Generalised method of moments estimation of structural mean models

Tom Palmer[1] Paul Clarke[2] Frank Windmeijer[2,3,4]

1. MRC CAiTE Centre, School of Social and Community Medicine, University of Bristol, UK
2. CMPO, University of Bristol, UK
3. Department of Economics, University of Bristol, UK
4. CEMMAP/IFS, London, UK

23 August 2011

MRC Centre for Causal Analyses in Translational Epidemiology

University of BRISTOL

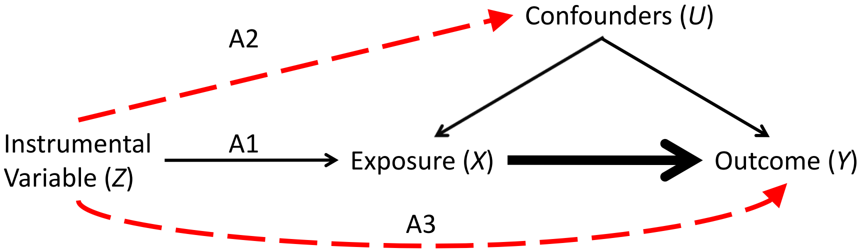
Outline

Generalised method of moments estimation of structural mean models . . . using instrumental variables

- Introduction to Mendelian randomization example
- Multiplicative structural mean model
  - Identification, G-estimation
  - GMM & Hansen over-id test
  - Implementation Stata & R, example estimates
  - Multiple instruments
- (double) Logistic SMM
  - Joint estimation of association & causal models
- Local risk ratios
- Summary

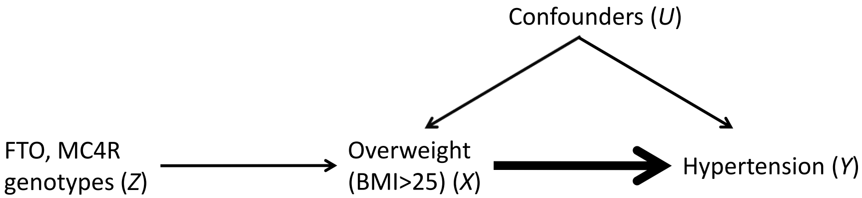
Introduction to Mendelian randomization example

- Mendelian randomization:
  use of genotypes robustly associated with exposures (from replicated genome-wide association studies, $P < 5 \times 10^{-8}$) as instrumental variables (Davey Smith & Ebrahim, 2003)



Introduction to Mendelian randomization example

- Mendelian randomization:
  use of genotypes robustly associated with exposures (from replicated genome-wide association studies, $P < 5 \times 10^{-8}$) as instrumental variables (Davey Smith & Ebrahim, 2003)



Copenhagen General Population study (N=55,523)

Multiplicative SMM

- Notation: X exposure/treatment, Y outcome, Z instrument, $Y\{X=0\}$ exposure/treatment free potential outcome

Robins, Rotnitzky, & Scharfstein, 1999; Hernán & Robins, 2006

$$\begin{aligned}
\log(E[Y|X,Z]) - \log(E[Y\{0\}]) &= (\psi + \psi_1 Z)X \\
\text{Identification NEM by } Z\text{: } \psi_1 = 0 & \\
&= \psi X \\
\frac{E[Y|X,Z]}{E[Y\{0\}|X,Z]} &= \exp(\psi X) \\
\psi : \text{ log causal risk ratio} & \\
\text{Rearrange: } Y\{0\} &= Y\exp(-\psi X)
\end{aligned}$$

MSMM G-estimation

Under the instrumental variable assumptions (Robins, 1989):

$$\begin{aligned} Y\{0\} &\perp\!\!\!\perp Z \\ Y\exp(-\psi X) &\perp\!\!\!\perp Z \\ Y\exp(-\psi X) - Y\{0\} &\perp\!\!\!\perp Z \end{aligned}$$

MSMM G-estimation

Under the instrumental variable assumptions (Robins, 1989):

$$Y\{0\} \perp\!\!\!\perp Z$$
$$Y \exp(-\psi X) \perp\!\!\!\perp Z$$
$$Y \exp(-\psi X) - Y\{0\} \perp\!\!\!\perp Z$$

