

Including multiple instrumental variables in Mendelian randomization analyses

Tom Palmer

MRC Centre for Causal Analyses in Translational Epidemiology,
Department of Social Medicine, University of Bristol

November 2009

MRC

Centre for Causal
Analyses in Translational
Epidemiology



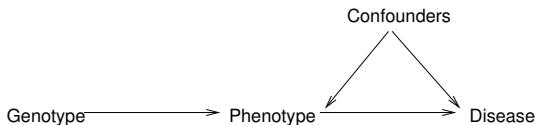
University of
BRISTOL

Outline

- ▶ Introduction:
 - the Mendelian randomization approach
 - rationale for multiple instruments
- ▶ Multiple instruments example:
 - increase precision
 - instrument strength
 - over-identification
 - use of allele score as IV
- ▶ Multiple instruments discussion

Introduction to the Mendelian randomization approach

- ▶ Use of genotypes as instrumental variables (Davey Smith & Ebrahim, 2003)
- ▶ Epi analyses - potential for unmeasured confounding, reverse causation
- ▶ Infer causal phenotype-disease association



IV assumptions; the genotype should be:

- (i) independent of confounders
- (ii) associated with phenotype
- (iii) independent of disease given phenotype and confounders

Rationale for using multiple instruments

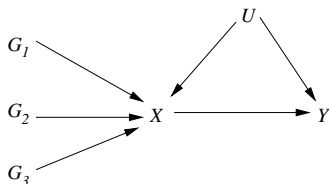
Problem: MR analyses - low power

- ▶ Weak instruments - bias IV estimate
- ▶ $F < 10$: $0.015 < p < 0.05$ GP coefficient stat. sig. & weak
(Lawlor, Harbord, Sterne, Timpson, & Davey Smith, 2008)
- ▶ Genotypes - small effects wrt phenotype st.dev. - wide IV CIs

Rationale for using multiple instruments

Problem: MR analyses - low power

- ▶ Weak instruments - bias IV estimate
- ▶ $F < 10$: $0.015 < p < 0.05$ GP coefficient stat. sig. & weak (Lawlor et al., 2008)
- ▶ Genotypes - small effects wrt phenotype st.dev. - wide IV CIs
- ▶ Ideal situation (Didelez & Sheehan, 2007):



Possible solutions:

- ▶ Increase study sample size
- ▶ Stronger instrument
- ▶ Multiple instruments
- ▶ (Meta-analysis)

Estimation

- ▶ Single instrument - ratio of coefficients: $PD = GD/GP$
- ▶ Multiple instruments - motivation for TSLS
(Theil, 1953; Basmann, 1957)
- ▶ TSLS, LIML, GMM estimators (`ivregress`, `ivreg2`, `condivreg`)
 - equivalent with 1 instrument & 1 phenotype
 - differ slightly with multiple instruments
 - LIML - smallest finite sample bias (Ullah, 2004)
 - Other types of CIs: AR (LIML), LM, CLR (`condivreg`)
(Mikusheva & Poi, 2006)
 - Bayesian approaches (Kleibergen & Zivot, 2003)

Multiple instruments

- ▶ Multiple instruments: Cragg-Donald F -statistic
 - inflates first stage F -statistic
 - (Cragg & Donald, 1993; Stock, Wright, & Yogo, 2002)
- ▶ Over-identification: Sargan & Hansen tests
 - if significant 1 or more instruments not valid
- ▶ Practical issue: increase in precision about IV estimate due to use of multiple IVs
 - countered by loss of precision due to extra missing data

Is fat mass causally related to bone mineral density?

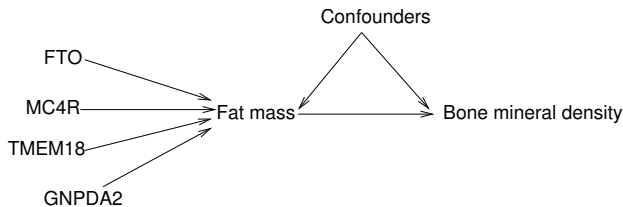
- ▶ Eligible sample: 5509 children, age 7-9yrs, ALSPAC cohort
- ▶ Outcome: bone mineral density
- ▶ Phenotype: fat mass (DXA scan)

Is fat mass causally related to bone mineral density?

