The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–21

Nonparametric bounds for the causal effect in a binary instrumental variable model

Tom M. Palmer,
MRC CAiTE Centre,
School of Social and Community Medicine,
University of Bristol, UK
tom.palmer@bristol.ac.uk

Roland R. Ramsahai,
Statistics Laboratory,
University of Cambridge, UK

Vanessa Didelez,
School of Mathematics,
University of Bristol, UK

Nuala A. Sheehan,
Departments of Health Sciences and Genetics,
University of Leicester, UK

Abstract. Instrumental variables can be used to make inferences about causal effects in the presence of unmeasured confounding. For a model in which the instrument, intermediate/treatment, and outcome variables are all binary, Balke and Pearl (Journal of the American Statistical Association, 1997, 92: 1172–1176) derived nonparametric bounds for the intervention probabilities and the average causal effect. We have implemented these bounds in two commands, `bpbounds` and `bpboundsi`. We have also implemented several extensions to these bounds. One of these is for the situation where the instrument and outcome are measured in one sample, and the instrument and intermediate are measured in another sample. We have also implemented the bounds for an instrument with three categories, as is common in Mendelian randomization analyses in epidemiology and for the case where a monotonic effect of the instrument on the intermediate can be assumed. In each case, we calculate the IV inequality constraints as a check for gross violations of the IV assumptions. The use of the commands is illustrated with a recreation of the original Balke and Pearl analysis and with a Mendelian randomization analysis. We also give a simulated example to demonstrate that the IV inequality constraints can both detect and fail to detect violations of the IV assumptions.

Keywords: st0001, bpbounds, bpboundsi, average causal effect, causal inference, instrumental variables, nonparametric bounds.

1 Introduction

Instrumental variables (IVs) can be used for inference on causal effects in the presence of unobserved confounding. One of their uses is for deriving upper and lower bounds for a causal effect in situations where we are interested in the effect of a binary exposure or treatment (endogenous variable) on a binary outcome and when we do not want to rely on any further assumptions apart from those defining an IV. These nonparametric bounds were derived independently by Robins (1989) and Manski (1990), and subsequently improved by Balke and Pearl (1997). A detailed overview is given in Pearl (2009, Chapter 8). They have also been generalised to cope with different data structures by Ramsahai (2007, 2011).

© *yyyy* StataCorp LP

Typical applications of this methodology include randomized controlled trials with partial compliance, where random assignment is the instrument and actual treatment taken is the intermediate variable (Balke and Pearl 1997). Another application is provided by Mendelian randomization studies in epidemiology where the instrument is a genetic predisposition (genotype) for the exposure of interest (Davey Smith and Ebrahim 2003; Lawlor et al. 2008).

We have implemented these bounds in a program `bpbounds` and an immediate version `bpboundsi`. We explain the instrumental variable assumptions and how they allow bounds to be obtained for a causal effect. This is followed by a description of the commands, and demonstration of their use on some examples.

2 The average causal effect

We define X to be the exposure variable and Y the outcome variable and assume that both are binary with the following interpretations: $X = 0$ was not exposed, $X = 1$ was exposed, $Y = 0$ did not experience the outcome, $Y = 1$ experienced the outcome. The average causal effect (ACE) of X on Y is the mean difference in Y if we set $X = 1$ as opposed to $X = 0$ by an intervention. This can be formally expressed using Pearl's $do(\cdot)$ notation (Pearl 2009): $P(Y|do(X = x))$ denotes the distribution of Y when we actively manipulate X fixing it at value x, while the usual $P(Y|X = x)$ denotes the distribution of Y when we passively observe that $X = x$. When there is confounding the latter will typically depend on X in a different way than the former. The ACE is then expressed as follows

$$ACE = E(Y|do(X = 1)) - E(Y|do(X = 0)).$$

Using potential outcome notation (Rubin 1974, 1978) this is expressed as $ACE = E(Y(1)) - E(Y(0))$, where $Y(x)$ denotes the potential outcome of Y when we set $X = x$ by an intervention. In other words, the ACE is the causal risk difference (Greenland 2000). In a randomised controlled trial (RCT), where X is randomly allocated, the ACE is the typical target of inference. More generally, we might be interested in other causal parameters which could be any functions of the intervention probabilities $P(Y = 1|do(X = x))$, e.g. the causal risk ratio $P(Y = 1|do(X = 1))/P(Y = 1|do(X = 0))$.

3 Instrumental variables

In observational studies or RCTs with imperfect compliance, it can often not be ruled out that unobserved confounding affects the association of X and Y. A causal effect is then usually not identifiable from data on (X, Y) alone. However, in the presence of an IV, Z, data can be at least partially informative for the causal effect in the sense that it imposes upper and lower bounds on $P(Y = y|do(X = x))$ and by extension on the ACE.

3.1 Definition of IVs

Assuming that the unobserved confounding can be represented by a variable or vector U, a valid IV Z satisfies the following core conditions (where $A \perp\!\!\!\perp B \mid C$ means that variables A and B are conditionally independent given C (Dawid 1979)):

(i) $Z \perp\!\!\!\perp U$;

(ii) $Z \not\perp\!\!\!\perp X$;

(iii) $Y \perp\!\!\!\perp Z \mid (X, U)$.

When data on (X, Y, Z) is available and all three variables are discrete, lower and upper bounds on the ACE can always be calculated provided the core IV assumptions are satisfied. This is because the IV conditions (i) $Z \perp\!\!\!\perp U$ and (iii) $Y \perp\!\!\!\perp Z \mid (X, U)$ impose certain constraints on the distribution of (X, Y, Z) addressed in the next section. However, point estimation of the ACE requires additional parametric assumptions (Didelez and Sheehan 2007).

