Generalised method of moments estimation of structural mean models

Tom Palmer[1,2] Roger Harbord[2] Paul Clarke[3]
Frank Windmeijer[3,4,5]

1. MRC Centre for Causal Analyses in Translational Epidemiology
2. School of Social and Community Medicine, University of Bristol
3. CMPO, University of Bristol
4. Department of Economics, University of Bristol, UK
5. CEMMAP/IFS, London

15 September 2011

MRC | Centre for Causal Analyses in Translational Epidemiology

University of BRISTOL

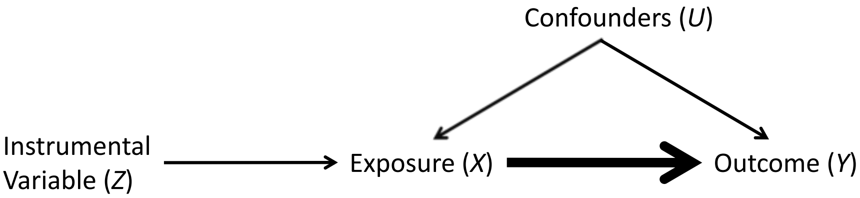
Outline

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

- Introduction to Mendelian randomization example
- Multiplicative structural mean model (MSMM)
  - G-estimation, identification, `gmm` syntax, example
- (double) Logistic SMM
  - `gmm` multiple equation syntax, example
- Summary
- MSMM: local risk ratios

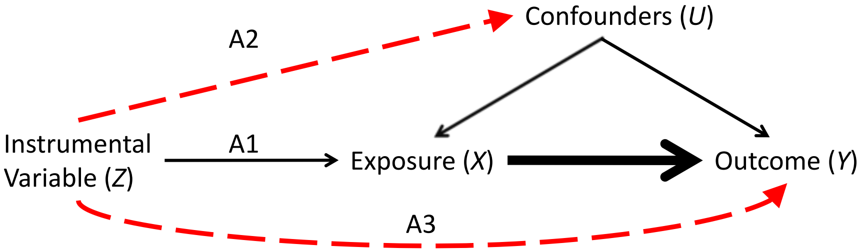
Introduction to Mendelian randomization example

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



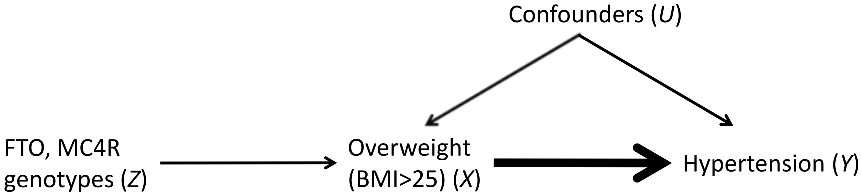
Introduction to Mendelian randomization example

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



Introduction to Mendelian randomization example

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



Copenhagen General Population study (N=55,523)

Multiplicative SMM

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

Robins, 1989, 1994; Robins, Rotnitzky, & Scharfstein, 1999; Hernán & Robins, 2006

$$\log(E[Y|X,Z]) - \log(E[Y\{0\}|X,Z]) = \psi X$$
$$\frac{E[Y|X,Z]}{E[Y\{0\}|X,Z]} = \exp(\psi X)$$

ψ : log causal risk ratio

Rearrange: $Y\{0\} = Y\exp(-\psi X)$

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):

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

MSMM `gmm` syntax

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

MSMM gmm output

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

Final GMM criterion Q(b) =  .0000425

GMM estimation

Number of parameters =   2
Number of moments    =   4
Initial weight matrix: Unadjusted                    Number of obs   =    55523
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |   .3104495   .1192332     2.60   0.009     .0767568    .5441423
        /ey0 |   .5758842   .0388716    14.82   0.000     .4996973    .6520711
------------------------------------------------------------------------------
Instruments for equation 1: z1 z2 z3 _cons
```

MSMM gmm output

Causal risk ratio exp(ψ) & Hansen over-id test

```
. lincom [psi]:_cons, eform

 ( 1)  [psi]_cons = 0

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.364038   .1626386     2.60   0.009     1.079779     1.72313
------------------------------------------------------------------------------

. estat overid

  Test of overidentifying restriction:

