The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2011)
11, Number 3, pp. 345–367

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

Tom M. Palmer
MRC Centre for Causal Analyses in Translational Epidemiology
School of Social and Community Medicine
University of Bristol, UK
tom.palmer@bristol.ac.uk

Roland R. Ramsahai
Statistical 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 (1997, *Journal of the American Statistical Association* 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 extensions applies when 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 instrumental-variable inequality constraints as a check for gross violations of the instrumental-variable conditions. The use of the commands is illustrated with a re-creation of the original Balke and Pearl analysis and with a Mendelian randomization analysis. We also give a simulated example to demonstrate that the instrumental-variable inequality constraints can both detect and fail to detect violations of the instrumental-variable conditions.

Keywords: st0232, 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 conditions 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, chap. 8). The nonparametric bounds have also been generalized by Ramsahai (2007, 2011) to cope with different data structures.

Typical applications of this methodology include randomized controlled trials (RCTs) with partial compliance, where random assignment is the instrument and the 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 called `bpbounds` and an immediate version called `bpboundsi`. We explain the IV conditions and how they allow bounds to be obtained for a causal effect. This explanation is followed by a description of the commands and a demonstration of their use on some examples.

2 The average causal effect

We define X to be the exposure variable and Y to be the outcome variable. We 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, and $Y = 1$ 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:

$$\text{ACE} = E\{Y|do(X = 1)\} - E\{Y|do(X = 0)\}$$

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 an 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 $P\{Y = 1|do(X = x)\}$, for example, the causal risk ratio (CRR) $P\{Y = 1|do(X = 1)\}/P\{Y = 1|do(X = 0)\}$.

3 IVs

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

3.2 Inequality constraints

The conditional independencies (i) and (iii) imply that $P(Y, X, U|Z) = P(Y|X, U) \times P(X|Z, U) \times P(U)$; this in turn implies that the observable marginal $P(Y, X|Z)$ is not unrestricted because 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 nontrivial constraints on $P(Y, X|Z)$ that can be expressed as a set of inequality constraints. "Nontrivial" 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)

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

For condition (ii), we simply need to check that $P(X = x|Z = z_1) \neq P(X = x|Z = z_2)$ for $z_1 \neq z_2$, that is, that 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

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

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 $p_{yx.z}$. 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; we will ignore this here, but Ramsahai and Lauritzen (forthcoming) 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 bounds 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.

$$\max \left\{ \begin{matrix} p_{10.1} \\ p_{10.0} \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{matrix} \right\} \leq \pi_0 \leq \min \left\{ \begin{matrix} 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{matrix} \right\} \quad (2)$$

and

$$\max \left\{ \begin{matrix} p_{11.0} \\ p_{11.1} \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1} \\ -p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \end{matrix} \right\} \leq \pi_1 \leq \min \left\{ \begin{matrix} 1 - p_{01.1} \\ 1 - p_{01.0} \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} \end{matrix} \right\} \quad (3)$$

Because ACE $= \pi_1 - \pi_0$, we can combine (2) and (3) to obtain bounds on the ACE; the lower bound is given by

$$\text{ACE} \geq \max \left\{ \begin{array}{c} p_{00.0} + p_{11.1} - 1 \\ p_{00.1} + p_{11.1} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{00.0} + p_{11.0} - 1 \\ 2p_{00.0} + p_{11.0} + p_{10.0} + p_{11.1} - 2 \\ p_{00.0} + 2p_{11.0} + p_{00.1} + p_{01.1} - 2 \\ p_{10.0} + p_{11.0} + 2p_{00.1} + p_{11.1} - 2 \\ p_{00.0} + p_{01.0} + p_{00.1} + 2p_{11.1} - 2 \end{array} \right\} \tag{4}$$

and the upper bound is given by

$$\text{ACE} \leq \min \left\{ \begin{array}{c} 1 - p_{10.0} - p_{01.1} \\ 1 - p_{01.0} - p_{10.1} \\ 1 - p_{01.0} - p_{10.0} \\ 1 - p_{01.1} - p_{10.1} \\ 2 - 2p_{01.1} - p_{10.0} - p_{10.1} - p_{11.1} \\ 2 - p_{01.0} - 2p_{10.0} - p_{00.1} - p_{01.1} \\ 2 - p_{10.0} - p_{11.0} - 2p_{01.1} - p_{10.1} \\ 2 - p_{00.0} - p_{01.0} - p_{01.1} - 2p_{10.1} \end{array} \right\} \tag{5}$$

Robins (1989) and Manski (1990) derived the first four lines of (4) and (5); Balke and Pearl (1997) tightened these by deriving the rest. Any combination of π_0 and π_1 in (2) and (3) is possible (Dawid 2003), and hence we can also obtain bounds for the CRR $(= \pi_1/\pi_0)$ as follows:

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

where π_x^L, π_x^U are the lower and upper bounds of π_x from (2) and (3).