3.2 Inequality constraints

The conditional independencies (i) and (iii) imply that
$P(Y, X, U|Z) = P(Y|X, U)P(X|Z, U)P(U)$; this in turn implies that the observable marginal $P(Y, X|Z)$ is not unrestricted as it has to be obtainable from
$P(Y|X, U)P(X|Z, U)P(U)$ by integrating out U. When X, Y, and Z are discrete (while U is entirely unrestricted) this leads to non-trivial constraints on $P(Y, X|Z)$ that can be expressed as a set of inequality constraints. 'Non-trivial' here means that there exist conditional distributions that do not satisfy the inequality constraints and hence cannot satisfy (i) and (iii). It is therefore necessary to check that these inequality constraints are supported by the observed data on X, Y, and Z. If we find that at least one inequality is violated we can conclude that Z is not a valid IV. The general form of these inequality constraints is (Pearl 1995a,b)

$$\max_x \sum_y [\max_z P(Y = y, X = x|Z = z)] \leq 1. \tag{1}$$

Note that for condition (ii) we simply need to check that $P(X = x|Z = z_1) \neq P(X = x|Z = z_2)$ for $z_1 \neq z_2$, i.e. X and Z are associated, which can easily be checked on the observed data.

In the particular case where all three variables are binary, we denote the conditional probability (as in Balke and Pearl (1997)) $p_{yx.z} = P(Y = y, X = x|Z = z)$. Then the constraints can be written out in detail as

$$\begin{aligned} p_{00.0} + p_{10.1} &\leq 1 \\ p_{10.0} + p_{00.1} &\leq 1 \\ p_{11.0} + p_{01.1} &\leq 1 \\ p_{01.0} + p_{11.1} &\leq 1, \end{aligned} \tag{2}$$

in addition to the usual $0 \leq p_{yx.z} \leq 1$ and $\sum_{y,x} p_{yx.z} = 1$. The above can be checked from data by substituting the corresponding relative frequency for $p_{yx.z}$. Note that this ‘checking’ of the IV conditions is not comparable to a statistical test because we only know that *if* the above inequalities fail, then the core conditions must be violated; however, it is possible that the core IV conditions are violated *without* failing the inequalities. It is therefore advisable to justify conditions (i) – (iii) based on subject matter background knowledge. Furthermore, simply plugging–in the relative frequencies to check the inequalities does not take sampling variation into account; however, we will ignore this here but Ramsahai and Lauritzen (2011) discuss the corresponding statistical test. Bonet (2001) shows that in the case where X is continuous there are no constraints comparable to (1) on the observable distribution $P(Y, X|Z)$ but some constraints can be found when Y and Z are continuous and X is discrete.

4 Bounds on causal effects

We first address bounds that are valid assuming only (i) – (iii). If an additional ‘monotonicity’ assumption is made, these can sometimes be tightened, see Section 4.2.

4.1 General bounds

For the case of binary variables (X, Y) and binary IV Z, Balke and Pearl (1997) derive bounds for the intervention probabilities $\pi_x = P(Y = 1|do(X = x))$ given as follows.

$$\max \left\{ \begin{matrix} p_{10.1} \\ p_{10.0} \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{matrix} \right\} \leq \pi_0 \leq \min \left\{ \begin{matrix} 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{matrix} \right\} \tag{3}$$

and

$$\max \left\{ \begin{matrix} p_{11.0} \\ p_{11.1} \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1} \\ -p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \end{matrix} \right\} \leq \pi_1 \leq \min \left\{ \begin{matrix} 1 - p_{01.1} \\ 1 - p_{01.0} \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} \end{matrix} \right\} \tag{4}$$

As ACE$= \pi_1 - \pi_0$, we can combine (3) and (4) to obtain bounds on the ACE; the lower bound is given by

$$ACE \geq \max \left\{ \begin{matrix} p_{00.0} + p_{11.1} - 1 \\ p_{00.1} + p_{11.1} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{00.0} + p_{11.0} - 1 \\ 2p_{00.0} + p_{11.0} + p_{10.0} + p_{11.1} - 2 \\ p_{00.0} + 2p_{11.0} + p_{00.1} + p_{01.1} - 2 \\ p_{10.0} + p_{11.0} + 2p_{00.1} + p_{11.1} - 2 \\ p_{00.0} + p_{01.0} + p_{00.1} + 2p_{11.1} - 2 \end{matrix} \right\}; \tag{5}$$

the upper bound is given by

$$ACE \leq \min \left\{ \begin{array}{c} 1 - p_{10.0} - p_{01.1} \\ 1 - p_{01.0} - p_{10.1} \\ 1 - p_{01.0} - p_{10.0} \\ 1 - p_{01.1} - p_{10.1} \\ 2 - 2p_{01.1} - p_{10.0} - p_{10.1} - p_{11.1} \\ 2 - p_{01.0} - 2p_{10.0} - p_{00.1} - p_{01.1} \\ 2 - p_{10.0} - p_{11.0} - 2p_{01.1} - p_{10.1} \\ 2 - p_{00.0} - p_{01.0} - p_{01.1} - 2p_{10.1} \end{array} \right\}. \qquad (6)$$

Robins (1989) and Manski (1990) derived the first four lines of (5) and (6), Balke and Pearl (1997) tightened these by deriving the rest. Note that any combination of π_0 and π_1 in (3) and (4) is possible (Dawid 2003) and hence we can also obtain bounds for the causal risk ratio ($CRR = \pi_1/\pi_0$) as follows

$$\frac{\pi_1^L}{\pi_0^U} \leq CRR \leq \frac{\pi_1^U}{\pi_0^L},$$

where π_x^L, π_x^U are the lower and upper bounds of π_x from (3) and (4).

4.2 The monotonicity assumption

In some applications it seems sensible to believe that for all values u of U

$$P(X = 1|Z = 1, U = u) \geq P(X = 1|Z = 0, U = u), \qquad (7)$$

which is a weaker version of the *monotonicity* assumption of Imbens and Angrist (1994) and Angrist et al. (1996). Note that we assume here that the levels of X are coded such that higher values are more likely given higher values of Z.

The constraints imposed by the IV conditions (i) and (iii) together with (7) lead to a tightening of the inequalities from Section 3.2 to (Balke and Pearl 1997)

$$\begin{aligned} p_{01.1} - p_{01.0} &\geq 0 \\ p_{11.1} - p_{11.0} &\geq 0 \\ p_{00.0} - p_{00.1} &\geq 0 \\ p_{10.0} - p_{10.1} &\geq 0. \end{aligned}$$

Furthermore, assuming (7) reduces the bounds on the ACE to

$$p_{00.0} - p_{00.1} - p_{01.1} - p_{10.1} \leq ACE \leq p_{00.0} + p_{01.0} + p_{11.0} - p_{01.1}.$$

These correspond to the bounds derived by Robins (1989) and Manski (1990).