  Hansen's J chi2(2) = 2.36125 (p = 0.3071)
```

MSMM gmm syntax including analytic first derivatives

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

Reduces runtime from 4.5 secs to 2.5 secs on 55000 obs

MSMM alternative parameterisation

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

- Same moment condition in `ivpois` (Mullahy, 1997; Nichols, 2007)
- Drukker, 2010: first syntax more numerically stable
- Also see Windmeijer & Santos Silva, 1997; Windmeijer, 2002, 2006; Clarke & Windmeijer, 2010
- Use X as instrument for itself $\equiv$ Gamma regression (log link)
- Slightly different to Poisson regression moment condition:

$$Y - \exp(X\beta) \perp\!\!\!\perp Z$$

MSMM 2nd syntax & ivpois output

```
. gmm (y*exp(-x*{psi} - {logey0}) - 1), instruments(z1 z2 z3) onestep nolog

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |    .290323   .1184236     2.45   0.014      .058217    .5224291
     /logey0 |  -.5404186   .0676225    -7.99   0.000    -.6729562   -.4078811
------------------------------------------------------------------------------

. ivpois y, endog(x) exog(z1 z2 z3)
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
           x |   .2903902   .1184242     2.45   0.014      .058283    .5224973
       _cons |   -.540463   .0676208    -7.99   0.000    -.6729974   -.4079286
------------------------------------------------------------------------------
```

MSMM 'endogenous' & Gamma (log link) output

```
. gmm (y*exp(-1*x*{psi} - {logey0}) - 1), instruments(x) onestep nolog

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |   .2974176   .0062505    47.58   0.000     .2851668    .3096684
     /logey0 |  -.5444755   .0054942   -99.10   0.000    -.5552439   -.5337072
------------------------------------------------------------------------------

. glm y x, family(gamma) link(log) robust nolog

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2974176   .0062506    47.58   0.000     .2851667    .3096685
       _cons |  -.5444755   .0054942   -99.10   0.000     -.555244   -.5337071
------------------------------------------------------------------------------
```

(double) Logistic SMM

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

Goetghebeur, 2010

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

ψ : log causal odds ratio

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

(double) Logistic SMM

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

Goetghebeur, 2010

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

ψ : log causal odds ratio

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

- LSMM can't be estimated in a single step (Robins et al., 1999)
- LSMM estimator with first stage association model (Vansteelandt & Goetghebeur, 2003; Bowden & Vansteelandt, 2010):
  - logistic regression of Y on X & Z (& interactions: saturated)
  - predict Y
  - estimate LSMM using predicted Y

(double) LSMM gmm syntax

$$\text{invlogit}(x) = \text{expit}(x) = e^x/(1 + e^x)$$

Association model gmm syntax - logistic regression using GMM

```
gmm (y - invlogit({b0} + {xb:x z1 z2 z3 xz1 xz2 xz3})), ///
    instruments(x z1 z2 z3 xz1 xz2 xz3)
predict prres
gen xblog = logit(y - prres)
```

(double) LSMM gmm syntax

$$\text{invlogit}(x) = \text{expit}(x) = e^x/(1+e^x)$$

Association model gmm syntax - logistic regression using GMM

```
gmm (y - invlogit({b0} + {xb:x z1 z2 z3 xz1 xz2 xz3})), ///
    instruments(x z1 z2 z3 xz1 xz2 xz3)
predict prres
gen xblog = logit(y - prres)
```

Causal model gmm syntax

```
gmm (invlogit(xblog - x*{psi}) - {ey0}), instruments(z1 z2 z3)
```

Problem: causal model SEs incorrect - need to incorporate uncertainty from association model

Association model output: gmm & logit

```
. gmm (y - invlogit({xb:x z1 z2 z3 xz1 xz2 xz3} + {b0})), instruments(x z1 z2 z3 xz1 xz2 xz3)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       /xb_x |   .9034697   .0419769    21.52   0.000     .8211965    .9857428
      /xb_z1 |   .0023852   .0346439     0.07   0.945    -.0655155     .070286
      /xb_z2 |   -.031613   .0375747    -0.84   0.400     -.105258     .042032
      /xb_z3 |   .0285799   .0598671     0.48   0.633    -.0887574    .1459173
     /xb_xz1 |   .0500118   .0509504     0.98   0.326    -.0498492    .1498728
     /xb_xz2 |     .06952   .0543206     1.28   0.201    -.0369464    .1759864
     /xb_xz3 |   .0412161   .0837708     0.49   0.623    -.1229716    .2054038
         /b0 |   .3295621   .0285043    11.56   0.000     .2736947    .3854295
------------------------------------------------------------------------------

