set of all stratified permutation matrices $\mathbf{p}$, so Ω is a set containing $|\Omega| = \Pi_{s=1}^{S} n_s!$ matrices, where $|\Omega|$ denotes the number of elements of the set Ω. Pick a random $\mathbf{P}$ from Ω where $\Pr(\mathbf{P} = \mathbf{p}) = 1/|\Omega|$ for each $\mathbf{p} \in \Omega$. Then $\mathbf{Z} = \mathbf{Ph}$ is a random permutation of $\mathbf{h}$ within strata, so the ith subject in stratum s received instrument setting Z_{si}. In effect, the draft lottery picked $\mathbf{P} \in \Omega$ at random and computed $\mathbf{Z} = \mathbf{Ph}$, and drafted the $n_1 - k$ men with $Z_{1i} = 1$. Violations of random instrument settings are

addressed by sensitivity analysis; see Rosenbaum (1999), Rosenbaum (2002a), chapter 5, and Rosenbaum (2002b).

For each such instrument setting $\mathbf{z}$ there is a dose $d_{si\mathbf{z}}$ that would be received by the ith subject in stratum s, who then exhibits response $r_{Csi} + \beta d_{si\mathbf{z}}$. In Angrist (1990), for a given pattern of draft lottery results, $\mathbf{z}$, the variable $d_{1i\mathbf{z}}$ indicates whether the ith man would serve in the military. Write D_{si} for the dose that is exhibited by the ith subject in stratum s, so $D_{si} = d_{si\mathbf{Z}}$, and let R_{si} be the response that is observed from this subject, so $R_{si} = r_{Csi} + \beta D_{si}$. Write $\mathbf{D} = (D_{11}, \ldots, D_{S,n_S})^{\mathrm{T}}$ and $\mathbf{R} = (R_{11}, \ldots, R_{S,n_S})^{\mathrm{T}}$.

In randomization inference, as developed by Fisher (1935) and later researchers (e.g. Pitman (1937), Welch (1937), Kempthorne (1955), Wilk (1955), Cox (1958), Robinson (1973), Tukey (1985), Gail *et al.* (1988) and Cox and Reid (2000)), a quantity whose value is determined by the random choice of $\mathbf{P}$ from Ω is a random variable because $\mathbf{P}$ is random. In contrast, a quantity that is not affected by the random choice of $\mathbf{P}$ from Ω is a fixed quantity describing the finite population of N subjects. For instance, the observed response R is a random variable because it depends on D which in turn depends on the random instrument settings $\mathbf{Z} = \mathbf{Ph}$; however, the potential response r_{Csi} that a subject would exhibit at dose 0 does not change with $\mathbf{P}$ and so is fixed. In this way, randomization creates the probability distributions that are used in inference and is the 'reasoned basis for inference' in randomized experiments, in Fisher's (1935) phrase. For an alternative view, see Sections 5 and 6. In contrast, in most econometric analyses with weak instruments, the potential responses r_{Csi} are viewed as random, and the analysis is conditional on the instruments Z (e.g. Staiger and Stock (1997)).

3.2. Randomization inference

Consider testing the hypothesis $H_0 : \beta = \beta_0$. Let $\mathbf{q}(\cdot)$ be some method of scoring responses, such as their ranks within strata or the aligned ranks of Hodges and Lehmann (1962), and let $\boldsymbol{\rho}(\mathbf{Z})$ be some way of scoring the instrument settings such that $\boldsymbol{\rho}(\mathbf{ph}) = \mathbf{p}\,\boldsymbol{\rho}(\mathbf{h})$ for each $\mathbf{p} \in \Omega$, e.g. Rosenbaum (1991). The test statistic is $T = \mathbf{q}(\mathbf{R} - \beta_0\mathbf{D})^{\mathrm{T}}\,\boldsymbol{\rho}(\mathbf{Z})$ and, for appropriate scores, T can be Wilcoxon's stratified rank sum statistic, the Hodges–Lehmann aligned rank statistic, the stratified Spearman rank correlation or the stratified version of Mood's median test statistic.

If H_0 were true, then $\mathbf{R} - \beta_0\mathbf{D} = \mathbf{r}_C$ would be fixed, not varying with $\mathbf{Z}$, so $\mathbf{q}(\mathbf{R} - \beta_0\mathbf{D}) = \mathbf{q}(\mathbf{r}_C) = \mathbf{q}$, say, would also be fixed. If the null hypothesis were false, $\beta \neq \beta_0$, then $\mathbf{R} - \beta_0\mathbf{D} = \mathbf{r}_C + (\beta - \beta_0)\mathbf{D}$ continues to be related to the dose $\mathbf{D}$, and possibly as a consequence related to the instruments $\mathbf{Z}$. We hope to recognize a correct or approximately correct value β_0 for β by an absence of a relationship between $\mathbf{R} - \beta_0\mathbf{D}$ and $\mathbf{Z}$.