In some applications it is impossible to observe $X = 1$ when $Z = 0$, for instance when subjects assigned to the control group ($Z = 0$) cannot possibly get hold of treatment ($X = 1$) and hence necessarily have to comply with their assignment, i.e. $P(X = 1|Z = 0) = 0$. This implies that monotonicity (7) necessarily holds. In such a case the general bounds for the ACE and the ones obtained under monotonicity are the same.

5 Other data structures

The bounds as stated above require joint prospective data on binary variables (X, Y, Z). However, modified bounds can be computed for different data structures, and the following structures can be used with the `bpbounds` and `bpboundsi` commands.

5.1 Instrument with three levels

The technique used to find the bounds can in principle be extended to discrete variables (X, Y, Z) with any finite number of levels, but the corresponding formulae quickly become prohibitive. The `bpbounds` and `bpboundsi` commands described below will also calculate the bounds when the IV Z has three levels. Dawid (2003), Ramsahai (2007), and Ramsahai (2011) describe the general technique how these can be obtained. An instrument with three levels is for instance relevant in Mendelian randomisation applications (Lawlor et al. 2008), where Z is a genotype coded as a risk allele count $\{0, 1, 2\}$.

5.2 Bivariate/marginal data

The above assumes that we have *jointly* observed all three variables (X, Y, Z). In some cases, however, data might have been obtained from separate studies, a first study where the pair (X, Z) was observed and a second study where (Y, Z) was observed. We call the case of joint data 'trivariate' and the case of separate (X, Z) and (Y, Z) data 'bivariate'. Such bivariate data provides less information and hence leads to different formulae for the bounds on π_x and hence on the ACE. Ramsahai (2007) derives the restrictions corresponding to the 'check' of Section 3.2 for bivariate data (see equation (5) of that paper), as well as the formulae for the bounds on the ACE corresponding to (5) and (6). Their calculation with the `bpboundsi` command is illustrated below.

5.3 Case-control data

The probabilities $p_{yx.z}$ required for the above bounds cannot be estimated from case–control data without additional information. Instead we can estimate $p^{cc}_{xz.y} = P(X = x, Z = z|Y = y)$ as the relative frequencies of (x, z) within cases $y = 1$ and within controls $y = 0$. If additional information on the marginal probability $P(Y = 1)$ is given, we can recover the required $p_{yx.z}$ as (Didelez and Sheehan 2007)

$$p_{yx.z} = \frac{p^{cc}_{xz.y} P(Y = y)}{\sum_{x,y} p^{cc}_{xz.y} P(Y = y)}.$$

Such additional information, for example, on the disease prevalence in the general population, may be available from other sources or databases. If it is not available, the researcher may still have a good idea of plausible values such as $P(Y = 1) \in [a, b]$ and one may then compute two sets of bounds, one for $P(Y = 1) = a$ and one for $P(Y = 1) = b$ in order to assess the sensitivity of the bounds to the assumed disease prevalence.

6 Interpretation of bounds

It is important to note that the bounds on the ACE (or on π_x) are *not* confidence intervals. If we find for example a lower and upper bound of [0.1, 0.3] this means that there exists some distribution involving the unobserved U that yields a true ACE as small as 0.1, while another choice of distribution involving U has a true ACE as large as 0.3, with both distributions satisfying the IV conditions and having the same observed marginal frequencies on (X, Y, Z) (or, in case of bivariate data on (X, Z) and (Y, Z)). As U is unobserved, it is impossible to decide where the ACE lies in the interval [0.1, 0.3] from the observable data without making further assumptions.

The bounds (5) and (6) are the tightest possible bounds if we make no other assumptions than the IV conditions (i)–(iii); they have therefore also been called the best *assumption-free* (or nonparametric) bounds for the ACE (Balke and Pearl 1994).

We have noted that the additional assumption of monotonicity (7) typically leads to tighter bounds. Another popular assumption is that $E(Y|X = x, U = u) = \beta x + h(u)$ for some function $h(\cdot)$, i.e. additivity of the outcome model (Didelez et al. 2010). In this case it can be shown that $\beta = \text{ACE}$ where,

$$\beta = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(X|Z = 1) - E(X|Z = 0)} = \frac{\text{cov}(Y, Z)}{\text{cov}(X, Z)} \tag{8}$$

which can be estimated using the ratio estimator or two-stage least squares (Angrist and Imbens 1995). Two-stage least squares is implemented in the official Stata command `ivregress` and also in the user-written command `ivreg2` (Baum et al. 2003, 2007, 2010). As the point estimate (8) relies on specific parametric assumptions it is always advisable to compare it with the assumption-free bounds to assess sensitivity to these additional assumptions.

7 The bpbounds command

The `bpbounds` command, and the immediate version `bpboundsi`, initially perform the inequality check of Section 3.2 and, if valid, proceed to calculate the bounds on the ACE as well as on the intervention probabilities and the CRR. The commands then also check the constraints under the additional assumption of monotonicity (7) and, if valid, compute the same set of bounds assuming monotonicity. The `bpbounds` command can only be applied to trivariate data (we assume that a Stata dataset comes from a single sample), whereas `bpboundsi` accepts frequencies or conditional probabilities from both trivariate and bivariate data as in Section 5.2. Both commands allow an instrument with either two or three categories.

The commands use the polytope transformation method devised by Bonet (2001) and Dawid (2003) and described in detail by Ramsahai (2007) and Ramsahai (2011). The relevant polytope transformations were calculated using polymake (Gawrilow and Joswig 2000) and PORTA (version 1.4.1).

7.1 Syntax

Syntax for `bpbounds` (trivariate data only):

`bpbounds` *depvar* `(`*varname*$_{\text{endog}}$ `=` *varname*$_{\text{iv}}$`)` [*if*] [*in*] [*weight*] [`, fmt(`*string*`)`]

This follows the standard syntax for Stata instrumental variable estimation commands such as `ivregress` where *depvar* is the outcome variable (Y), *varname*$_{\text{endog}}$ is the exposure or treatment received or endogenous variable (X), and *varname*$_{\text{iv}}$ is the instrumental variable (Z). There are restrictions on how these variables are coded: the categories of *depvar* and *varname*$_{\text{endog}}$ must be coded {0, 1}, and the categories of *varname*$_{\text{iv}}$ must be coded {0, 1} for a two category instrument and {0, 1, 2} for a three category instrument. Note unlike other Stata instrumental variable estimation commands exogenous covariates are not allowed. Frequency weights are allowed.

