

IV-estimators of the causal odds ratio for a continuous exposure in prospective and retrospective designs

Jack Bowden ¹, Stijn Vansteelandt ²

¹MRC Biostatistics Unit, Cambridge

²Department of Applied Mathematics and Computer Science, Ghent University

August 21, 2009



Introduction

- Motivation

- Mendelian randomisation - prospective data

- IV and Adjusted IV approaches

- Implementing the Adjusted IV approach

Application to retrospective data

- Invalidity of IV methods

- Exact and approximate solutions

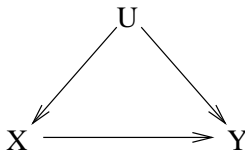
Simulation

Summary

Public health motivation

- ▶ For observed binary Y , continuous X and unobserved U :

$$\text{logitPr}(Y = 1|X = x, U = u) = \beta_0 + \beta_1x + \beta_2u$$

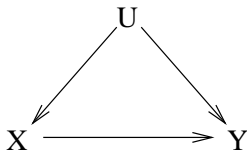


- ▶ "What would be the [population] effect of intervening and changing the level of risk factor X on the incidence of outcome Y ?"
- ▶ NOT interested in confounded $X - Y$ association!

Public health motivation

- ▶ For observed binary Y , continuous X and unobserved U :

$$\text{logitPr}(Y = 1|X = x, U = u) = \beta_0 + \beta_1x + \beta_2u$$

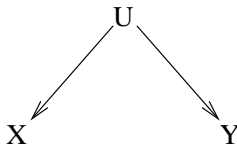


- ▶ "What would be the [population] effect of intervening and changing the level of risk factor X on the incidence of outcome Y ?"
- ▶ **NOT** interested in confounded $X - Y$ association!

Public health motivation

- ▶ For observed binary Y , continuous X and unobserved U :

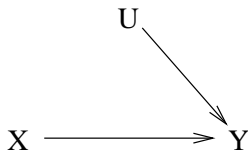
$$\text{logitPr}(Y = 1|X = x, U = u) = \beta_0 + \beta_1x + \beta_2u$$



- ▶ "What would be the [population] effect of intervening and changing the level of risk factor X on the incidence of outcome Y ?"
- ▶ **NOT** interested in confounded $X - Y$ association!

Our modest aim: to obtain an 'RCT' style estimate

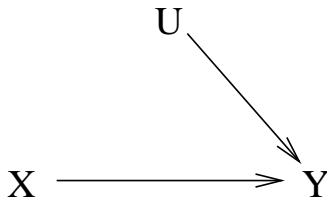
- ▶ If we could **fix** X then:



- ▶ $\text{logitPr}(Y = 1 | \text{do}(X = x_0), U = u) = \beta_0 + \beta_1 x_0 + \beta_2 u$
- ▶ x_0 now evidently $\perp\!\!\!\perp u \dots$
- ▶ ...but u still affects y

Pearl (1995)

This...



$$\begin{aligned}
 CLOR(x_0, x_0 + 1) &= \log \left\{ \frac{\text{odds } Pr(Y = 1 | do(X = x_0 + 1))}{\text{odds } Pr(Y = 1 | do(X = x_0))} \right\} \\
 &= \beta_1 \{ \beta_2^2 \text{Var}(U) \}
 \end{aligned}$$

...**NOT** necessarily this

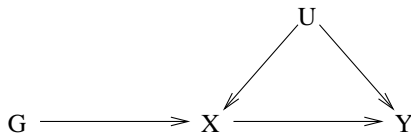
U

X \longrightarrow Y

The conditional causal effect

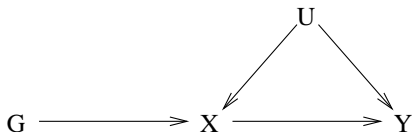
$$\beta_1 = \log \left\{ \frac{\text{odds } Pr(Y=1|do(X=x_0+1),U)}{\text{odds } Pr(Y=1|do(X=x_0),U)} \right\}$$

A possible solution - Mendelian randomisation



- ▶ Genetic quantity ' G ' must:
 1. be independent of U
 2. be predictive of X
 3. only affect Y through X

