Estimation of structural mean models with multiple instruments

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

[1]MRC CAiTE Centre, School of Social and Community Medicine, University of Bristol

[2]Department of Economics and CMPO, University of Bristol

Royal Statistical Society, 26 May 2011

Aim

Combine two strands of literature:

- Structural mean models [Biostatistics]
- Generalised Method of Moments estimation [Econometrics]

Rationale:

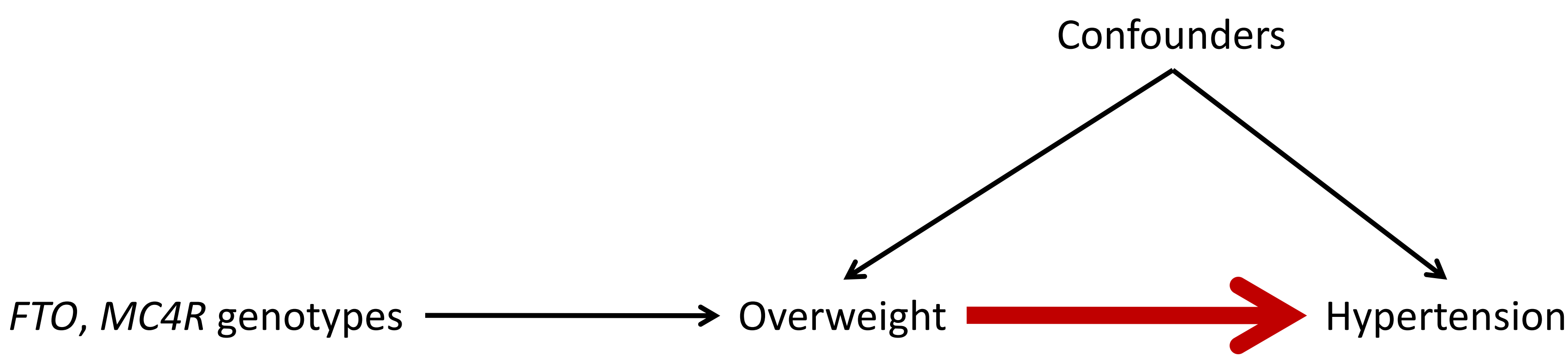
- Concepts such as G-estimation intimidating
- Estimation with multiple instruments
- Straightforward implementation in Stata and R

Outline

- Introduction to example
- Causal parameters & potential outcomes
- Multiplicative SMM
  - What is GMM?
  - Over-identification test
  - Combining multiple instruments
  - Two step GMM
  - Implementation in Stata
  - Local risk ratios
  - MSMM and MGMM
- Logistic SMM
  - Joint estimation
- Summary

Introduction to example

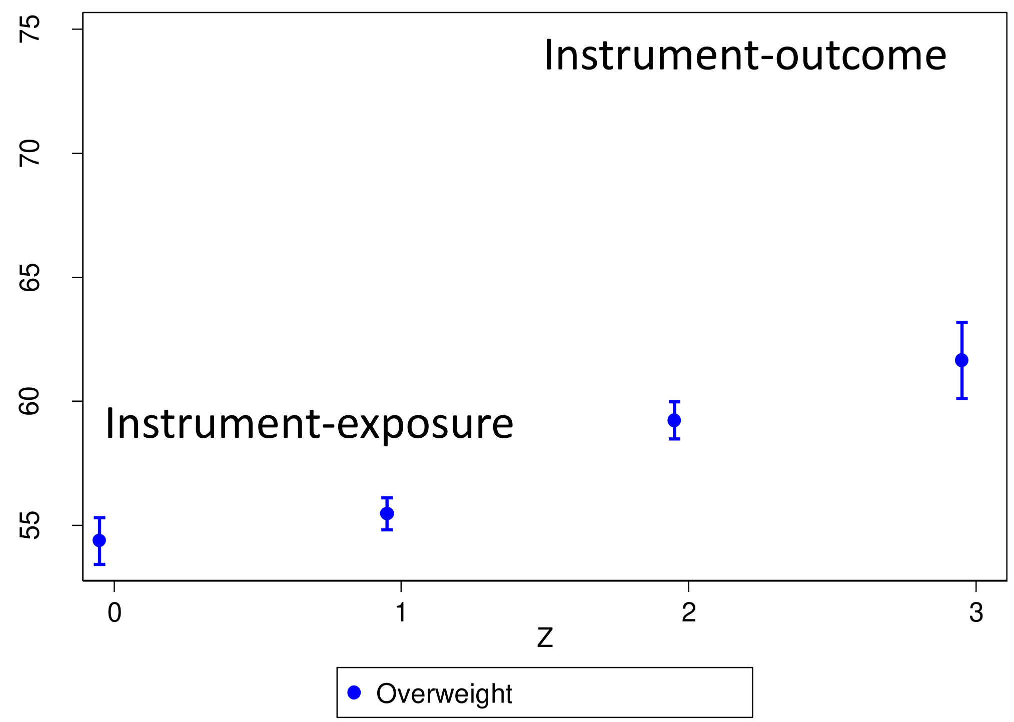
- Copenhagen General Population study
  - N=55,523
- Instruments:
  - *FTO* (rs9939609) chr16, *MC4R* (rs17782313) chr18 genotypes
  - Associated with obesity in GWAS (0.4, 0.2 BMI units). Frayling 2007, Loos 2008
- Exposure:
  - Overweight (body mass index BMI [weight/height^2] >25)
- Outcome:
  - Hypertension (high blood pressure [SBP>140mmHg, or DBP>90mmHg, or taking anti-hypertensives])