4.2 The monotonicity assumption

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

$$P(X = 1|Z = 1, U = u) \geq P(X = 1|Z = 0, U = u) \tag{6}$$

which is a weaker version of the monotonicity assumption of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). 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 IV conditions (i) and (iii) together with (6) 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 (6) reduces the bounds on the ACE to

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

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, that is, $P(X = 1|Z = 0) = 0$. This implies that monotonicity (6) 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 formula quickly becomes prohibitive. The `bpbounds` and `bpboundsi` commands described below will also calculate the bounds when IV Z has three levels. Dawid (2003) and Ramsahai (2007, 2011) describe the general technique for how these can be obtained. An instrument with three levels is, for instance, relevant in Mendelian randomization 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 provide less information and hence lead to a different formula for the bounds on π_x and hence on the ACE. Ramsahai (2007, eq. 5) derives

the restrictions corresponding to the "check" of section 3.2 for bivariate data, as well as the formula for the bounds on the ACE corresponding to (4) and (5). 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 on, for example, 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 $P(Y = 1) = b$—to assess the sensitivity of the bounds to the assumed disease prevalence.

6 Interpretation of bounds

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)]. Because 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 (4) and (5) 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 (6) typically leads to tighter bounds. Another popular assumption is that $E(Y|X = x, U = u) = \beta x + h(u)$ for some function $h(\cdot)$, that is, additivity of the outcome model (Didelez, Meng, and Sheehan 2010). In this case, it can be shown that $\beta =$ ACE, where

$$\beta = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(X|Z = 1) - E(X|Z = 0)} = \frac{\mathrm{cov}(Y, Z)}{\mathrm{cov}(X, Z)} \tag{7}$$

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, Schaffer, and Stillman 2003, 2007, 2010). Because the point estimate (7) relies on specific parametric conditions, it is always advisable to compare it with the assumption-free bounds to assess sensitivity to these additional conditions.

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 (6) 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, 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) is as follows:

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

This follows the standard syntax for Stata's IV 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 IV (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 IV estimation commands, exogenous covariates are not allowed. Frequency weights, however, 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 is as follows:

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

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

Syntax for `bpboundsi` with an instrument with three categories is as follows:

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

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

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

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 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.

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

7.2 Options

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

`fmt(`*format*`)` changes the displayed 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 or conditional probabilities input in matrices. For trivariate data: The X categories must be the rows, and the Y categories must be the columns. The matrices must also be listed by ordered categories of Z, that is, conditional on $Z = 0$, $Z = 1$, and $Z = 2$. For 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.

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, with $Z = 1$ being the treatment group and $Z = 0$ being the control group. Children in the treatment group received two large doses of vitamin A, while those in the control group received no treatment. Not every child in the treatment group complied with the assignment, so $X = 1$ denotes treatment actually taken and $X = 0$ means no treatment taken. The control group necessarily had to comply because vitamin A supplements were not available to them. As noted in section 4.2, this automatically implies that the monotonicity assumption is satisfied because $P(X = 1|Z = 0) = 0$. The outcome Y was the number of deaths in both groups (where $Y = 1$ denotes survival). Table 3 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.

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

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

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

```
. 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
```

Causal parameter		Bounds 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 conditions, is satisfied. Then the command gives the bounds for the ACE, which are $-0.1946 \leq \text{ACE} \leq 0.0054$, as reported by Balke and Pearl (1997). We multiply the results by 100 to express in percentages, and 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 situation, the upper and lower bounds for $P\{Y = 1|do(X = 0)\}$ are equal because there was no noncompliance 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 noncompliance in the control group.

We could also use the immediate version of the 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.

```
. by z, sort: tabulate x y [fw=count], cell
```

```
-> z = 0
```

Key
frequency *cell percentage*

x	y 0	1	Total
0	74 0.64	11,514 99.36	11,588 100.00
Total	74 0.64	11,514 99.36	11,588 100.00

```
-> z = 1
```

Key
frequency *cell percentage*

x	y 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
```

Causal parameter		Bounds 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
```

Causal parameter		Bounds 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 (7).

```
. quietly correlate y z [fw=count], covariance
. scalar covyz = r(cov_12)
. quietly correlate x z [fw=count], covariance
. scalar covxz = r(cov_12)
. display "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` command, 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`.

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

Key
frequency *row percentage*

z	y 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

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

Key
frequency *row percentage*

z	x 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, matrices(zy zx) bivariate
Data:                              Bivariate
Instrument categories:             2
```

Causal parameter		Bounds 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 $-0.1974 \leq \text{ACE} \leq 0.0064$. 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 IVs (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. Such genotypes are promising candidates for IVs because the randomization of alleles at conception means that genotypes are very unlikely to be associated with potential confounding factors that 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), and Palmer et al. (forthcoming). 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 gene, involved in folate metabolism, as an IV (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 genome-wide association studies (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.