An exact test of hypothesis $H_0 : \beta = \beta_0$ computes $\mathbf{q}(\mathbf{R} - \beta_0\mathbf{D})$, which is the fixed value $\mathbf{q} = \mathbf{q}(\mathbf{r}_C)$ if H_0 is true, in which case $T = \mathbf{q}^{\mathrm{T}}\mathbf{P}\,\boldsymbol{\rho}(\mathbf{h})$. The chance that $T \geqslant t$ under H_0 is simply the proportion of $\mathbf{p} \in \Omega$ such that $\mathbf{q}^{\mathrm{T}}\mathbf{p}\,\boldsymbol{\rho}(\mathbf{h}) \geqslant t$, or

$$\frac{|\{\mathbf{p} \in \Omega : \mathbf{q}^{\mathrm{T}}\mathbf{p}\,\boldsymbol{\rho}(\mathbf{h}) \geqslant t\}|}{|\Omega|}. \tag{1}$$

Using the known null distribution (1) of T, an exact, distribution-free $100(1-\alpha)\%$ confidence set for β is the set of all hypotheses $H_0 : \beta = \beta_0$ that are not rejected at level α. Moreover, as we discuss in Section 6, this is the *only* basis for distribution-free inference. Typically, this confidence set is formed by rejecting hypothesis H_0 for either large or small T with each tail having null probability $\alpha/2$ by expression (1).

The exact confidence set addresses the issue of identification in a natural way. The exact $100(1-\alpha)\%$ confidence set for β always has coverage $100(1-\alpha)\%$ but, when identification is

absent, so that the data are without useful information, the interval may achieve this coverage by becoming infinite in length. Nonparametric confidence intervals of infinite length are not new: the 95% confidence set for the upper 1% point of a distribution based on an independent and identically distributed sample of size 20 will of necessity not be a finite interval, but rather a half-line, correctly reflecting the obvious fact that 20 observations place a lower limit but not an upper limit on the upper 1% point of an unspecified distribution. Weak identification may result in a very long confidence set. An attractive feature of the method is that speculation about identification is replaced by a confidence set that is long or short depending on the evidence that is actually available in the data at hand.

If identification is lacking because the instruments are irrelevant, then the exact $100(1-\alpha)\%$ confidence set for β derived from expression (1) will none-the-less maintain its stated coverage of $100(1-\alpha)\%$. The confidence set may do this by including the entire real line, but it need not include the entire real line. The promise, after all, is not coverage 100% of the time, but rather $100(1-\alpha)\%$ coverage.

The confidence set can be empty: the test may reject every value of β_0, i.e. $\mathbf{R}-\beta_0\mathbf{D}$ may be significantly related to the instrument $\mathbf{Z}$ for every choice of β_0. This is strong evidence of a specification error, i.e. evidence that the effect of the treatment is not correctly modelled by a multiple of dose, perhaps because the exclusion restriction is false, so that the instrument directly affects the response. Indeed, the rule—reject the specification if the confidence set is empty—is a particular type of exact specification test, one that falsely rejects a correct specification with probability at most α.

The exact confidence set for β derived from expression (1) will often be an interval, but it need not be. If we desire an interval, we can *define* the $100(1-\alpha)\%$ confidence interval to be the shortest interval that includes the $100(1-\alpha)\%$ confidence set—adding points to a confidence set cannot decrease its coverage probability.

3.3. Exact moments and approximate distributions

The expectation and variance of T in proposition 1 are used in

(a) a large sample approximation to the null distribution of T and
(b) the estimating equation that defines the Hodges and Lehmann (1963) estimate of β: essentially, we solve $T=E(T)$ for β; see Hodges and Lehmann (1963) for details. The proof is analogous to that of theorem 3.3.3 of Hájek *et al.* (1999) and is omitted.

Proposition 1. Under the null hypothesis $H_0:\beta=\beta_0$, the expectation and variance of $T=\mathbf{q}^{\mathrm{T}}\boldsymbol{\rho}(\mathbf{Z})$ with $\mathbf{q}=\mathbf{q}(\mathbf{R}-\beta_0\mathbf{D})$ are

$$E(T)=\mu=\sum_{s=1}^{S}\bar{q}_s\bar{\rho}_s, \tag{2}$$

and

$$\mathrm{var}(T)=\sigma^2=\sum_{s=1}^{S}\frac{1}{n_s-1}\left\{\sum_{i=1}^{n_s}(q_{si}-\bar{q}_s)^2\right\}\sum_{i=1}^{n_s}(\rho_{si}-\bar{\rho}_s)^2, \tag{3}$$

where $\bar{q}_s=(1/n_s)\,\Sigma_{i=1}^{n_s}q_{si}$ and $\bar{\rho}_s=(1/n_s)\,\Sigma_{i=1}^{n_s}\rho_{si}$.