Conventional statistical approach



1. Regress X on G to obtain $E[X|G]$
2. fit model

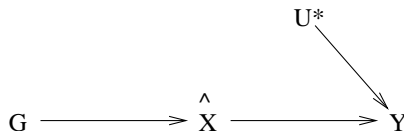
$$\text{logitPr}(Y = 1|X) = \alpha_0 + \alpha_1 E[X|G]$$

3. take $\hat{\alpha}_1$ as CLOR

► Simple to implement, poor performance ($\text{Var}[U] \neq \text{Var}[U^*]$)

Didelez et al. (2008); Palmer et al. (2008)

Conventional statistical approach



1. Regress X on G to obtain $E[X|G]$
2. fit model

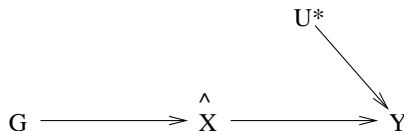
$$\text{logitPr}(Y = 1|X) = \alpha_0 + \alpha_1 E[X|G]$$

3. take $\hat{\alpha}_1$ as CLOR

► Simple to implement, poor performance ($\text{Var}[U] \neq \text{Var}[U^*]$)

Didelez et al. (2008); Palmer et al. (2008)

Conventional statistical approach



1. Regress X on G to obtain $E[X|G]$
2. fit model

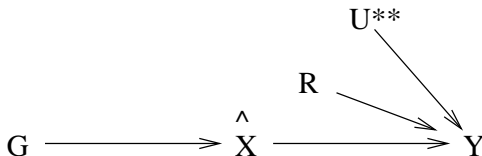
$$\text{logitPr}(Y = 1|X) = \alpha_0 + \alpha_1 E[X|G]$$

3. take $\hat{\alpha}_1$ as CLOR
- ▶ Simple to implement, poor performance ($\text{Var}[U] \neq \text{Var}[U^*]$)

Didelez et al. (2008); Palmer et al. (2008)

Adjusted IV approach

Palmer et al. (2008); Nagelkerke et al. (2000)

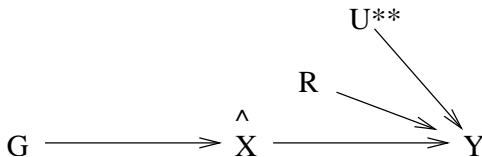


1. Define residual $R = X - E[X|G]$
2. fit model

$$\text{logitPr}(Y = 1|X, R) = \alpha_0 + \alpha_1 E[X|G] + \alpha_2 R$$

3. take $\hat{\alpha}_1$ as CLOR

Comments: $\text{logitPr}(Y = 1|X, R) = \alpha_0 + \alpha_1 E[X|G] + \alpha_2 R$



- ▶ Better performance ($\text{Var}[U] \geq \text{Var}[U]^{**} \geq 0$)
- ▶ Can be equivalent to fitting a Logistic SMM
- ▶ What is $\text{Var}(\hat{\alpha}_1)$?
- ▶ What if parametric model for $E[X|G]$ wrong?

Suggested implementation: Score function the adjusted IV

- ▶ Specify $x_i(\theta) = \theta_0 + \theta_1 g_i$, $r_i(\theta) = x_i - \theta_0 + \theta_1 g_i$
- ▶ Solve score equation $\sum_{i=1}^n S_i(\theta) = 0$

$$S_i(\theta) = \begin{pmatrix} \begin{pmatrix} 1 \\ g_i \end{pmatrix} & W_i r_i(\theta) \\ \begin{pmatrix} 1 \\ x_i(\theta) \\ r_i(\theta) \end{pmatrix} & [y_i - \text{expit}\{\theta_3 + \theta_4 x_i(\theta) + \theta_5 r_i(\theta)\}] \end{pmatrix}$$

- ▶ CLOR is $\hat{\theta}_4$, $\text{var}(CLOR)$ is 4th element of 'sandwich variance'
- ▶ $W_i = 1$ for all subjects with prospective observational data

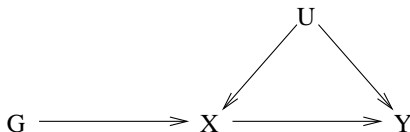
Suggested implementation: Score function the adjusted IV