Moment conditions

Z=0,1

$$E[(Y \exp(-\psi X) - Y\{0\})1] = 0$$
$$E[(Y \exp(-\psi X) - Y\{0\})Z_1] = 0$$

MSMM G-estimation

Under the instrumental variable assumptions (Robins, 1989):

$$Y\{0\} \perp\!\!\!\perp Z$$
$$Y \exp(-\psi X) \perp\!\!\!\perp Z$$
$$Y \exp(-\psi X) - Y\{0\} \perp\!\!\!\perp Z$$

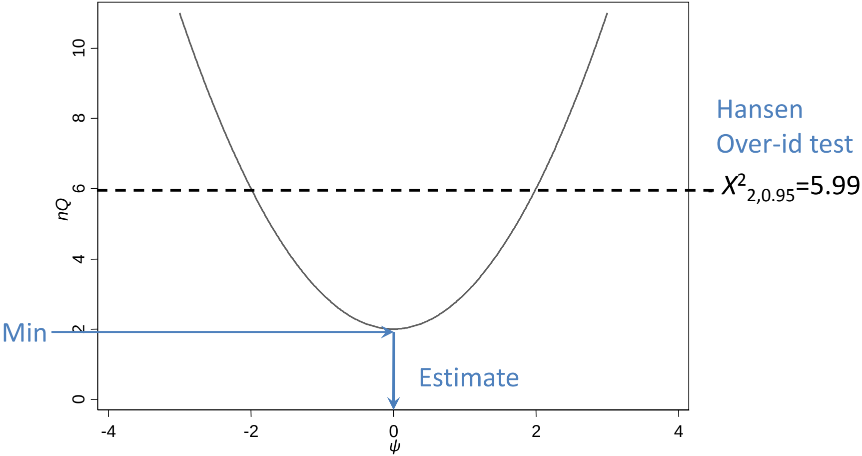
Moment conditions

Z=0,1,2,3 Over-identified

$$E[(Y \exp(-\psi X) - Y\{0\})1] = 0$$
$$E[(Y \exp(-\psi X) - Y\{0\})Z_1] = 0$$
$$E[(Y \exp(-\psi X) - Y\{0\})Z_2] = 0$$
$$E[(Y \exp(-\psi X) - Y\{0\})Z_3] = 0$$

What is GMM?

Minimises quadratic form: $Q = m'W^{-1}m$



Implementation

Stata: `gmm` command

```
gmm (y*exp(-1*x*{psi}) - {ey0}), instruments(z1 z2 z3)
```

Implementation

Stata: gmm command

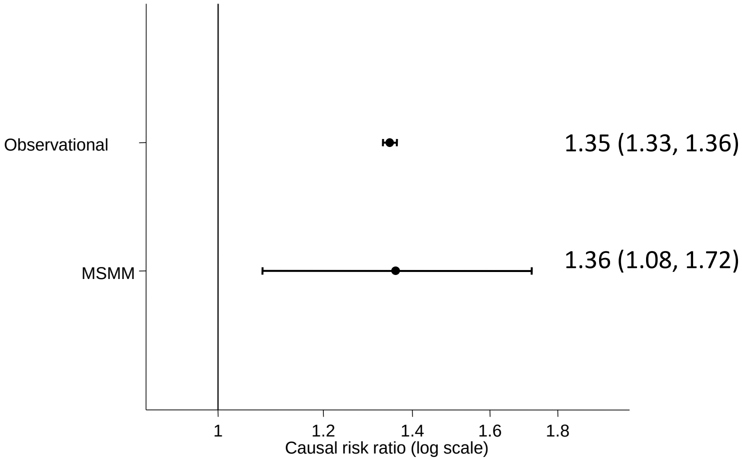
```
gmm (y*exp(-1*x*{psi}) - {ey0}), instruments(z1 z2 z3)
```

R: gmm package (Chaussé, 2010)

```
library(gmm)
msmmMoments <- function(theta,x){
  # extract variables from x
  Y <- x[,1]; X <- x[,2]; Z1 <- x[,3]; Z2 <- x[,4]; Z3 <- x[,5]
  # moments
  m1 <- (Y*exp(- X*theta[2]) - theta[1])
  m2 <- (Y*exp(- X*theta[2]) - theta[1])*Z1
  m3 <- (Y*exp(- X*theta[2]) - theta[1])*Z2
  m4 <- (Y*exp(- X*theta[2]) - theta[1])*Z3
  return(cbind(m1,m2,m3,m4))
}
fit <- gmm(msmmMoments, data, t0=c(0,0))
```

Specifying analytic first derivatives improves runtimes

MSMM example estimates



MSMM: Hansen over-identification test $P = 0.31$

How does GMM deal with multiple instruments?