In large samples, with either a few large strata or many small strata, $(T-\mu)/\sigma$ is approximately standard normal under very mild conditions on the rank score functions $\mathbf{q}(\cdot)$ and $\boldsymbol{\rho}(\cdot)$. (Formally, if each $n_s\to\infty$, with S fixed, then apply theorem 6.1.1 of Hájek *et al.* (1999), whereas, if $N\to\infty$ and $S\to\infty$ with n_s bounded, then apply standard central limit theorems.)

The limiting normal null distribution of $(T-\mu)/\sigma$ is based on the limiting properties of permutation tests and does not depend on whether β is identified. Identification only arises when the test is inverted to obtain confidence intervals and point estimates, which may not behave as if constructed from an asymptotically normal estimate of β. In particular, without identification, confidence intervals may include the entire real line, and the moment equation that defines the Hodges–Lehmann estimate may admit a wide range of solutions. Weak or absent identification does not disrupt the asymptotic normal null distribution of the test statistic, so inferences that are based on this distribution remain valid, but inversion of the distribution to make inferences about β may reveal that the data provide little or no information about β.

4. An application to the Angrist–Krueger quarter-of-birth data

We apply the permutation method to estimate returns to schooling in the data of Angrist and Krueger (1991). There is an absolute minimum number δ of years of schooling that every student in a state must have, regardless of birth date; however, because birth dates vary, individual students are *required* to attend between δ and $\delta+1$ years of schooling, with most students receiving many years more than required.

Let D be the number of years of schooling beyond δ. Let R be log-earnings actually achieved with $D+\delta$ years of schooling, and let r_C be log-earnings with δ years of schooling. Angrist and Krueger (1991) focused on a model in which log-earnings increase linearly with years of schooling, so that $R=r_C+\beta D$. The log-transformation is motivated by technical and conceptual issues. The distribution of earnings is extremely skewed, with a small number of men earning very large amounts, and the log-transformation reduces the skewness. On the log-scale, the coefficient β may be interpreted as a relative increase or rate of return to a year of education. If log-earnings R are regressed on years of schooling, $D+\delta$, the estimated coefficient is 0.0709 with standard error 0.0003. This suggests that we predict about 7% higher earnings associated with an additional year of schooling; however, this prediction does not distinguish the effect of schooling on earnings from the tendency of brighter, wealthier, better motivated students both to stay in school and subsequently to earn more.

Angrist and Krueger (1991) used quarter-of-birth indicators, denoted by z, as an instrument, and used census data for estimation. In this section we use their data on 329509 men born between 1930 and 1939. We observe their years of schooling, D, log-earnings in 1980, R, and state and year of birth. The mean number of years of education is 12.75. The instrument that we use is an indicator for being born in the fourth quarter of the year, which is 1 for 24.5% of men and 0 for the rest. If the number of years of education is regressed on this quarter-of-birth indicator, the least squares regression coefficient is 0.092 with standard error 0.013, so being born in the fourth quarter of the year is associated with, on average, about a tenth of a year of additional education, an association that is small but clearly not due to chance. Moreover, if log-earnings are regressed on the quarter-of-birth indicator, the coefficient is 0.0068 with standard error 0.0027, so being born in the fourth quarter is associated with about $\frac{2}{3}$% higher earnings, which is a very weak relationship.

First we consider estimators ignoring information on year and state of birth in Table 1. Table 1 gives the point estimate $\hat{\beta}$, the upper and lower end points of the 95% confidence interval and a pseudostandard error. For the permutation procedures, $\hat{\beta}$ is the Hodges–Lehmann estimate, and the confidence interval is based on the large sample approximation in Section 3.3. The 'standard error' is the length of the confidence interval divided by 2×1.96, so it focuses attention on the length rather than the location of the confidence intervals. For instance, conventional TSLS estimates about 7.4% higher earnings with an additional year of schooling, although the confi-

Table 1. Comparison of instrumental variable estimates without covariates in the quarter-of-birth data

Procedure	$\hat{\beta}$	*95% confidence interval low*	*95% confidence interval high*	*'Standard error'*
TSLS	0.074	0.019	0.129	0.028
Permute log-earnings	0.073	0.017	0.132	0.029
Permute ranks	0.058	0.014	0.102	0.023

Table 2. Comparison of instrumental variable estimates with year and state covariates in the quarter-of-birth data

Procedure	$\hat{\beta}$	*95% confidence interval low*	*95% confidence interval high*	*'Standard error'*
TSLS	0.074	0.058	0.090	0.008
Permute log-earnings	0.077	0.036	0.139	0.026
Permute ranks	0.067	−0.015	0.162	0.045

dence interval ranges from about 2% to about 13%. The three estimates are quite similar, but the rank estimates are slightly lower, with a slightly narrower confidence interval, perhaps because of some individuals with extremely high log-earnings. Combined with what we know from the simulations in Bound *et al.* (1995), Table 1 suggests that all three estimates are performing reasonably, as theory suggests they should.