The `bpboundsi` command is an immediate command. It accepts inputs as either frequency counts or conditional probabilities entered directly or in matrices. Syntax for `bpboundsi` with an instrument with two categories:

`bpboundsi` [$\#_1$ $\#_2$ $\#_3$ $\#_4$ $\#_5$ $\#_6$ $\#_7$ $\#_8$] [`, fmt(`*string*`)` `bivariate` `matrices(`*matlist*`)`]

The inputs ($\#_1$–$\#_8$) are as described in Table 1 or can be in matrices using the matrices option.

Syntax for `bpboundsi` with an instrument with three categories:

`bpboundsi` [$\#_1$ $\#_2$ $\#_3$ $\#_4$ $\#_5$ $\#_6$ $\#_7$ $\#_8$ $\#_9$ $\#_{10}$ $\#_{11}$ $\#_{12}$] [`, fmt(`*string*`)` `bivariate` `matrices(`*matlist*`)`]

The inputs ($\#_1$–$\#_{12}$) are as described in Tables 1 and 2 or can be in matrices using the matrices option.

7.2 Options

`bivariate` specifies bivariate data. The default is trivariate data.

`fmt(`*string*`)` specifies the format of the results. The default is `fmt(%5.4f)`. See `help format` or [U] **12.5 Formats: Controlling how data are displayed**.

`matrices(`*matlist*`)` specifies frequencies/conditional probabilities input in matrices. Trivariate data: the X categories must be the rows and the Y categories the columns. The matrices must also be listed by ordered categories of Z, i.e. conditional on $Z = 0$, $Z = 1$, $Z = 2$. Bivariate data: matrices must be listed in the following order; (Z by Y) then (Z by X).

The commands return their results in scalars and matrices as detailed in the help-file.

Two category instrument			Three category instrument		
Input	Freq.	Cond. prob.	Input	Freq.	Cond. prob.
$\#_1$	$ng_{0.0}$	$\gamma_{0.0}$	$\#_1$	$ng_{0.0}$	$\gamma_{0.0}$
$\#_2$	$ng_{1.0}$	$\gamma_{1.0}$	$\#_2$	$ng_{1.0}$	$\gamma_{1.0}$
$\#_3$	$ng_{0.1}$	$\gamma_{0.1}$	$\#_3$	$ng_{0.1}$	$\gamma_{0.1}$
$\#_4$	$ng_{1.1}$	$\gamma_{1.1}$	$\#_4$	$ng_{1.1}$	$\gamma_{1.1}$
$\#_5$	$nt_{0.0}$	$\theta_{0.0}$	$\#_5$	$ng_{0.2}$	$\gamma_{0.2}$
$\#_6$	$nt_{1.0}$	$\theta_{1.0}$	$\#_6$	$ng_{1.2}$	$\gamma_{1.2}$
$\#_7$	$nt_{0.1}$	$\theta_{0.1}$	$\#_7$	$nt_{0.0}$	$\theta_{0.0}$
$\#_8$	$nt_{1.1}$	$\theta_{1.1}$	$\#_8$	$nt_{1.0}$	$\theta_{1.0}$
			$\#_9$	$nt_{0.1}$	$\theta_{0.1}$
			$\#_{10}$	$nt_{1.1}$	$\theta_{1.1}$
			$\#_{11}$	$nt_{0.2}$	$\theta_{0.2}$
			$\#_{12}$	$nt_{1.2}$	$\theta_{1.2}$

Table 1: `bpboundsi` inputs for bivariate data; $ng_{y.z} = \#(Y = y|Z = z)$, $\gamma_{y.z} = P(Y = y|Z = z)$, $nt_{x.z} = \#(X = x|Z = z)$, $\theta_{x.z} = P(X = x|Z = z)$.

Input	Freq. $n_{yx.z}$	Cond. prob. $p_{yx.z}$
$\#_1$	$n_{00.0}$	$p_{00.0}$
$\#_2$	$n_{10.0}$	$p_{10.0}$
$\#_3$	$n_{01.0}$	$p_{01.0}$
$\#_4$	$n_{11.0}$	$p_{11.0}$
$\#_5$	$n_{00.1}$	$p_{00.1}$
$\#_6$	$n_{10.1}$	$p_{10.1}$
$\#_7$	$n_{01.1}$	$p_{01.1}$
$\#_8$	$n_{11.1}$	$p_{11.1}$
$\#_9$	$n_{00.2}$	$p_{00.2}$
$\#_{10}$	$n_{10.2}$	$p_{10.2}$
$\#_{11}$	$n_{01.2}$	$p_{01.2}$
$\#_{12}$	$n_{11.2}$	$p_{11.2}$

Table 2: `bpboundsi` inputs for trivariate data; $n_{yx.z} = \#(Y = y, X = x|Z = z)$, $p_{yx.z} = P(Y = y, X = x|Z = z)$. For a two category instrument $\#_1$–$\#_8$ are required. For a three category instrument $\#_1$–$\#_{12}$ are required.

8 Use of `bpbounds` and `bpboundsi`

8.1 Balke-Pearl Vitamin A supplementation example

Balke and Pearl (1997) illustrate their methodology with data described by Sommer et al. (1986), assessing the impact of vitamin A supplementation on childhood mortality. In the trial, 450 villages in northern Sumatra were randomized to either receive vitamin A supplementation or act as a control group for a year. This randomized assignment provides the IV, $Z = 1$ being the treatment group and $Z = 0$ the control. Children in the treatment group received two large doses of vitamin A, whilst controls received no treatment. Not every child in the treatment group complied with the assignment, so that $X = 1$ denotes treatment actually taken, and $X = 0$ means no treatment taken. The control group necessarily had to comply as vitamin A supplements were not available to them. As noted in Section 4.2, this automatically implies that the monotonicity assumption is satisfied as $P(X = 1|Z = 0) = 0$. The outcome Y was the number of deaths in both groups (where $Y = 1$ denotes survival). Table 4 shows the results of the trial. We can see from the two zero cell counts that children who were randomized to the control group had to comply.