. logit y x z1 z2 z3 xz1 xz2 xz3, nolog
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .9034696   .0419769    21.52   0.000     .8211964    .9857428
          z1 |   .0023852   .0346439     0.07   0.945    -.0655155     .070286
          z2 |   -.031613   .0375747    -0.84   0.400     -.105258     .042032
          z3 |   .0285799   .0598671     0.48   0.633    -.0887574    .1459173
         xz1 |   .0500117   .0509504     0.98   0.326    -.0498493    .1498727
         xz2 |     .06952   .0543206     1.28   0.201    -.0369465    .1759864
         xz3 |    .041216   .0837708     0.49   0.623    -.1229717    .2054037
       _cons |   .3295621   .0285043    11.56   0.000     .2736947    .3854295
------------------------------------------------------------------------------
. matrix from = e(b)
. predict xblog, xb
```

Causal model output

```
. gmm (invlogit(xblog - x*{psi}) - {ey0}), instruments(z1 z2 z3) nolog

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |   .6331413   .0362588    17.46   0.000     .5620754    .7042073
        /ey0 |   .6226167    .004652   133.84   0.000      .613499    .6317344
------------------------------------------------------------------------------
Instruments for equation 1: z1 z2 z3 _cons
. matrix from = (from,e(b))
```

Problem: causal model SEs incorrect - need to incorporate uncertainty from association model

LSMM joint estimation

Joint estimation of association and causal models = correct SEs

(Gourieroux, Monfort, & Renault, 1996)

LSMM gmm multiple equation syntax

```
gmm (y - invlogit({xb:x z1 z2 z3 xz1 xz2 xz3} + {b0})) ///
    (invlogit({xb:} + {b0} - x*{psi}) - {ey0}), ///
    instruments(1:x z1 z2 z3 xz1 xz2 xz3) ///
    instruments(2:z1 z2 z3) ///
    winitial(unadjusted, independent) ///
    from(from)
```

LSMM gmm multiple equation output

```
Number of parameters =   10
Number of moments    =   12
Initial weight matrix: Unadjusted                 Number of obs   =    55523
GMM weight matrix:     Robust
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       /xb_x |   .9091545   .0418464    21.73   0.000     .8271371    .9911719
      /xb_z1 |  -.0207159   .0279367    -0.74   0.458    -.0754708     .034039
      /xb_z2 |  -.0339566   .0343049    -0.99   0.322     -.101193    .0332797
      /xb_z3 |  -.0058356   .0550491    -0.11   0.916    -.1137299    .1020586
     /xb_xz1 |    .039923   .0502901     0.79   0.427    -.0586438    .1384898
     /xb_xz2 |   .0687247   .0542023     1.27   0.205    -.0375099    .1749592
     /xb_xz3 |   .0262868   .0826922     0.32   0.751     -.135787    .1883605
         /b0 |   .3425951   .0253272    13.53   0.000     .2929547    .3922354
        /psi |    1.05276   .4217043     2.50   0.013     .2262351    1.879286
        /ey0 |   .5656666   .0592065     9.55   0.000     .4496241    .6817091
------------------------------------------------------------------------------
```

Causal model SEs $\times 10$!

LSMM gmm multiple equation output

Causal odds ratio $\exp(\psi)$ & Hansen over-id test

```
. lincom [psi]:_cons, eform

 ( 1)  [psi]_cons = 0

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    2.86555   1.208415     2.50   0.013      1.25387    6.548825
------------------------------------------------------------------------------

. estat overid

  Test of overidentifying restriction:

  Hansen's J chi2(2) =   2.459 (p = 0.2924)
```

LSMM gmm multiple equation syntax with derivatives