Table 2 adjusts for differences in year of birth and state. Because lifetime earnings and education vary with both state and cohort, we would be hesitant to attribute to education variations in earnings that could be predicted from state and year. In the first row of Table 2, TSLS estimates were performed as in Angrist and Krueger (1991), with state and year dummy variables as covariates and interactions of these dummy variables with the quarter-of-birth variable as additional instruments. Now it is well documented in the literature that adding many instruments in this way inaccurately appears to increase the precision—see Bound *et al.* (1995) and Staiger and Stock (1997)—so the much narrower confidence interval here is neither a surprise nor a comfort. The two randomization-based estimators take account of the covariates by using the permutation distribution within strata defined by year and state. The randomization-based confidence intervals are not extremely narrow, and the rank-based interval includes slightly negative returns as well as markedly positive returns to a year of schooling. If we believed the rank-based intervals, it would suggest that the pattern in Table 1 might be due to variations in earnings and education that can be predicted from state and year alone.

In Table 3, we replace the actual quarter-of-birth variable by a randomly generated instrument that carries no information because it is unrelated to years of education. With a useless instrument, the data contain no information about β, and an accurate method would report this. As first reported by Bound *et al.* (1995), the TSLS estimate incorrectly suggests that the data are informative, indeed, very informative when there are many instruments. The randomization-based estimators show no such spurious precision: in all cases the 95% confidence intervals include all values between −1 and 1. (The interval $[-1, 1]$ for β is extremely long and entirely uninformative: if $\beta = 1$, then completing 4 years of college would raise earnings by a factor of

Table 3. Comparison of instrumental variable estimates with uninformative data†

Procedure	*95% confidence interval*
Without covariates	
TSLS	[−0.109, 0.648]
Permute log-earnings	Includes [−1, 1]
Permute ranks	Includes [−1, 1]
With state and year covariates	
TSLS	[0.042, 0.078]
Permute log-earnings	Includes [−1, 1]
Permute ranks	Includes [−1, 1]

†Permutation methods reveal that the data contain no information, but TSLS is misleading.

$\exp(4\beta) = 54.6$ times the earnings of a high school graduate, so a college education would raise a minimum wage of perhaps \$10000 per year to more than \$500000, which of course it does not; a doctoral degree taking 8 years beyond high school would yield \$30 million per year; similarly, $\beta = -1$ would reduce \$10000 to \$183 per year as a consequence of 4 years of college.)

Fig. 1 presents the same story graphically, again indicating that permutation methods correctly reflect the absence of information in the random data, whereas TSLS misleadingly suggests that information is present. Fig. 1 presents the $(T-\mu)^2/\sigma^2$ for randomization inferences or $(\hat{\beta}-\beta_0)^2/\widehat{\text{var}}(\hat{\beta})$ for TSLS as a function of the parameter value β_0 for four of the leading cases (random *versus* real data, no covariates *versus* state and year dummy variables). The full horizontal line is at $3.84 = 1.96^2$, and if $(T-\mu)^2/\sigma^2 \leqslant 3.84$ then the corresponding value of β_0 is not rejected at the 0.05-level and is in the 95% confidence interval. In Fig. 1(d), with the random quarter of birth data, the test statistic is very flat as a function of the parameter value for the permutation tests, correctly reflecting the absence of information, whereas it is very curved for TSLS with many instruments, incorrectly suggesting that the random data are informative.

5. A small simulation: weak instruments and long tails

In this section we study the performance of the test statistic in a controlled environment. We consider a standard two-simultaneous-equation model

$$R = r_C + \beta D,$$

$$D = \gamma z + \nu,$$

where (ν, r_C) are independent of the instrument z but, because of the potential correlation between r_C and ν, the potential response under control, r_C, is not necessarily independent of the dose D, as it would be if doses had been randomly assigned. (Incidentally, with the model just identified, the TSLS estimator is identical to the limited information maximum likelihood estimator, so the simulation covers one instance of both types of estimator.) We focus on the power function of the test of the null hypothesis $\beta = \beta_0$, as a function of β_0. The first test that we consider is the usual test based on the normal approximation to the distribution of the TSLS estimator. In addition we consider two versions of the randomization inference: one that permutes the adjusted responses $R - \beta_0 D$ and the other which permutes their ranks. In all figures,