- ▶ Specify $x_i(\theta) = \theta_0 + \theta_1 g_i$, $r_i(\theta) = x_i - \theta_0 + \theta_1 g_i$
- ▶ Solve score equation $\sum_{i=1}^n S_i(\theta) = 0$

$$S_i(\theta) = \begin{pmatrix} \begin{pmatrix} 1 \\ g_i \end{pmatrix} & W_i r_i(\theta) \\ \begin{pmatrix} 1 \\ x_i(\theta) \\ r_i(\theta) \end{pmatrix} & [y_i - \text{expit}\{\theta_3 + \theta_4 x_i(\theta) + \theta_5 r_i(\theta)\}] \end{pmatrix}$$

- ▶ CLOR is $\hat{\theta}_4$, $\text{var}(CLOR)$ is 4th element of 'sandwich variance'
- ▶ $W_i = 1$ for all subjects with prospective observational data

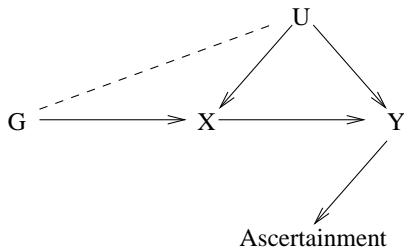
Application to retrospective data



$G \perp\!\!\!\perp U$ assumption highly plausible for prospective data

G NOT $\perp\!\!\!\perp U$ for case-control data

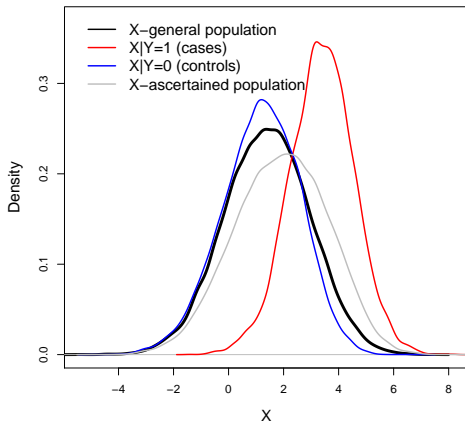
Didelez and Sheehan (2007)



- ▶ 'Moral' edge induced by a sampling conditional on Y .
- ▶ IV analysis invalid (wrt estimation)
- ▶ Most IV analyses are performed using case-control data!

solution: re-weighting the exposure distribution

Distribution of X for 1:2 case-control data ($\pi = \frac{1}{3}, P(Y = 1) = 5\%$)

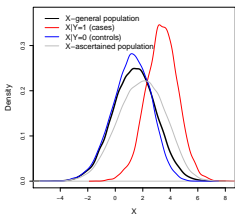


$$f_{gen}(X) = f(X|Y = 1)Pr(Y = 1) + f(X|Y = 0)Pr(Y = 0)$$

$$f_{asc}(X) = f(X|Y = 1)\pi + f(X|Y = 0)(1 - \pi)$$

solution: re-weighting the exposure distribution

Exact and Crude solutions



- ▶ Employ 'exact' weights W :

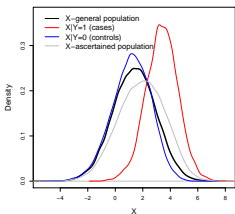
$$f_{gen} = f(X|Y = 1)\pi \frac{Pr(Y=1)}{\pi} + f(X|Y = 0)(1 - \pi) \frac{Pr(Y=0)}{1-\pi}$$

- ▶ Or 'crude' weights W :

$$f_{gen} \approx f(X|Y = 1)\pi 0 + f(X|Y = 0)(1 - \pi) 1$$

solution: re-weighting the exposure distribution

Exact and Crude solutions



- ▶ Employ 'exact' weights W :

$$f_{gen} = f(X|Y = 1)\pi \frac{Pr(Y=1)}{\pi} + f(X|Y = 0)(1 - \pi) \frac{Pr(Y=0)}{1-\pi}$$