	$Z = 0$		$Z = 1$	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$
$X = 0$	74	11514	34	2385
$X = 1$	0	0	12	9663

Table 3: Vitamin A supplementation data from Balke and Pearl (1997, Table 1).

We enter the data into Stata and run the `bpbounds` command.

```
. clear

. input z x y count

             z          x          y      count
  1. 0 0 0 74
  2. 0 0 1 11514
  3. 1 0 0 34
  4. 1 0 1 2385
  5. 1 1 0 12
  6. 1 1 1 9665
  7. end

. bpbounds y (x = z) [fw=count]

Data:                              Trivariate
Instrument categories:             2

──────────────────────────────┬───────────────────────────────────────────────
                              │                               Bounds
            Causal parameter  │                           Lower      Upper
──────────────────────────────┼───────────────────────────────────────────────
   IV inequality constraints  │    satisfied
                         ACE  │                           -0.1946    0.0054
               P(Y|do(X=0))   │                           0.9936     0.9936
               P(Y|do(X=1))   │                           0.7990     0.9990
                         CRR  │                           0.8042     1.0054
──────────────────────────────┼───────────────────────────────────────────────
   Assuming monotonicity:     │
    Monotonicity constraints  │    satisfied
                         ACE  │                           -0.1946    0.0054
               P(Y|do(X=0))   │                           0.9936     0.9936
               P(Y|do(X=1))   │                           0.7990     0.9990
                         CRR  │                           0.8042     1.0054
──────────────────────────────┴───────────────────────────────────────────────
```

The command lists that we have trivariate data and an instrument with two categories. The IV inequality, the 'check' of the IV assumptions, is satisfied. Then the command gives the bounds for the ACE, which are $-0.1946 \leq ACE \leq 0.0054$ as reported by Balke and Pearl (1997). We multiply the results by 100 to express in percentages, hence the ACE lies between -19.5% and 0.5%. The command then reports the bounds for: the intervention probabilities, $P(Y=1|do(X=0))$ and $P(Y=1|do(X=1))$, and the CRR. In this data situation, the upper and lower bounds for $P(Y=1|do(X=0))$ are equal because there was no non-compliance in the control group.

Next the command checks the monotonicity inequality. As mentioned above, this is necessarily satisfied, and the command reports the bounds for the ACE, intervention probabilities, and CRR under monotonicity. Again, we note that in this particular example all the bounds under monotonicity are the same as those without assuming monotonicity because there was no non-compliance in the control group.

We could also use the immediate command `bpboundsi` to calculate the bounds, being careful to enter the eight numbers in the appropriate order. First we calculate the required frequencies using the `tabulate` command. The `bpboundsi` command alternatively accepts conditional probabilities, as reported by Table 2 of Balke and Pearl (1997), which we also demonstrate below.

```
. bysort z: tabulate x y [fw=count], cell
-------------------------------------------------------------------------------
-> z = 0

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  cell percentage  |
+-------------------+

           |           y
         x |         0          1 |     Total
-----------+----------------------+----------
         0 |        74     11,514 |    11,588 
           |      0.64      99.36 |    100.00 
-----------+----------------------+----------
     Total |        74     11,514 |    11,588 
           |      0.64      99.36 |    100.00 
-------------------------------------------------------------------------------
-> z = 1

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  cell percentage  |
+-------------------+

           |           y
         x |         0          1 |     Total
-----------+----------------------+----------
         0 |        34      2,385 |     2,419 
           |      0.28      19.72 |     20.00 
-----------+----------------------+----------
         1 |        12      9,665 |     9,677 
           |      0.10      79.90 |     80.00 
-----------+----------------------+----------
           |                      |
     Total |        46     12,050 |    12,096 
           |      0.38      99.62 |    100.00 

. * input frequencies
. bpboundsi 74 11514 0 0 34 2385 12 9665
Data:                             Trivariate
Instrument categories:            2
------------------------------------------------------------------------------
                             |                            Bounds
            Causal parameter |                        Lower      Upper
-----------------------------+------------------------------------------------
   IV inequality constraints |    satisfied
                         ACE |                        -0.1946    0.0054
               P(Y|do(X=0))  |                        0.9936     0.9936
               P(Y|do(X=1))  |                        0.7990     0.9990
                         CRR |                        0.8042     1.0054
-----------------------------+------------------------------------------------
   Assuming monotonicity:    |
    Monotonicity constraints |    satisfied
                         ACE |                        -0.1946    0.0054
               P(Y|do(X=0))  |                        0.9936     0.9936
               P(Y|do(X=1))  |                        0.7990     0.9990
                         CRR |                        0.8042     1.0054
------------------------------------------------------------------------------

. * input conditional probabilities
. bpboundsi .0064 .9936 0 0 .0028 .1972 .001 .799
Data:                             Trivariate
Instrument categories:            2
------------------------------------------------------------------------------
                             |                            Bounds
            Causal parameter |                        Lower      Upper
-----------------------------+------------------------------------------------
   IV inequality constraints |    satisfied
                         ACE |                        -0.1946    0.0054
               P(Y|do(X=0))  |                        0.9936     0.9936
               P(Y|do(X=1))  |                        0.7990     0.9990
                         CRR |                        0.8041     1.0054
-----------------------------+------------------------------------------------
   Assuming monotonicity:    |
    Monotonicity constraints |    satisfied
                         ACE |                        -0.1946    0.0054
               P(Y|do(X=0))  |                        0.9936     0.9936
               P(Y|do(X=1))  |                        0.7990     0.9990
                         CRR |                        0.8041     1.0054
------------------------------------------------------------------------------
```

We obtain the same results as before. We now estimate the ACE as in (8).

```
. qui corr y z [fw=count], cov
. sca covyz = r(cov_12)
. qui corr x z [fw=count], cov
. sca covxz = r(cov_12)
. di "ACE:", %5.4f covyz/covxz
ACE: 0.0032
```

This means that the additional assumption of linearity and additivity $E(Y|X = x, U = u) = \beta x + h(u)$ allows us to estimate an ACE of 0.3% which is close to the upper bound calculated earlier. The same estimate can be obtained from the `ivregress` or `ivreg2` commands but the standard errors are not generally appropriate for binary outcomes.