GMM estimator solution to:

$$\frac{\partial m'(\psi)}{\partial \psi} W^{-1} m(\psi) = 0$$

- MSMM: instruments combined into linear projection of $YX \exp(-X\psi)$ on $Z = (1, Z_1, Z_2)'$ (Bowden & Vansteelandt, 2010)
- LSMM: GMM also equivalent to their optimal instruments approach

(double) Logistic SMM

$\mathrm{logit}(p) = \log(p/(1-p)), \mathrm{expit}(x) = e^x/(1+e^x)$

Goetghebeur, 2010

$$\mathrm{logit}(E[Y|X,Z]) - \mathrm{logit}(E[Y\{0\}]) = \psi X$$

ψ : log causal odds ratio

Rearrange for $Y\{0\} = \mathrm{expit}(\mathrm{logit}(Y) - \psi X)$

(double) Logistic SMM

$\text{logit}(p) = \log(p/(1-p)), \text{expit}(x) = e^x/(1+e^x)$

Goetghebeur, 2010

$$\text{logit}(E[Y|X,Z]) - \text{logit}(E[Y\{0\}]) = \psi X$$

ψ : log causal odds ratio

Rearrange for $Y\{0\} = \text{expit}(\text{logit}(Y) - \psi X)$

- Can't be estimated in a single step (Robins et al., 1999)
- First stage association model (Vansteelandt & Goetghebeur, 2003):
  (i) logistic regression of Y on X & Z & interactions
  (ii) predict Y, estimate LSMM using predicted Y

(double) Logistic SMM moment conditions

Association model moment conditions

Logistic regression using GMM

$$
\begin{aligned}
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))1] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))X] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))Z] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))XZ] &= 0
\end{aligned}
$$

(double) Logistic SMM moment conditions

Association model moment conditions

Logistic regression using GMM

$$
\begin{aligned}
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))1] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))X] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))Z] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))XZ] &= 0
\end{aligned}
$$

Causal model moment conditions

$$
\begin{aligned}
E[(\text{expit}(\text{logit}(\widehat{p}) - \psi X) - Y\{0\})1] &= 0 \\
E[(\text{expit}(\text{logit}(\widehat{p}) - \psi X) - Y\{0\})Z] &= 0
\end{aligned}
$$

Problem: SEs incorrect - need association model uncertainty

LSMM joint estimation

Joint estimation = correct SEs (Gourieroux, Monfort, & Renault, 1996)

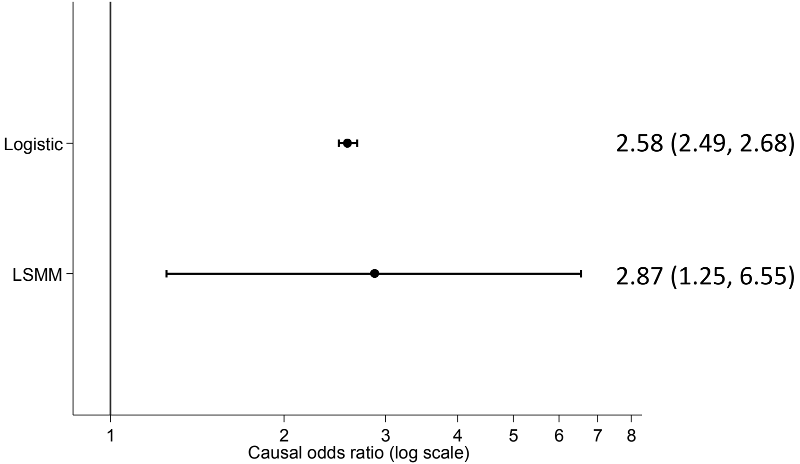
Vansteelandt & Goetghebeur, 2003; Bowden & Vansteelandt, 2010

$$
\begin{aligned}
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))1] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))X] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))Z] &= 0 \\
E[(Y - \text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ))XZ] &= 0 \\
E[(\text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ - \psi X) - Y\{0\})1] &= 0 \\
E[(\text{expit}(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ - \psi X) - Y\{0\})Z] &= 0
\end{aligned}
$$