- ▶ Or 'crude' weights W :

$$f_{gen} \approx f(X|Y = 1)\pi 0 + f(X|Y = 0)(1 - \pi) 1$$

Performance of crude and exact weighting

β_2	Prospective data estimates		Retrospective data: Av.diff(variance)coverage					
	Alt IV	LSMM	exact W			crude W		
			Adjusted IV	LSMM	Adjusted IV	LSMM	Adjusted IV	LSMM
(Other parameter values - ($\beta_0 = -5, \alpha_0 = 0, \alpha_1 = 1, \beta_1 = 1$))								
0.00	1.010	1.010	0.001 (0.015)0.944	0.001 (0.015)0.944	0.000 (0.017)0.946	0.001 (0.017)0.944	0.000 (0.017)0.946	0.001 (0.017)0.944
0.25	1.009	1.009	0.003 (0.015)0.952	0.003 (0.015)0.952	0.012 (0.017)0.950	0.013 (0.018)0.950	0.012 (0.017)0.950	0.013 (0.018)0.950
0.50	0.998	0.998	0.003 (0.015)0.960	0.002 (0.015)0.960	0.025 (0.017)0.952	0.029 (0.018)0.948	0.025 (0.017)0.952	0.029 (0.018)0.948
0.75	0.978	0.978	0.004 (0.015)0.947	0.003 (0.015)0.949	0.041 (0.017)0.939	0.050 (0.018)0.930	0.041 (0.017)0.939	0.050 (0.018)0.930
1.00	0.944	0.943	0.000 (0.015)0.942	0.000 (0.016)0.945	0.052 (0.017)0.938	0.069 (0.018)0.924	0.052 (0.017)0.938	0.069 (0.018)0.924

- ▶ Exact weighting of retrospective data as good as standard analysis with prospective data
- ▶ Crude weighting harder to justify and more poorly performing with LSMM

Summary/Further work

- ▶ Robust & General implementation of Adjusted IV approach
- ▶ Adjusted IV and Logistic SMM equivalent in particular setting
 - ▶ Both can handle Case-control data, given $P(Y = 1)$
 - ▶ Crude method better for adjusted IV method
- ▶ Extension: IV methods for matched case-control data

- BOWDEN, J., VANSTEELANDT, S. (2009) IV-estimators of the causal odds ratio for a continuous exposure in prospective and retrospective designs. submitted to *Biostatistics*.
- DIDELEZ, V. AND SHEEHAN, N. (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330.
- DIDELEZ, V., S.MENG, AND N.SHEEHAN (2008). On the bias of iv estimators for mendelian randomisation. Technical report, University of Bristol.
- NAGELKERKE, N., FIDLER, V., R.BERNSSEN, AND BORGDORFF, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19:1849–64.
- PALMER, T., THOMPSON, J., TOBIN, M., SHEEHAN, N., AND BURTON, P. (2008). Adjusting for bias and unmeasured confounding in mendelian randomisation studies with binary responses. *IJE*, 37: 1161–1168
- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:699–710.
- VANSTEELANDT, S. AND GOETGHEBEUR, E. (2003). Causal inference with generalized structural mean models. *JRSSB*, 65:817–835.
- WHITTEMORE, A. (1995). Logistic regression of family data from case-control studies. *Biometrika*, 82:57–67.

Equivalence of Adjusted IV and Logistic SMM

Vansteelandt and Goetghebeur (2003)

- ▶ Let $E[X|G] = \theta_0 + \theta_1 G$

$$\begin{aligned}
 \text{logitPr}(Y = 1|X, R) &= \alpha_0 + \alpha_1 E[X|G] + \alpha_2 (X - E[X|G]) \\
 &= \gamma_0 + \alpha_2 X + \gamma_2 G \\
 &= \text{logitPr}(Y = 1|X, G) \\
 &= \text{Association model for LSMM}
 \end{aligned}$$

- ▶ When X normal & $\text{Var}(X|G) \perp\!\!\!\perp G$
LSMM and Adj IV target same CLOR

Equivalence of Adjusted IV and Logistic SMM

Vansteelandt and Goetghebeur (2003)

- ▶ Let $E[X|G] = \theta_0 + \theta_1 G$

$$\begin{aligned}
 \text{logitPr}(Y = 1|X, R) &= \alpha_0 + \alpha_1 E[X|G] + \alpha_2 (X - E[X|G]) \\
 &= \gamma_0 + \alpha_2 X + \gamma_2 G \\
 &= \text{logitPr}(Y = 1|X, G) \\
 &= \text{Association model for LSMM}
 \end{aligned}$$