```
local p1 "invlogit({xb:} + {b0})"
local d1 "-1*`p1'*(1 - `p1')"
local p2 "invlogit({xb:} + {b0} - x*{psi})"
local d2 "`p2'*(1 - `p2')"
gmm (y - invlogit({xb:x z1 z2 z3 xz1 xz2 xz3} + {b0})) ///
    (invlogit({xb:} + {b0} - x*{psi}) - {ey0}), ///
    instruments(1:x z1 z2 z3 xz1 xz2 xz3) ///
    instruments(2:z1 z2 z3) ///
    winitial(unadjusted, independent) from(from) ///
    deriv(1/xb = `d1') ///
    deriv(1/b0 = `d1') ///
    deriv(2/xb = `d2') ///
    deriv(2/b0 = `d2') ///
    deriv(2/psi = -1*x*`d2') ///
    deriv(2/ey0 = -1)
```

Stata applies last step of chain rule to derivates of {xb:} i.e. $\frac{\partial u}{\partial \beta_j} = \frac{\partial u}{\partial (\mathbf{x}'\beta)} \times \frac{\partial (\mathbf{x}'\beta)}{\partial \beta_j}$

See `help gmm` & manual P593–5

Reduces runtime from 155secs to 32secs on 55000 obs

Summary

- Structural Mean Models estimated using IVs by G-estimation

$$Y\{0\} \perp\!\!\!\perp Z$$

- GMM estimation using multiple instruments
- Multiplicative SMM = `ivpois`
- Specifying analytic derivatives in `gmm` = faster!
- (double) logistic SMM estimation using multiple equations
- `estat overid`: Hansen J-test of joint validity of instruments
- 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, 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.

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.

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.

References II

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*.

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.

References III

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. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics: Theory and Methods*, *23*(8), 2379–2412.

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.

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.

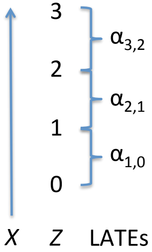
Windmeijer, F. (2006). *GMM for panel count data models* (Bristol Economics Discussion Papers No. 06/591). Department of Economics, University of Bristol, UK. Available from `http://ideas.repec.org/p/bri/uobdis/06-591.html`

References IV

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.

Local risk ratios for MSMM

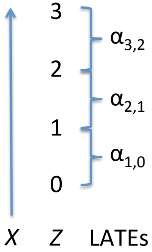
- Identification depends on NEM ... what if it doesn't hold?
- Alternative assumption of monotonicity: $X(Z_k) \geq X(Z_{k-1})$
- Local Average Treatment Effect (LATE) (Imbens & Angrist, 1994)
  - effect among those whose exposures are changed (upwardly) by changing (counterfactually) the IV from Z_{k-1}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

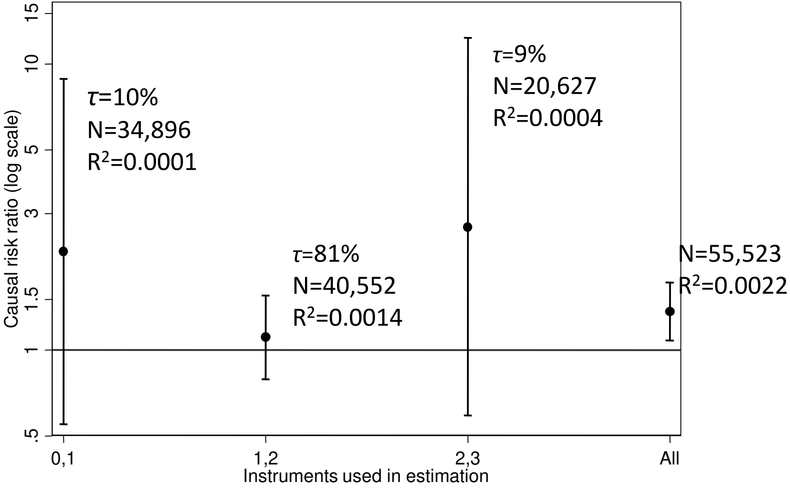
- Identification depends on NEM ... what if it doesn't hold?
- Alternative assumption of monotonicity: $X(Z_k) \geq X(Z_{k-1})$
- Local Average Treatment Effect (LATE) (Imbens & Angrist, 1994)
  - effect among those whose exposures are changed (upwardly) by changing (counterfactually) the IV from Z_{k-1}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

Local risk ratios in example



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

Compare SMMs with other estimators

	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
(double) LSMM	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)	