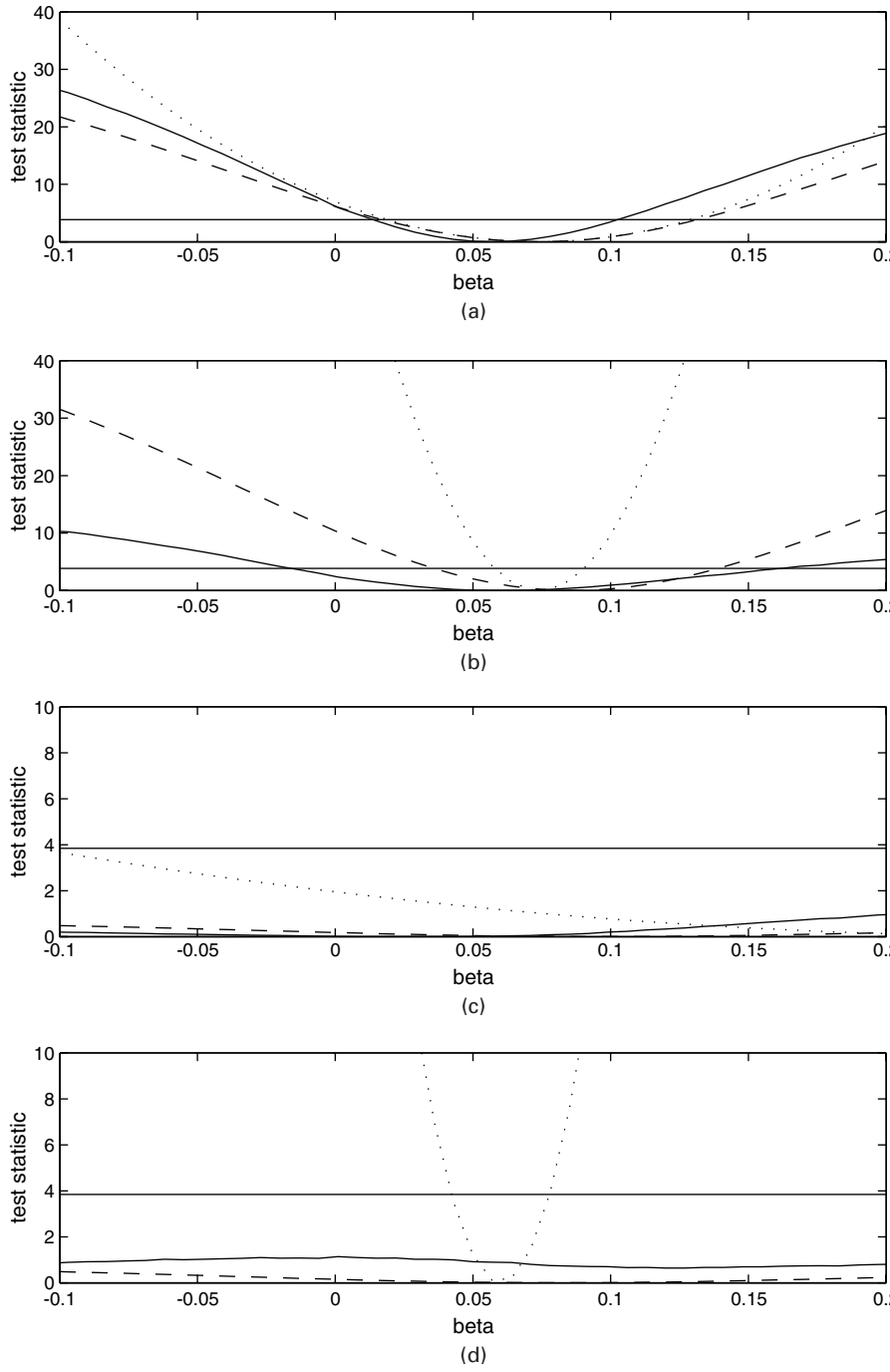

Fig. 1. (a) Quarter-of-birth data without covariates, (b) quarter-of-birth data with state and year-of-birth interactions as covariates, (c) random quarter-of-birth data without covariates and (d) random quarter-of-birth data with state and year-of-birth interactions as covariates: ———, randomization test using ranks; — —, randomization test using the observed data; · · · · · · · , TSLS

TSLS is the *dotted curve*, the randomization test using the observed data is the *broken curve* and the randomization test using the ranks is the *full curve*.

The simulation considers four different situations, listed below. In all four, the instruments are normally distributed with zero mean and unit variance. In each case we draw 40 observations per sample and carry out 100 000 replications of the sampling process. The first three cases differ in the error distributions which can be normal or thick tailed. In addition we consider one data-generating process where the instrument is very weak and the correlation between the errors is high, as in the data-generating processes in Nelson and Startz (1990).

(a) *Strong instrument, thin tails*, $\beta = 1$ *and* $\gamma = 1$: here, (r_C, v) are bivariate normal, specifically, $r_C = \rho\nu + \sqrt{(1-\rho^2)}\omega$ where ν is standard normal and ω is an independent standard normal distribution, and $\rho = 0.5$.

(b) *Strong instrument and thick tails for the response*: the modification from the first case is the distribution of $r_C = \rho\nu + \sqrt{(1-\rho^2)}\omega$ where ν is standard normal and ω has a t-distribution with 2 degrees of freedom, and $\rho = 0.5$.