To demonstrate the use of the bivariate option we next assume that (X, Z) were collected in one sample and (Y, Z) in another. The following code also demonstrates passing frequencies to `bpboundsi` in matrices, which we generate using `tabulate`.

```
. tab z y [fw=count], row matcell(zy)

  ┌────────────────┐
  │ Key            │
  ├────────────────┤
  │   frequency    │
  │ row percentage │
  └────────────────┘

           │          y
         z │         0          1 │     Total
───────────┼──────────────────────┼──────────
         0 │        74     11,514 │    11,588 
           │      0.64      99.36 │    100.00 
───────────┼──────────────────────┼──────────
         1 │        46     12,050 │    12,096 
           │      0.38      99.62 │    100.00 
───────────┼──────────────────────┼──────────
     Total │       120     23,564 │    23,684 
           │      0.51      99.49 │    100.00 

. tab z x [fw=count], row matcell(zx)

  ┌────────────────┐
  │ Key            │
  ├────────────────┤
  │   frequency    │
  │ row percentage │
  └────────────────┘

           │          x
         z │         0          1 │     Total
───────────┼──────────────────────┼──────────
         0 │    11,588          0 │    11,588 
           │    100.00       0.00 │    100.00 
───────────┼──────────────────────┼──────────
         1 │     2,419      9,677 │    12,096 
           │     20.00      80.00 │    100.00 
───────────┼──────────────────────┼──────────
     Total │    14,007      9,677 │    23,684 
           │     59.14      40.86 │    100.00 

. bpboundsi, mat(zy zx) biv
Data:                             Bivariate
Instrument categories:            2
──────────────────────────────┬─────────────────────────────────────────────────
                              │                              Bounds
              Causal parameter│                          Lower     Upper
──────────────────────────────┼─────────────────────────────────────────────────
     IV inequality constraints│   satisfied
                           ACE│                          -0.1974   0.0064
```

```
               P(Y|do(X=0)) |                          0.9936    0.9936
               P(Y|do(X=1)) |                          0.7962    1.1962
                        CRR |                          0.8013    1.2039
----------------------------+----------------------------------------------
   Assuming monotonicity:   |
    Monotonicity constraints|     satisfied
                        ACE |                         -0.1974    0.0064
               P(Y|do(X=0)) |                          0.9936    0.9936
               P(Y|do(X=1)) |                          0.7962    1.0026
                        CRR |                          0.8013    1.0090
----------------------------+----------------------------------------------
```

In the case of bivariate data the bounds for the ACE are now $-0.1974 \leq ACE \leq 0.0064$. As expected these are slightly wider, because bivariate data are less informative than trivariate data.

8.2 Mendelian randomization example with a three category instrument

In epidemiology the ‘Mendelian randomization’ approach represents the use of genotypes as instrumental variables (Davey Smith and Ebrahim 2003). Importantly the chosen genotypes in such a study should have been shown to be robustly associated with the exposure in previous replicated genome-wide association studies (GWAS). Such genotypes are promising candidates for instrumental variables because the randomization of alleles at conception means genotypes are very unlikely to be associated with potential confounding factors which can bias traditional observational studies (Davey Smith et al. 2007). For a more detailed discussion of the Mendelian randomization approach see Didelez and Sheehan (2007); Lawlor et al. (2008); Palmer et al. (2011). For a biallelic polymorphism there are three genotypes, hence we have implemented the extension of the bounds for a three category instrument in the `bpbounds` and `bpboundsi` commands.

We perform a Mendelian randomization analysis using the 677CT polymorphism (rs1801133) in the Methylenetetrahydrofolate Reductase (*MTHFR*) gene, involved in folate metabolism, as an instrumental variable (Z) to investigate the effect of homocysteine (X) on cardiovascular disease (CVD, Y) risk using data published by Meleady et al. (2003, Table 3). This polymorphism has subsequently been found to be robustly associated with homocysteine in GWAS (Tanaka et al. 2009) although it was identified prior to this. The ‘T’ allele is associated with higher average homocysteine levels.

In our analysis we combine the six homocysteine categories into two categories (low: $< 15\mu$mol/L; high: $\geq 15\mu$mol/L). The analysis is further complicated because it is a case-control study ($Y = 0$ denotes controls and $Y = 1$ denotes CVD cases). The original case-control data are shown in Table 4.

As we commented in Section 5.3, to calculate the bounds we must first convert the data back to the corresponding population frequencies assuming a prevalence of CVD. In the following we calculate the bounds assuming a prevalence of 6.5% and also 2% to illustrate both ‘extremes’. First the output assuming a prevalence of 6.5%.

	$Z = 0$ (CC)		$Z = 1$ (CT)		$Z = 2$ (TT)	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$
$X = 0$ (Low)	341	272	297	269	63	56
$X = 1$ (High)	47	41	17	38	18	35

Table 4: Case-control (Y) frequencies by homocysteine (X) and *MTHFR* genotypes (Z) from Meleady et al. (2003, Table 3).

```
. mata
------------------------------------------------- mata (type end to exit) -----
: p = .065
: controls = (341, 47, 297, 17, 63, 18)
: cases = (272, 41, 269, 38, 56, 35)
:
: py0 = controls:*(1 - p)/sum(controls)
: py1 = cases:*p/sum(cases)
:
: z0 = sum(py0[1::2]) + sum(py1[1::2])
: z1 = sum(py0[3::4]) + sum(py1[3::4])
: z2 = sum(py0[5::6]) + sum(py1[5::6])
:
: pyxz0 = ((py0[1::2])/z0 \ py1[1::2]/z0)´
: pyxz1 = ((py0[3::4])/z1 \ py1[3::4]/z1)´
: pyxz2 = ((py0[5::6])/z2 \ py1[5::6]/z2)´
:
: st_matrix("pyxz0",pyxz0)
: st_matrix("pyxz1",pyxz1)
: st_matrix("pyxz2",pyxz2)
: end
--------------------------------------------------------------------------------

.
. bpboundsi , mat(pyxz0 pyxz1 pyxz2)
Data:                             Trivariate
Instrument categories:            3
-------------------------------------------------------------------------------
                                |
                                |                         Bounds
             Causal parameter   |                     Lower      Upper
--------------------------------+----------------------------------------------
   IV inequality constraints    |    satisfied
                          ACE   |                     -0.0895    0.7344
                 P(Y|do(X=0))   |                     0.0610     0.1200
                 P(Y|do(X=1))   |                     0.0305     0.7954
                          CRR   |                     0.2538     13.0348
--------------------------------+----------------------------------------------
   Assuming monotonicity:       |
    Monotonicity constraints    |    not satisfied
-------------------------------------------------------------------------------
```