Table 4. Case–control (Y) frequencies by homocysteine (X) and Methylenetetrahydrofolate Reductase genotypes (Z) from Meleady et al. (2003, table 3)

	$Z = 0$ (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

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%:

```
. 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, matrices(pyxz0 pyxz1 pyxz2)
Data:                           Trivariate
Instrument categories:          3
```

Causal parameter		Bounds 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		

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

```
. bpboundsi, matrices(pyxz0 pyxz1 pyxz2)
Data:                           Trivariate
Instrument categories:          3
```

Causal parameter		Bounds 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 $-0.0895 \leq \text{ACE} \leq 0.7344$; and the monotonicity inequality constraints are not satisfied. With a prevalence of 2%, the IV inequality constraints are satisfied; the bounds are slightly wider, so the ACE lies between $-0.0650 \leq \text{ACE} \leq 0.7644$; 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 conditions. We simulate two outcome variables, Y_1 and 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 it is for Y_2. We simulate data from the following algorithm, where U is the confounder, X is the exposure, Y_i are the outcomes, and Z is the instrument.

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

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

```
. clear
. set seed 2232011
. set obs 10000
obs was 0, now 10000
```

```
. generate z = rbinomial(1,.5)
. generate u = rbinomial(1,.5)
. generate px = .05 + .1*z + .1*u
. generate x = rbinomial(1,px)
. generate p1 = .1 + .2*z + .05*x + .1*u
. generate y1 = rbinomial(1,p1)
. generate p2 = .1 + .05*z + .05*x + .1*u
. generate y2 = rbinomial(1,p2)
. bpbounds y1 (x = z)
Data:                           Trivariate
Instrument categories:          2
```

Causal parameter		Bounds Lower	Upper
IV inequality constraints	not satisfied		

```
. bpbounds y2 (x = z)
Data:                           Trivariate
Instrument categories:          2
```

Causal parameter		Bounds 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. As such, we do not 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 conditions.

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 ACE 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 conditions on the observable data should be checked; however, as illustrated in section 8.3, we should only expect this check to detect gross violations of the conditions. 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 assumptions other than (i)–(iii) of section 3.1 [or the additional monotonicity assumption (6)]. The nonparametric bounds have been criticized because they will often be wide and contain ACE $= 0$ (that is, 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, for example, 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.

Acknowledgments

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).

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: 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: 444–472.

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. L. de Mantaras and D. Poole, 46–54. San Francisco: Morgan Kaufmann.

———. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92: 1171–1176.

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

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

———. 2010. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. Statistical Software Components S425401, Department of Economics, Boston College.
http://ideas.repec.org/c/boc/bocode/s425401.html.

Bonet, B. 2001. Instrumentality tests revisited. In *Proceedings of the Seventeenth Annual Conference on Uncertainty in Artificial Intelligence*, ed. J. Breese and D. Koller, 48–55. San Francisco: Morgan Kaufmann.

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

Clarke, P., and F. Windmeijer. 2010. Instrumental variable estimators for binary outcomes. Working Paper 10/239, Centre for Market and Public Organisation, University of 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: e352.

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

———. 2003. Causal inference using influence diagrams: The problem of partial compliance. In *Highly Structured Stochastic Systems*, ed. P. J. Green, N. L. Hjort, and S. Richardson, 45–65. Oxford: Oxford University Press.

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

Didelez, V., and N. A. 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. Basil, Switzerland: 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: 467–475.

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: 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, A.-L. Bjorke Monsen, and The EC Concerted Action Project: Homocysteine and Vascular Disease. 2003. Thermolabile methylenetetrahydrofolate reductase, homocysteine, and cardiovascular disease risk: The European Concerted Action Project. *American Journal of Clinical Nutrition* 77: 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. Forthcoming. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*.

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. San Francisco: Morgan Kaufmann.

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

———. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: 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, OR: AUAI Press. http://uai.sis.pitt.edu/papers/07/p310-ramsahai.pdf.

———. 2011. Causal bounds and observable constraints for non-deterministic models. Unpublished manuscript.

Ramsahai, R. R., and S. L. Lauritzen. Forthcoming. Likelihood analysis of the binary instrumental variable model. *Biometrika*.

Richardson, T. S., and J. M. Robins. 2010. Analysis of the binary instrumental variable model. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, ed. R. Dechter, H. Geffner, and J. Y. Halpern, 415–444. London: College Publications.

Robins, J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Services Research Methodology: A Focus on AIDS*, ed. L. Sechrest, H. Freeman, and A. Mulley. Washington, DC: U.S. Public Health Service.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomised and non-randomised studies. *Journal of Educational Psychology* 66: 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. P. West, R. Tilden, L. Mele, and The Aceh Study Group. 1986. Impact of vitamin A supplementation on childhood mortality: A randomised controlled community trial. *Lancet* 327: 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. *American Journal of Human Genetics* 84: 477–482.

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. She 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.