(c) *Strong instrument and thick tails for the dose*: in the third case the distribution of $\nu = \rho r_C + \sqrt{(1-\rho^2)}\omega$ where r_C is standard normal and again ω has a t-distribution with 2 degrees of freedom, and $\rho = 0.5$.

(d) *Weak instrument and thin tails*: in the fourth case (r_C, v) are bivariate normal with correlation 0.95, so the instrument contributes only slightly to the correlation of dose and responses. The coefficient γ is changed to 0.229 so that R^2 in the first stage is only 0.05.

Fig. 2 presents the power functions for the four data-generating processes for the three tests. Here, we are testing $H_0 : \beta = \beta_0$ for various values of β_0 when in fact $\beta = 1$. The power curve at $\beta_0 = 1$ gives the level of the test, so we hope to see empirically 0.05 for our theoretically 0.05-level test. In Fig. 2(a), with a strong instrument and thin tails, all three tests have approximately the right size (in fact, the size for the TSLS test is slightly high at 0.055, whereas for the other tests the estimated size is 0.049, which cannot be distinguished from 0.05 given that it is based on 100 000 replications), and the standard TSLS-based test is somewhat more powerful under the alternative, as it should be. If the error in the first equation is thick tailed (Fig. 2(b)), then the randomization-based test using the ranks is much more powerful, with more than three times the power of TSLS when testing hypothesis $H_0 : \beta = 2$. With the error in the second equation thick tailed the standard test is again slightly more powerful. The rank-based test is superior to the randomization test based on levels. With a weak instrument (Fig. 2(d)), the standard test has the wrong size, rejecting true hypotheses approximately 15% of the time, as is well known (e.g. Nelson and Startz (1990) and Maddala and Jeong (1992)). The randomization-based tests continue to have the right size in this case.

6. All distribution-free instrumental variable tests are permutation tests

A test of hypothesis $H_0 : \beta = \beta_0$ in Section 4 can be distribution free *only* if it is a permutation test. The logic closely parallels that in Lehmann (1959), section 5.7, for an additive effect, so it will be only sketched here.

Lehmann began by assuming that the r_{Csi} are independent and identically distributed within each stratum s, so their joint density is