- ▶ When X normal & $\text{Var}(X|G) \perp\!\!\!\perp G$
LSMM and Adj IV target same CLOR

Implementation of logistic SMM

- Find ψ via *G-estimation*:

$$\Pr(Y(0) = 1|X, G) \approx \Pr(Y = 1|X, G)\exp(-\psi X) \perp\!\!\!\perp G$$

Score equation:

$$S_i(\theta) = \begin{pmatrix} W_i(g_i - E_W[g])(\text{expit}\{\theta_1 + (\theta_2 - \psi)x_i + \theta_3 g_i\} - q) \\ \begin{pmatrix} 1 \\ x_i \\ g_i \end{pmatrix} [y_i - \text{expit}\{\theta_1 W + \theta_2 x_i + \theta_3 g_i\}] \end{pmatrix}$$

where $q = E_W[\text{expit}\{\theta_1 + (\theta_2 - \psi)x + \theta_3 g\}]$

$$\theta_{1,W} = \theta_1 + \frac{\pi \Pr(Y=0)}{(1-\pi)\Pr(Y=1)}$$

Whittemore (1995)

Implementation of logistic SMM

- Find ψ via *G-estimation*:

$$Pr(Y(0) = 1|X, G) \approx Pr(Y = 1|X, G) \exp(-\psi X) \perp\!\!\!\perp G$$

Score equation:

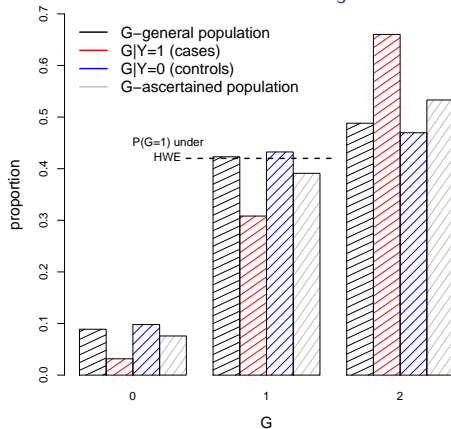
$$S_i(\theta) = \begin{pmatrix} W_i(g_i - E_W[g]) (\text{expit} \{ \theta_1 + (\theta_2 - \psi)x_i + \theta_3 g_i \} - q) \\ \begin{pmatrix} 1 \\ x_i \\ g_i \end{pmatrix} [y_i - \text{expit} \{ \theta_1 W + \theta_2 x_i + \theta_3 g_i \}] \end{pmatrix}$$

where $q = E_W[\text{expit} \{ \theta_1 + (\theta_2 - \psi)x + \theta_3 g \}]$

$$\theta_{1,W} = \theta_1 + \frac{\pi Pr(Y=0)}{(1-\pi)Pr(Y=1)}$$

Whittemore (1995)

Distribution of G for 1:2 case-control data ($\pi = \frac{1}{3}$)



$$f_{gen}(G) = f(G|Y=1)Pr(Y=1) + f(G|Y=0)Pr(Y=0)$$

$$f_{asc}(G) = f(G|Y=1)\pi + f(G|Y=0)(1-\pi)$$

Whittemore (1995)

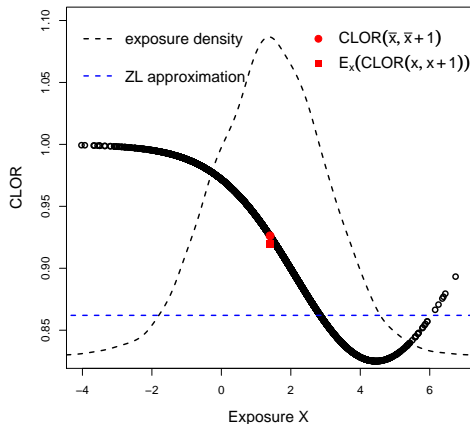


Figure: Exposure level x versus $CLOR(x, x + 1)$ for a distribution of 5000 exposures generated as in Section ?? . Highlighted in red are $CLOR(\bar{x}, \bar{x} + 1)$ and $E_x[CLOR(x, x + 1)]$ for this data.