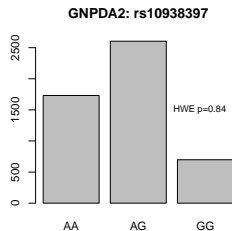
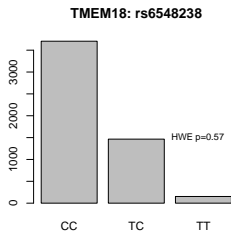
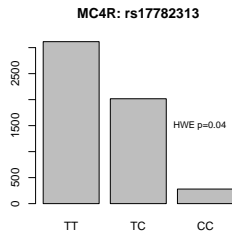
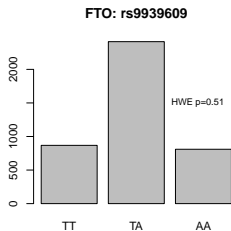
- ▶ Eligible sample: 5509 children, age 7-9yrs, ALSPAC cohort
 - ▶ Outcome: bone mineral density
 - ▶ Phenotype: fat mass (DXA scan)
 - ▶ IVs: FTO, MC4R, TMEM18, GNPDA2
 - Chromosomes 16, 18, 2, 4
- (Frayling et al., 2007; Loos et al., 2008; Willer et al., 2009)

Is fat mass causally related to bone mineral density?

- ▶ Eligible sample: 5509 children, age 7-9yrs, ALSPAC cohort
 - ▶ Outcome: bone mineral density
 - ▶ Phenotype: fat mass (DXA scan)
 - ▶ IVs: FTO, MC4R, TMEM18, GNPDA2
 - Chromosomes 16, 18, 2, 4
- (Frayling et al., 2007; Loos et al., 2008; Willer et al., 2009)



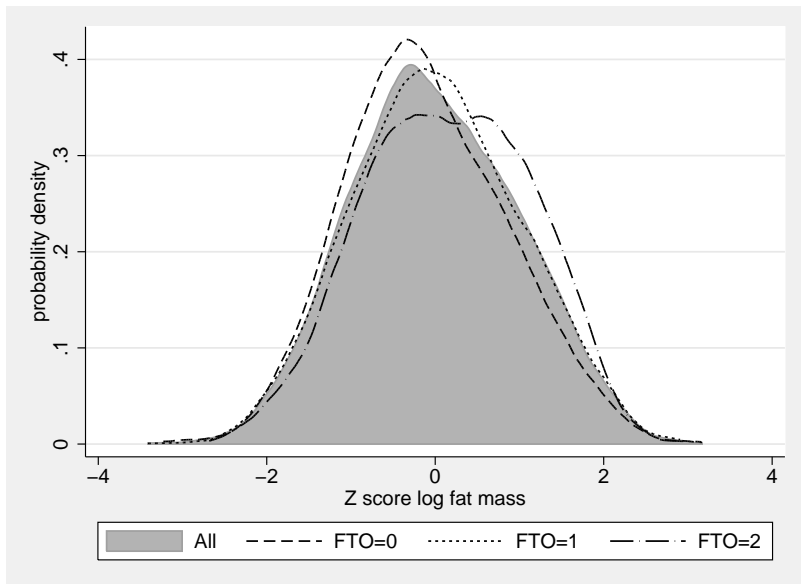
Distribution of SNPs in sample



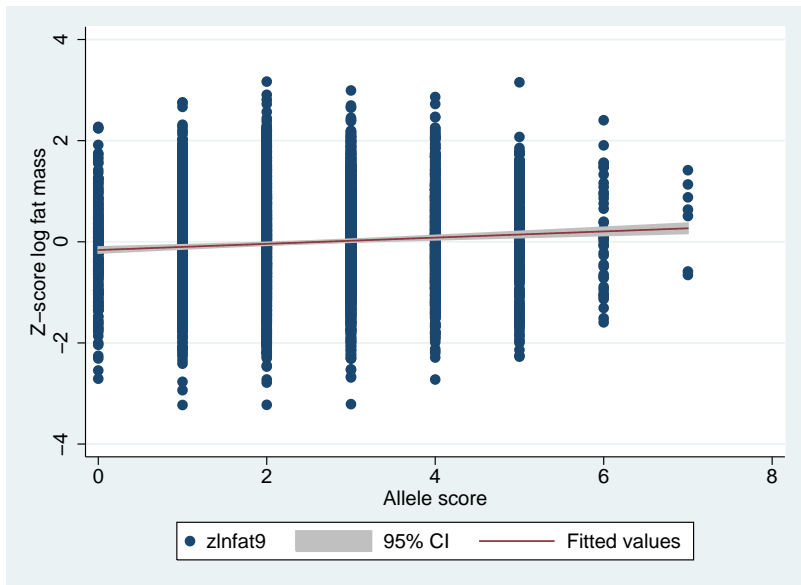
Example continued

- ▶ IV estimation:
 - each SNP separately
 - all four SNPs
 - allele score as IV - sum of risk alleles (Weedon et al., 2008)
- ▶ Fat mass & bone mineral density:
 - +vely skewed; logged & z-scored
 - $\exp(\beta)$; ratio of geometric means (RGM)