$$g(\mathbf{r}_C) = \prod_{s=1}^{S} \prod_{i=1}^{n_s} f_s(r_{Csi}), \tag{4}$$

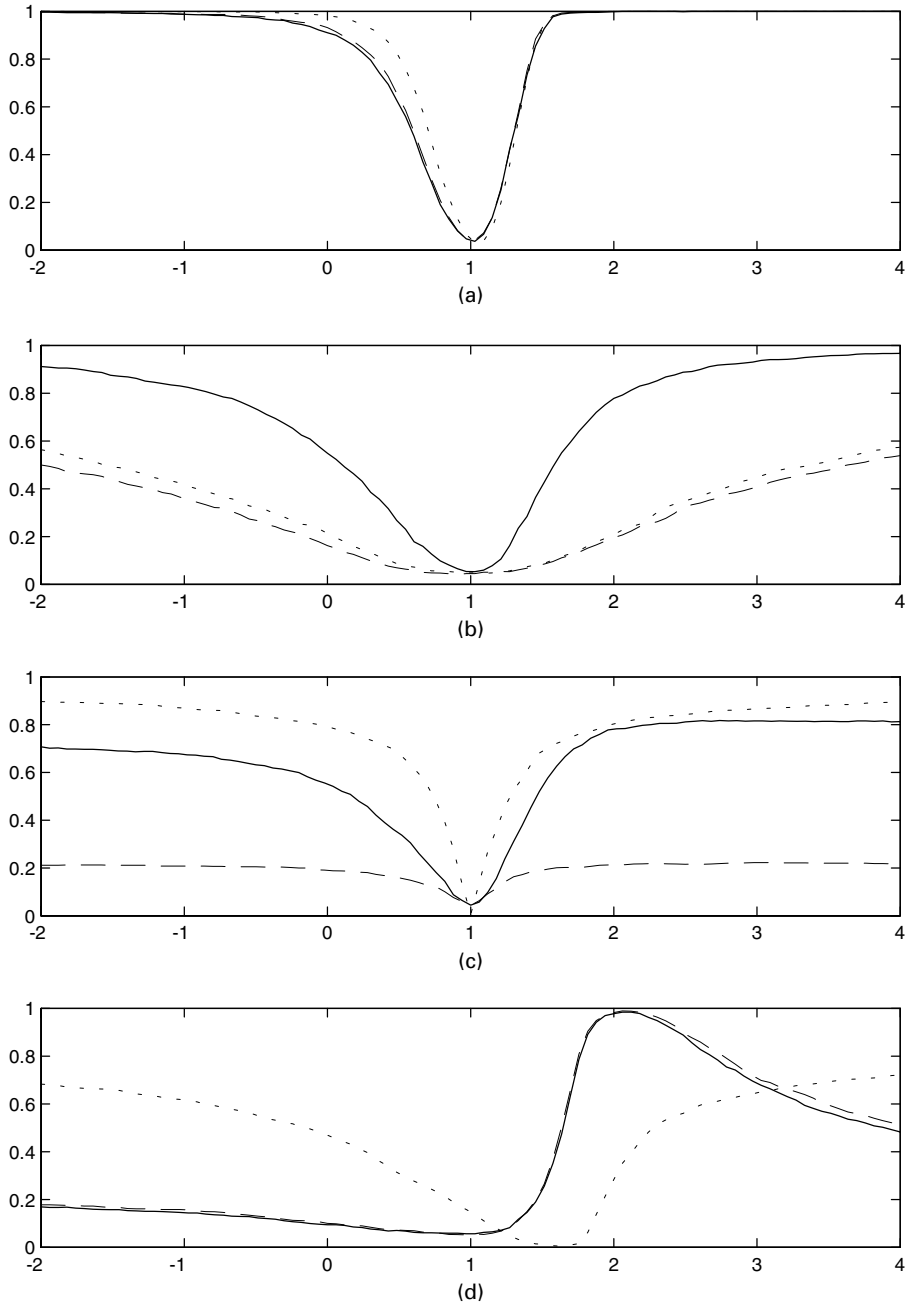

Fig. 2. (a) Power function design (a), (b) power function design (b), (c) power function design (c) and (d) power function design (d): ———, randomization test using ranks; — —, randomization test using the observed data; · · · · · · ·, TSLS

where $f_s(\cdot)$ is unknown for $s = 1, \ldots, S$. The instrumental variable model adds to this the following assumptions:

(a) the binary quarter-of-birth indicators Z_{si} are independent of r_{Csi},
(b) person i in stratum s would receive $d_{si\mathbf{z}}$ years of education beyond the minimum if the quarters of birth were $\mathbf{z}$, so this person actually receives $D_{si} = d_{si\mathbf{Z}}$ years beyond the minimum, and
(c) actual log-earnings are determined by the structural equation $R_{si} = r_{Csi} + \beta D_{si}$.

From this, it follows that the conditional distribution of $R_{si} - \beta_0 D_{si}$ given $\mathbf{Z}$ is equation (4) if $H_0 : \beta = \beta_0$ is true. Now Lehmann showed that a test of the hypothesis that $R_{si} - \beta_0 D_{si}$ given $\mathbf{Z}$ has distribution (4) will have level α for all $f_s(\cdot)$, $s = 1, \ldots, S$, *if and only if* the test rejects the hypothesis for $\alpha|\Omega|$ permutations $\mathbf{p} \in \Omega$ of $\mathbf{R} - \beta_0\mathbf{D}$, i.e. if it divides the orbit $\{\mathbf{p}(\mathbf{R} - \beta_0\mathbf{D}) : \mathbf{p} \in \Omega\}$ into a rejection region Ω_1 containing $\alpha|\Omega|$ permutations and an acceptance region Ω_0 containing $(1 - \alpha)|\Omega|$ permutations. In our simplified description, we have assumed that $\alpha|\Omega|$ is an integer, but Lehmann showed that this is not needed.

7. Conclusion

A common practice with an instrumental variable is to assume that the instrument is informative and that the problem is identified, and to apply asymptotic theory to justify an approximately normal distribution for the estimate $\hat{\beta}$, from methods such as TSLS. This turns out badly when the identification is weak or in doubt, because the resulting methods can perform very poorly, yet the associated confidence intervals wrongly suggest that they have performed well. In contrast, the permutation approach performs well in all the cases that we considered. In favourable situations, with adequate identification and short-tailed responses, the permutation approach is not very different from TSLS. With adequate identification and long-tailed responses, the permutation method yields correct coverage with shorter confidence intervals than does TSLS. With inadequate identification, the permutation method maintains correct coverage, yielding long intervals, correctly reflecting the limited information in the data, whereas TSLS gives misleadingly narrow confidence intervals with coverage rates that are too low.

Rather than assume that the parameter is identified, it is better to let the data speak to the issue of identification. With permutation methods, if identification is weak or non-existent, the confidence interval accurately reflects this by becoming appropriately longer.

Acknowledgements

This work was supported by grants from the US National Science Foundation. The hospitality and support of the Center for Advanced Study in the Behavioral Sciences are gratefully acknowledged.

References

Amemiya. T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.
Angrist, J. D. (1990) Lifetime earnings and the Vietnam era draft lottery: evidence from Social Security administrative records. *Am. Econ. Rev.*, **80**, 313–336.
Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999) Jackknife instrumental variables estimation. *J. Appl. Econometr.*, **14**, 57–67.
Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–455.
Angrist, J. D. and Krueger, A. B. (1991) Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.*, **106**, 979–1014.