Secondly, the output assuming a prevalence of 2% (omitting the output converting to population frequencies).

```
. bpboundsi , mat(pyxz0 pyxz1 pyxz2)
Data:                             Trivariate
Instrument categories:            3
─────────────────────────────────────────────────────────────────────────────
                             │
                             │                              Bounds
            Causal parameter │                          Lower     Upper
─────────────────────────────┼───────────────────────────────────────────────
   IV inequality constraints │    satisfied
                         ACE │                          -0.0650   0.7644
               P(Y|do(X=0))  │                          0.0188    0.0745
               P(Y|do(X=1))  │                          0.0095    0.7833
                         CRR │                          0.1272    41.5740
─────────────────────────────┼───────────────────────────────────────────────
   Assuming monotonicity:    │
    Monotonicity constraints │    not satisfied
─────────────────────────────────────────────────────────────────────────────
```

With a prevalence of 6.5% the IV inequality constraints are satisfied, the ACE lies between $-0.0895 \leq ACE \leq 0.7344$, and the monotonicity inequality constraints are not satisfied. With a prevalence of 2% the IV inequality constraints are satisfied, the bounds are slightly wider and the ACE lies between $-0.065 \leq ACE \leq 0.7644$, and the monotonicity inequality constraints are again not satisfied.

8.3 Simulated example that does not satisfy the IV conditions

We use simulated data to show that the IV inequality constraint check can both detect and fail to detect violations of the IV assumptions. We simulate two outcome variables, Y_1, Y_2, assuming a direct effect of the instrument on the outcome which violates assumption (iii) of Section 3.1. The strength of the direct effect is larger for Y_1 than Y_2. We simulate data from the following algorithm where U is the confounder, X the exposure, Y_i the outcomes, and Z the instrument.

$$
\begin{aligned}
Z &\sim Bern(0.5) \\
U &\sim Bern(0.5) \\
p_X &= 0.05 + 0.1Z + 0.1U, \quad X \sim Bern(p_X) \\
p_1 &= 0.1 + 0.2Z + 0.05X + 0.1U, \quad Y_1 \sim Bern(p_1) \\
p_2 &= 0.1 + 0.05Z + 0.05X + 0.1U, \quad Y_2 \sim Bern(p_2)
\end{aligned}
$$

We simulate 10,000 observations and run the `bpbounds` command.

```
. clear
. set seed 2232011
. set obs 10000
obs was 0, now 10000
. gen z = rbinomial(1,.5)
. gen u = rbinomial(1,.5)
. gen px = .05 + .1*z + .1*u
. gen x = rbinomial(1,px)
. gen p1 = .1 + .2*z + .05*x + .1*u
. gen y1 = rbinomial(1,p1)
. gen p2 = .1 + .05*z + .05*x + .1*u
. gen y2 = rbinomial(1,p2)
. bpbounds y1 (x = z)
Data:                              Trivariate
Instrument categories:             2
------------------------------------------------------------------------------
                              |
                              |                      Bounds
             Causal parameter |                 Lower     Upper
------------------------------+-----------------------------------------------
    IV inequality constraints |    not satisfied
------------------------------------------------------------------------------
. bpbounds y2 (x = z)
Data:                              Trivariate
Instrument categories:             2
------------------------------------------------------------------------------
                              |
                              |                      Bounds
             Causal parameter |                 Lower     Upper
------------------------------+-----------------------------------------------
    IV inequality constraints |    satisfied
                          ACE |                 -0.1767   0.6922
                P(Y|do(X=0))  |                 0.1542    0.2352
                P(Y|do(X=1))  |                 0.0585    0.8464
                          CRR |                 0.2488    5.4897
------------------------------+-----------------------------------------------
    Assuming monotonicity:    |
     Monotonicity constraints |    not satisfied
------------------------------------------------------------------------------
```

Running the analysis for Y_1 the IV inequality constraints are not satisfied and as such we don't continue with the IV analysis in this case. However, for Y_2 the IV inequality constraints are satisfied even though assumption (iii) is violated in this simulation. It is therefore always recommended to use subject matter background knowledge to justify the IV assumptions.

9 Discussion

We have described and implemented various versions and extensions of the nonparametric bounds originally proposed by Balke and Pearl (1997). The `bpbounds` and `bpboundsi` commands compute these for the average causal effect for an instrument with two or three categories, with and without assuming monotonicity, and for bivariate and trivariate data (`bpboundsi` only).

Before calculating these bounds, the inequality constraints imposed by the IV assumptions on the observable data should be checked, but as illustrated in Section 8.3 we should only expect this check to detect gross violations of the assumptions. It is therefore always recommended to draw on additional subject matter knowledge to justify the IV conditions because even small violations invalidate the IV analysis.

The upper and lower bounds on the ACE (or on the CRR, or intervention probabilities) must not be confused with confidence intervals. They are in fact the range of all 'physically' possible values, given the data, if we do not make any other assumptions than (i)–(iii) of Section 3.1 (or the additional monotonicity assumption (7)). The nonparametric bounds have been criticized as they will often be wide and contain ACE=0 (i.e. no causal effect of X on Y), as in all our examples. This is especially true when the association between IV and exposure X is weak (Clarke and Windmeijer 2010). Also, Greenland (2000) makes the point that some additional knowledge, e.g. about the direction of a possible causal effect, is usually available. However, any point estimates, with their corresponding confidence intervals, will rely on specific parametric assumptions on the distributions of X, Y and (usually implicitly) on U, which are difficult to verify from the observational data. We therefore find that it is generally advisable and important to compute the nonparametric bounds *in addition* to any point estimates as an indication of how much information the data contain on their own, as opposed to the information gained by additional parametric assumptions.

Further work could investigate bounds on the ACE for the four compliance types as discussed by Richardson and Robins (2010). An alternative set of bounds has also been proposed by Chesher (2010) based on a 'nearly' nonparametric model.