Stata `gmm` command - allows multiple equations - still 1 line of code

Example: causal model SEs $\times 10$

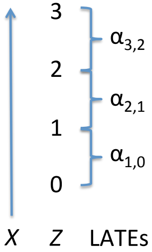
LSMM example estimates



LSMM: Hansen over-identification test $P = 0.29$

Local risk ratios for MSMM

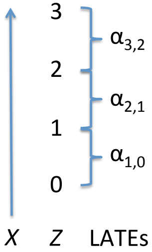
- Identification depends on NEM by Z ... what if it doesn't hold?
- Alternative assumption of monotonicity: $X(Z_k) \geq X(Z_{k-1})$
- Local Average Treatment Effect (LATE) (Imbens & Angrist, 1994)
  - effect among those whose exposures are changed (upwardly) by changing (counterfactually) the IV from Z_{k-1} to Z_k



$$\alpha_{All} = \lambda_1\alpha_{1,0} + \lambda_2\alpha_{2,1} + \lambda_3\alpha_{3,2}$$

Local risk ratios for MSMM

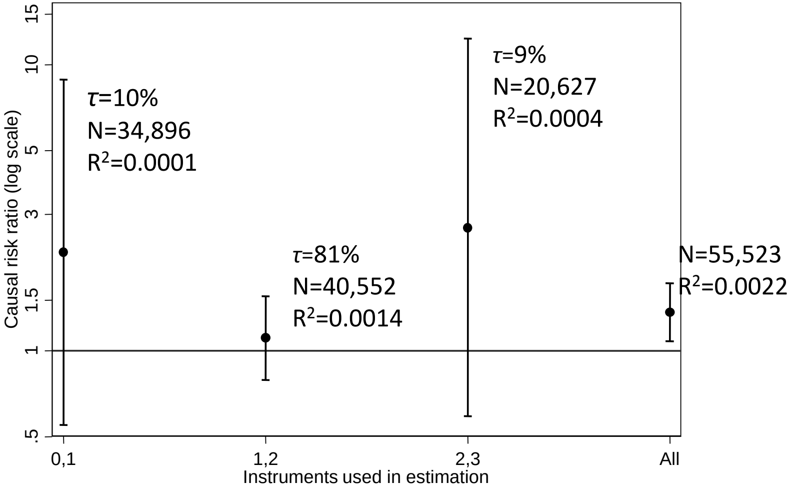
- Identification depends on NEM by Z ... what if it doesn't hold?
- Alternative assumption of monotonicity: $X(Z_k) \geq X(Z_{k-1})$
- Local Average Treatment Effect (LATE) (Imbens & Angrist, 1994)
  - effect among those whose exposures are changed (upwardly) by changing (counterfactually) the IV from Z_{k-1} to Z_k



$$\alpha_{All} = \lambda_1\alpha_{1,0} + \lambda_2\alpha_{2,1} + \lambda_3\alpha_{3,2}$$

Similar result holds for MSMM: $e^{\psi}_{\text{All}} = \sum_{k=1}^{K} \tau_k e^{\psi}_{k,k-1}$

Local risk ratios in example



Check: $(0.10 \times 2.21) + (0.81 \times 1.11) + (0.09 \times 2.69) = 1.36$

Summary

- Structural Mean Models estimated using IVs by G-estimation
  $$Y\{0\} \perp\!\!\!\perp Z$$
- GMM estimation approach:
  - Estimate $Y\{0\}$
  - Hansen over-id test of joint validity of instruments
  - Optimal combination of multiple instruments
  - Two-step GMM gives efficient SEs
  - LSMM: joint estimation approach
  - Straightforward implementation in Stata and R
  - Discussion by Tan, 2010
- SMMs: subtly different to additive residual IV estimators
  - RR: $Y - \exp(\psi X) \perp\!\!\!\perp Z$ (Cameron & Trivedi, 2009; Johnston, Gustafson, Levy, & Grootendorst, 2008)
  - OR: $Y - \text{expit}(\psi X) \perp\!\!\!\perp Z$ (Foster, 1997; Rassen, Schneeweiss, Glynn, Mittleman, & Brookhart, 2009)
- Review of some of the methods (Palmer et al., 2011)

Acknowledgements

- MRC Collaborative grant G0601625
- MRC CAiTE Centre grant G0600705
- ESRC grant RES-060-23-0011
- With thanks to Nuala Sheehan, Vanessa Didelez, Debbie Lawlor, Jonathan Sterne, George Davey Smith, Sha Meng, Neil Davies, Roger Harbord, Nic Timpson, Borge Nordestgaard.

References I

Bowden, J., & Vansteelandt, S. (2010). Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*. (in press)

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics Using Stata*. College Station, Texas: Stata Press.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*(3), 305–334.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software*, *34*(11), 1–35. Available from `http://www.jstatsoft.org/v34/i11/`

Clarke, P. S., & Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, *11*(4), 756–770.