Angrist, J. D. and Krueger, A. B. (1995) Split-sample instrumental variables estimates of the returns to schooling. *J. Bus. Econ. Statist.*, **13**, 225–235.
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.
Bekker, P. (1994) Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, **62**, 657–681.
Bound, J., Jaeger, D. A. and Baker, R. M. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Statist. Ass.*, **90**, 443–450.
Card, D. (2001) The causal effect of education. In *Handbook of Labor Economics* (eds O. Ashenfelter and D. Card). New York: North-Holland.
Chamberlain, G. and Imbens, G. (1996) Hierarchical Bayes models with many instruments. *Technical Working Paper 204*. National Bureau of Economic Research, Cambridge.
Cox, D. R. (1958) The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.
Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*. New York: CRC Press.
Durbin, J. (1954) Errors in variables. *Rev. Int. Statist. Inst.*, **22**, 23–32.
Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
Gail, M., Tan, W. and Piantadosi, S. (1988) Tests for no treatment effect in randomized clinical trials. *Biometrika*, **75**, 57–64.
Goetghebeur, E. and Molenberghs, G. (1996) Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *J. Am. Statist. Ass.*, **91**, 928–934.
Hájek, J., Šidák, Z. and Sen, P. K. (1999) *Theory of Rank Tests*, 2nd edn. New York: Academic Press.
Halliwell, J. (1966) Reviewing the reviews on entrance age and school success. *J. Educ. Res.*, **59**, 395–401.
Han, C. and Schmidt, P. (2001) The asymptotic distribution of the instrumental variable estimators when the instruments are not correlated with the regressors. *Econ. Lett.*, **74**, 61–66.
Heitjan, D. F. (1999) Causal inference in a clinical trial: a comparative example. *Contr. Clin. Trials*, **20**, 309–318.
Hodges, J. L. and Lehmann, E. L. (1962) Rank methods for combination of independent experiments in the analysis of variance. *Ann. Math. Statist.*, **33**, 482–497.
Hodges, J. L. and Lehmann, E. L. (1963) Estimates of location based on ranks. *Ann. Math. Statist.*, **34**, 598–611.
Holland, P. W. (1988) Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.*, **18**, 449–484.
Imbens, G. W. and Rubin, D. B. (1997) Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.*, **25**, 305–327.
Kempthorne, O. (1955) Randomization theory of experimental practice. *J. Am. Statist. Ass.*, **50**, 946–967.
Kleibergen, F. (2002) Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, **70**, 1781–1803.
Lehmann, E. (1959) *Testing Statistical Hypotheses*. New York: Wiley.
Lehmann, E. and Stein, C. (1949) On the theory of some nonparametric hypotheses. *Ann. Math. Statist.*, **20**, 28–45.
Maddala, G. S. and Jeong, J. (1992) On the exact small sample distribution of the instrumental variable estimator. *Econometrica*, **60**, 181–183.
Moreira, M. J. (2003) A conditional likelihood ratio test for structural models. *Econometrica*, **71**, 1027–1803.
Nelson, C. R. and Startz, R. (1990) Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, **58**, 967–976.
Pitman, E. J. G. (1937) Significance tests which may be applied to samples from any populations. *J. R. Statist. Soc.*, suppl., **4**, 119–130.
Robinson, J. (1973) The large sample power of permutation tests for randomization models. *Ann. Statist.*, **1**, 291–296.
Rosenbaum, P. R. (1991) Some poset statistics. *Ann. Statist.*, **19**, 1091–1097.
Rosenbaum, P. R. (1996) Comment on "Identification of causal effects using instrumental variables" by Angrist, Imbens and Rubin. *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) Replicating effects and biases. *Am. Statistn*, **55**, 223–227.
Rosenbaum, P. R. (2002a) *Observational Studies*, 2nd edn. New York: Springer.
Rosenbaum, P. R. (2002b) Covariance adjustment in randomized experiments and observational studies (with discussion). *Statist. Sci.*, **17**, 286–327.
Sheiner, L. B. and Rubin, D. B. (1995) Intention-to-treat analysis and the goals of clinical trials. *Clin. Pharm. Therp.*, **57**, 6–15.
Sommer, A. and Zeger, S. L. (1991) On estimating efficacy from clinical trials. *Statist. Med.*, **10**, 45–52.
Staiger, D. and Stock, J. H. (1997) Instrumental variables regression with weak instruments. *Econometrica*, **65**, 557–586.
Tukey, J. W. (1985) Improving crucial randomized experiments—especially in weather modification—by double randomization and rank combination. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (eds L. Le Cam and R. Olshen), vol. 1, pp. 79–108. Belmont: Wadsworth.

Wald, A. (1940) The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, **11**, 284–300.
Welch, B. L. (1937) On the z-test in randomized blocks and Latin squares. *Biometrika*, **29**, 21–52.
Wilk, M. B. (1955) The randomization analysis of a generalized randomized block design. *Biometrika*, **42**, 70–79.
Zelen, M. (1979) A new design for randomized clinical trials. *New Engl. J. Med.*, **300**, 1242–1245.