An adjusted instrumental-variable model for Mendelian randomization

Tom Palmer, Paul Burton, John Thompson and Martin Tobin

Department of Health Sciences, University of Leicester

tmp8@le.ac.uk

Summary

- *The standard approach to the instrumental-variable analysis of a binary outcome is biased.*
- *The proposed adjusted approach has better properties as demonstrated via simulations.*

Introduction

MENDELIAN randomization uses the associations between genotype and disease and between genotype and phenotype to make inferences about the association between phenotype and disease [1].

In the case where all variables are continuous measures, traditional two-stage least squares instrumental variable (IV) methods are appropriate, however, the majority of genetic epidemiological studies have binary responses.

A standard approach to Mendelian randomization would be to use a logistic regression model at the second stage of the IV procedure, with perhaps some adjustment of the standard errors [2].

However with a binary disease variable the logistic regression is affected by shrinkage bias and unmeasured confounding [3]. An adjusted IV estimator is proposed and investigated through a simulation study.

Modelling approaches

THREE modelling approaches were considered, termed; direct, standard IV and adjusted IV. The notation used for the approaches is given in Table 1.

For an individual i, the direct approach is given by the logistic regression of the phenotype on disease status,

$$\textbf{direct approach:} \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i.$$

The first stage of the standard and adjusted IV approaches is given by the regression of the genotype on the phenotype,

$$\text{first stage:} \quad x_i = \alpha_0 + \alpha_1 g_i.$$

At the second stage the standard IV approach uses the logistic regression of the predicted phenotype on disease status,

$$\textbf{standard IV approach:} \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \widehat{x}_i.$$

The second stage of the adjusted IV approach is given by the logistic regression of the predicted phenotype and the estimated residuals on disease status,

$$\text{estimated residuals:} \quad r_i = x_i - \widehat{x}_i,$$

$$\textbf{adjusted IV approach:} \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \widehat{x}_i + \beta_r r_i.$$

Simulations

SIMULATIONS were undertaken to investigate the properties of the approaches. The structure of the simulation study is given in Figure 1 and the notation is explained in Table 1.

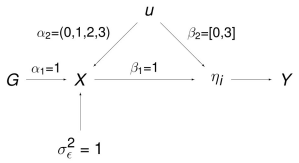

Figure 1: *The relationship between the variables and the parameter values used in the simulations (η_i denotes the linear predictor of the logistic regression).*

G	Genotype	X	Phenotype
Y	Disease	U	Confounder
α_1	gene-phenotype association	α_2	confounder-phenotype association
β_1	phenotype-disease log odds ratio	β_2	confounder-disease association
σ^2_ϵ	variance of phenotype error term	p_i	probability of disease
β_0	Baseline risk of disease	α_0	gene-phenotype intercept

Table 1: *Notation used to describe the approaches and the simulations*

For a cohort of 10,000 individuals, the genotype variable was generated using a minor allele frequency of 30% and by assuming Hardy–Weinberg Equilibrium. The phenotype variable was generated to be Normally distributed.

Results

FOUR scenarios of confounding were simulated by varying the magnitude of confounder-phenotype coefficient α_2. In Figure 2 the correct value of β_1 is 1.

The difference in the bias in the three approaches was consistent over the three scenarios in Figure 2 where α_2 was non-zero. In these scenarios the direct approach provided an overestimate of β_1 whilst the standard approach provided an underestimate. The adjusted approach provided the best estimate of β_1.

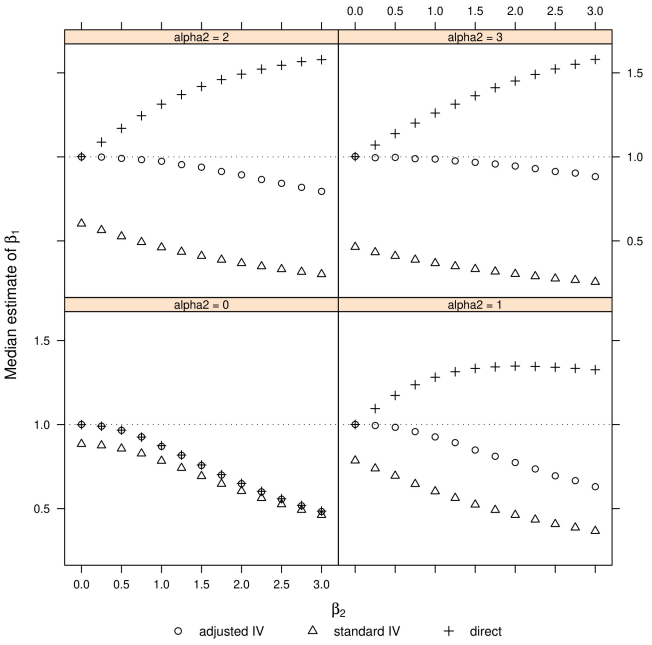

Figure 2: *Median estimates of the phenotype-disease log odds-ratio.*

Discussion

IN these simulations the direct and standard IV modelling approaches have been shown to provide positively and negatively biased parameter estimates respectively in the presence of unmeasured confounding factors.

The adjusted IV approach is superior in terms of reducing the bias in the parameter estimates by accounting for unmeasured confounding factors, and, mitigating the shrinkage bias.

Similar results hold if the Logistic regressions in the modelling approaches are replaced by Probit regressions [3].

References

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

[2] J. W. Hardin and R. J. Carroll. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *The Stata Journal*, 3(4):342–350, 2003.

[3] S. L. Zeger, K-Y. Liang, and P. S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060, 1988.