10 References

Angrist, J. D., and G. W. Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90(430): 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444–455.

Balke, A., and J. Pearl. 1994. Counterfactual probabilities: Computational methods, bounds, and applications. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, ed. R. de Mantaras and D. Poole, 46–54.

———. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439): 1172–1176.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *The Stata Journal* 3(1): 1–31.

———. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal* 7(4): 465–506.

———. 2010. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. http://ideas.repec.org/c/boc/bocode/s425401.html.

Bonet, B. 2001. Instrumentality Tests Revisited. In *Proc. 17th Conf. on Uncertainty in Artificial Intelligence*, ed. J. Breese and D. Koller, 48–55. Seattle, WA: Morgan Kaufmann.

Chesher, A. 2010. Instrumental variable models for discrete outcomes. *Econometrica* 78(2): 575–601.

Clarke, P., and F. Windmeijer. 2010. Instrumental variable estimators for binary outcomes. Technical Report WP 10/239, CMPO, University of Bristol, Bristol, UK. http://www.bristol.ac.uk/cmpo/publications/papers/2010/abstract239.html.

Davey Smith, G., and S. Ebrahim. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology* 32: 1–22.

Davey Smith, G., D. A. Lawlor, R. M. Harbord, N. J. Timpson, I. Day, and S. Ebrahim. 2007. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Medicine* 4(12): e352.

Dawid, A. P. 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1): 1–31.

———. 2003. *Highly Structured Stochastic Systems*, chap. Causal inference using influence diagrams: The problem of partial compliance, 45–65. Oxford, UK: Oxford University Press.

Didelez, V., S. Meng, and N. A. Sheehan. 2010. Assumptions of IV methods for observational epidemiology. *Statistical Science* 25(1): 22–40.

Didelez, V., and N. Sheehan. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 16: 309–330.

Gawrilow, E., and M. Joswig. 2000. polymake: a Framework for Analyzing Convex Polytopes. In *Polytopes — Combinatorics and Computation*, ed. G. Kalai and G. M. Ziegler, 43–74. Birkhäuser.

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 29: 722–729.

Imbens, G. W., and J. D. Angrist. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 62: 467–467.

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

Manski, C. F. 1990. Nonparametric bounds on treatment effects. *American Economic Review* 80: 319–323.

Meleady, R., P. M. Ueland, H. Blom, A. S. Whitehead, H. Refsum, L. E. Daly, S. E. Vollset, C. Donohue, B. Giesendorf, I. M. Graham, A. Ulvik, Y. Zhang, and A.-L. Bjorke Monsen. 2003. Thermolabile methylenetetrahydrofolate reductase, homocysteine, and cardiovascular disease risk: the European Concerted Action Project. *The American Journal of Clinical Nutrition* 77(1): 63–70.

Palmer, T. M., D. A. Lawlor, R. M. Harbord, N. A. Sheehan, J. H. Tobias, N. J. Timpson, G. Davey Smith, and J. A. C. Sterne. 2011. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research* Published online before print January 7, 2011. http://dx.doi.org/10.1177/0962280210394459.

Pearl, J. 1995a. On the testability of causal models with latent and instrumental variables. In *Uncertainty in Artificial Intelligence*, ed. P. Besnard and S. Hanks, vol. 11, 435–443. Morgan Kaufman, San Francisco.

———. 1995b. Causal inference from indirect experiments. *Artificial Intelligence in Medicine* 7(6): 561–582.

———. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York, US: Cambridge University Press.

Ramsahai, R. R. 2007. Causal Bounds and Instruments. In *Proceedings of the Twenty-Third Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, 310–317. Corvallis, Oregon: AUAI Press. http://uai.sis.pitt.edu/papers/07/p310-ramsahai.pdf.

———. 2011. Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research* (under revision).

Ramsahai, R. R., and S. L. Lauritzen. 2011. Likelihood analysis of the binary instrumental variable model. *Biometrika* (in press).

Richardson, T. S., and J. M. Robins. 2010. *Heuristics, Probability and Causality: A tribute to Judea Pearl*, chap. Analysis of the binary instrumental variable model, 415–444. London, UK: College Publications.

Robins, J. 1989. *Health Services Research Methodology: A focus on AIDS*, chap. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Washington DC, US: US Public Health Service.

Rubin, D. B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(5): 688–701.

———. 1978. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 6: 34–58.

Sommer, A., E. Djunaedi, A. A. Loeden, I. Tarwotjo, K. West, R. Tilden, L. Mele, and T. A. S. Group. 1986. Impact of vitamin A supplementation on childhood mortality: a randomised controlled community trial. *The Lancet* 327(8491): 1169 – 1173.

Tanaka, T., P. Scheet, B. Giusti, S. Bandinelli, M. G. Piras, G. Usala, S. Lai, A. Mulas, A. M. Corsi, A. Vestrini, F. Sofi, A. M. Gori, R. Abbate, J. Guralnik, A. Singleton, G. R. Abecasis, D. Schlessinger, M. Uda, and L. Ferrucci. 2009. Genome-wide Association Study of Vitamin B6, Vitamin B12, Folate, and Homocysteine Blood Concentrations. *The American Journal of Human Genetics* 84(4): 477–482.

Acknowledgements

This work was support by UK Medical Research Council Collaborative Project Grant G0601625 entitled, “Inferring epidemiological causality using Mendelian randomisation”. Tom Palmer is funded by the MRC Centre for Causal Analyses in Translational Epidemiology (Medical Research Council Grant G0600705). Vanessa Didelez acknowledges financial support from the Leverhulme Trust (grant RF-2011-320). Nuala Sheehan acknowledges financial support from the Leverhulme Trust (Research Fellowship RF/9/RFG/2009/0062).

About the authors

Tom Palmer is a Research Associate in Biostatistics, he assisted with the `winbugsfromstata` package and is the author of the `confunnel` command. Roland Ramsahai is a Research Fellow in Statistics, his PhD thesis was entitled “Causal inference with instruments and other supplementary variables”. Vanessa Didelez is a Senior Lecturer in Statistics and has ten years experience in research on causal inference from observational data. Nuala Sheehan is a Reader in Statistical Genetics with research interests in causal inference and genetic related problems in statistics.