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.

Drukker, D. (2010). An introduction to GMM estimation using Stata. In *German stata users group meeting.* Berlin.

Foster, E. M. (1997). Instrumental variables for logistic regression: an illustration. *Social Science Research*, *26*, 487–504.

Goetghebeur, E. (2010). Commentary: To cause or not to cause confusion vs transparency with Mendelian Randomization. *International Journal of Epidemiology*, *39*(3), 918–920.

References II

Gourieroux, C., Monfort, A., & Renault, E. (1996). Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference*, *50*(1), 37–63.

Hernán, M. A., & Robins, J. M. (2006). Instruments for Causal Inference. An Epidemiologist's Dream? *Epidemiology*, *17*, 360–372.

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

Johnston, K. M., Gustafson, P., Levy, A. R., & Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, *27*, 1539–1556.

Mullahy, J. (1997). Instrumental–variable estimation of count data models: Applications to models of cigarette smoking behaviour. *The Review of Economics and Statistics*, *79*(4), 568–593.

Nichols, A. (2007). *ivpois: Stata module for IV/GMM Poisson regression.* Statistical Software Components, Boston College Department of Economics. (available at http://ideas.repec.org/c/boc/bocode/s456890.html)

Palmer, T. M., Sterne, J. A. C., Harbord, R. M., Lawlor, D. A., Sheehan, N. A., Meng, S., et al. (2011). Instrumental variable estimation of causal risk ratios and causal odds ratios in mendelian randomization analyses. *American Journal of Epidemiology*, *173*, 1392–1403.

References III

Rassen, J. A., Schneeweiss, S., Glynn, R. J., Mittleman, M. A., & Brookhart, M. A. (2009). Instrumental Variable Analysis for Estimation of Treatment Effects With Dichotomous Outcomes. *American Journal of Epidemiology*, *169*(3), 273–284.

Robins, J. M. (1989). Health services research methodology: A focus on aids. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), (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.

Robins, J. M., Rotnitzky, A., & Scharfstein, D. O. (1999). Statistical models in epidemiology: The environment and clinical trials. In M. E. Halloran & D. Berry (Eds.), (pp. 1–92). New York, US: Springer.

Tan, Z. (2010). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association*, *105*(489), 157–169.

Vansteelandt, S., & Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B*, *65*(4), 817–835.

Windmeijer, F. (2002). *ExpEnd, A Gauss program for non-linear GMM estimation of exponential models with endogenous regressors for cross section and panel data* (Tech. Rep.). Centre for Microdata Methods and Practice.

References IV

Windmeijer, F. (2006). *GMM for panel count data models* (Bristol Economics Discussion Papers No. 06/591). Department of Economics, University of Bristol, UK. Available from http://ideas.repec.org/p/bri/uobdis/06-591.html

Windmeijer, F., & Santos Silva, J. (1997). Endogeneity in Count Data Models: An Application to Demand for Health Care. *Journal of Applied Econometrics*, *12*(3), 281–294.

Two-step GMM

1. Minimize quadratic form: $m' W^{-1} m$
2. Estimate $\widehat{W_1}$, minimize quadratic form starting from $\widehat{W_1}$

- Two-step GMM gives efficient SEs (Chamberlain, 1987)
- Stata Hansen test command (`estat overid`) requires this

MSMM alternative parameterisation

$$Y \exp(-X\psi - \log(Y\{0\})) - 1 = 0$$

- Same as moments used by Mullahy, 1997; Nichols, 2007
- First parameterisation more numerically stable (Drukker, 2010)
- Also see Windmeijer & Santos Silva, 1997; Windmeijer, 2002, 2006; Clarke & Windmeijer, 2010
- Use X as instrument for itself = Gamma regression (log link)

Example estimates

	RR (95% CI)	P over-id
MSMM	1.36 (1.08, 1.72)	0.31
$Y - \exp(\psi X) \perp\!\!\!\perp Z$	1.36 (1.07, 1.75)	0.30
Control function	1.36 (1.08, 1.71)	
	OR (95% CI)	**P over-id**
LSMM two-stage	1.88 (1.75, 2.02)	
LSMM joint	2.87 (1.25, 6.55)	0.29
$Y - \text{expit}(\psi X) \perp\!\!\!\perp Z$	2.69 (1.23, 5.90)	0.30
Control function	2.69 (1.21, 5.97)	