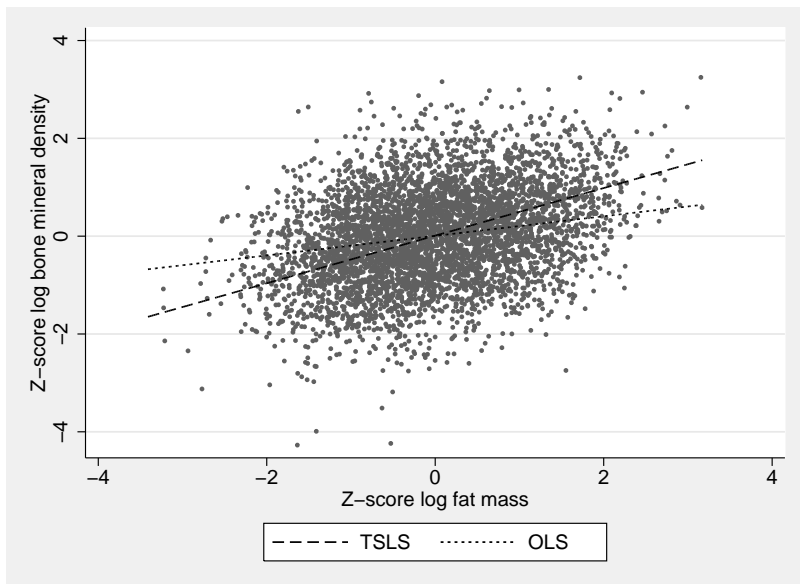
Distribution of log fat mass by FTO genotypes



First stage regression



Second stage regression



OLS: 1.22 (1.19, 1.26); IV allele score: 1.40 (0.99, 1.98)

IV estimates

Model	RGM (95% CI)	F
FTO	1.44 (1.05, 1.97)	39.8
MC4R	2.33 (1.34, 4.05)	17.9
TMEM18	2.27 (0.98, 5.28)	7.5
GNPDA2	0.98 (0.47, 2.03)	7.6
4 SNPs: TSLS	1.63 (1.28, 2.06)	18.6 _{16.9}
AR(LIML)	1.66 (1.29, 2.23)	
LM	(1.30, 2.21)	
CLR	(1.30, 2.20)	
Allele sc.	1.40 (0.99, 1.98)	33.2

- precision

- strength

- missing data: $N = 4796$

FTO: 5091; MC4R: 5412; TMEM18: 5323; GNPDA2: 5303

- over-id: 4 SNPs Sargan test, $P=0.16$

Multiple instruments discussion

- ▶ One way to increase precision of IV estimates
 - given each instrument meets IV assumptions
- ▶ Investigate:
 - joint strength - Cragg-Donald F -statistic
 - over-identification - Sargan test

Acknowledgements

MRC collaborative grant G0601625: Methods for Mendelian randomization

Collaborators: Nuala Sheehan, Vanessa Didelez, Sha Meng, Roger Harbord, John Thompson, Paul Clarke, Frank Windmeijer, Paul Burton, George Davey Smith.

References I

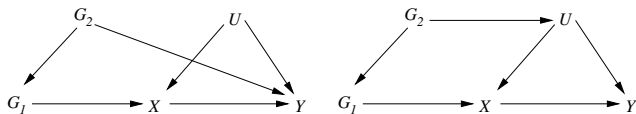
- Basman, R. L. (1957). A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. *Econometrica*, 25(1), 77–83.
- Cragg, J. G., & Donald, S. G. (1993). Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, 9, 222–240.
- Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *International Journal of Epidemiology*, 32, 1–22.
- Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16, 309–330.
- Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, 316(5826), 889–894.
- Kleibergen, F., & Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1), 29–72.
- Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8), 1133–1163.
- Loos, R. J. F., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., et al. (2008). Common variants near mc4r are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6), 768–775. Available from <http://dx.doi.org/10.1038/ng.140>

References II

- Mikusheva, A., & Poi, B. (2006). Tests and confidence sets with correct size when instruments are potentially weak. *The Stata Journal*, 6(3), 335–347.
- Stock, J. H., Wright, J. H., & Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics*, 20(4), 518–529.
- Theil, H. (1953). Repeated Least Squares Applied to Complete Equation Systems. *The Hague: Central Planning Bureau*.
- Ullah, A. (2004). *Finite sample econometrics*. Oxford, UK: Oxford University Press.
- Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., et al. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, 40(5), 575–583.
- Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics*, 41, 25–34.

Possible practical problems

- ▶ Population stratification
 - mixture of genetic popns in sample
 - could confound IV estimates
- ▶ Linkage disequilibrium
 - problematic if a gene in LD with instrument is associated with the outcome and/or confounders



- ▶ Pleiotropy
 - SNP has multiple functions
 - problematic if one of these functions is associated with the outcome

Missing data: multiple imputation & IV estimation

- ▶ Misleading to think about causal model in context of MI
 - more important to think about missing data mechanism - MNAR/MAR/MAR-CD/MCAR
- ▶ Although including all available variables in all chained equations would appear to violate IV DAG:
 - this doesn't matter if MAR
- ▶ SNPs: Maternal genotypes available in ALSPAC - impute offspring genotypes
 - HapMap type imputations (`impute/mach`)