	No Hypertension	Hypertension	Total
Not Overweight	10,066 42%	13,909 58%	23,975
Overweight	6,906 22%	24,642 78%	31,548
Total	16,972 31%	38,551 69%	55,523 χ^2 P<0.001

Risk ratio 1.35 (1.32, 1.37)

FTO	*MC4R*	*Z*	Freq
0	0	0	0.20
0	1	1	0.15
1	0	1	0.27
1	1	2	0.21
2	0	2	0.09
2	1	3	0.07



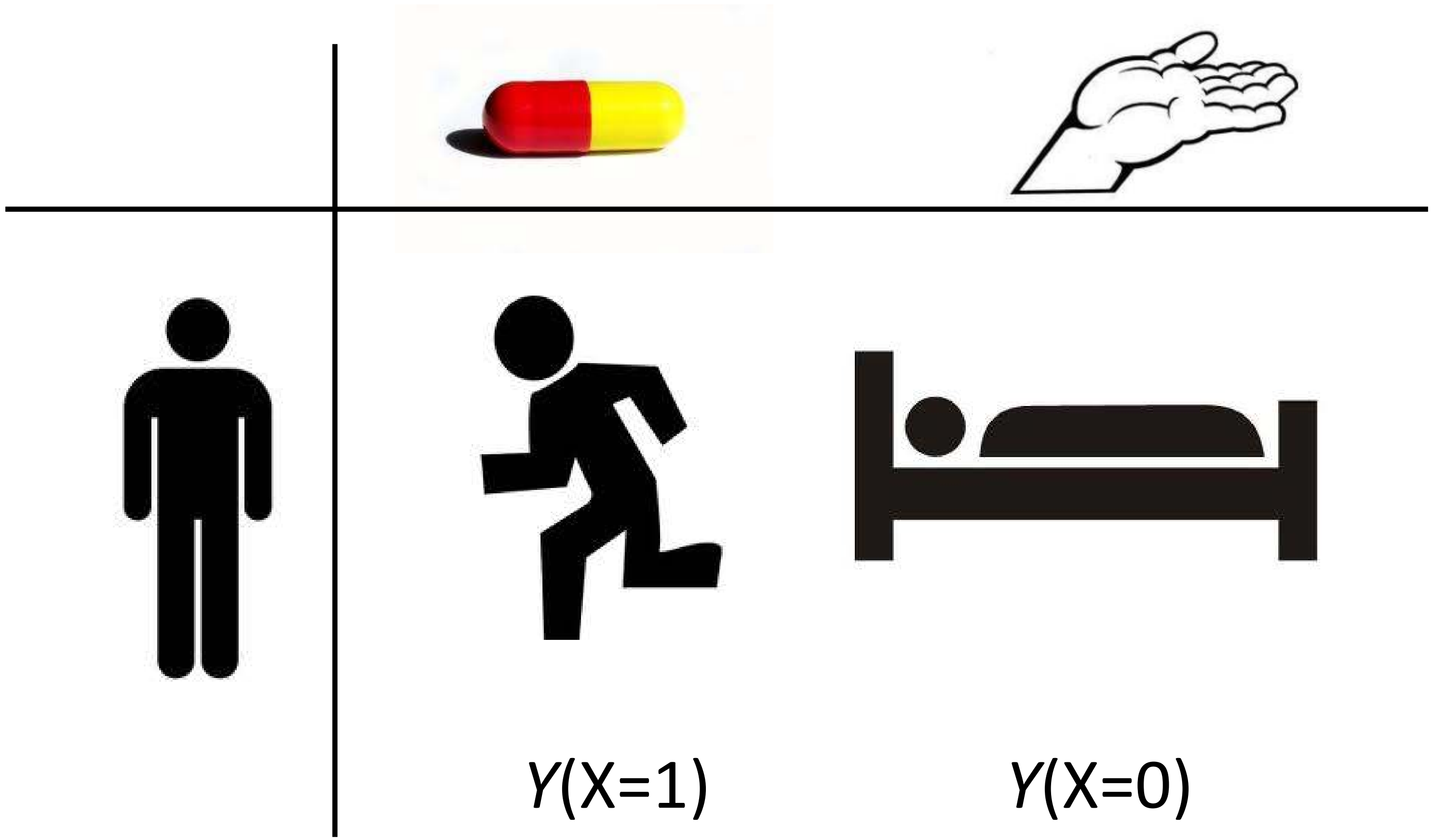
P=0.007

P<0.001
R^2=0.002

Causal parameters and potential outcomes

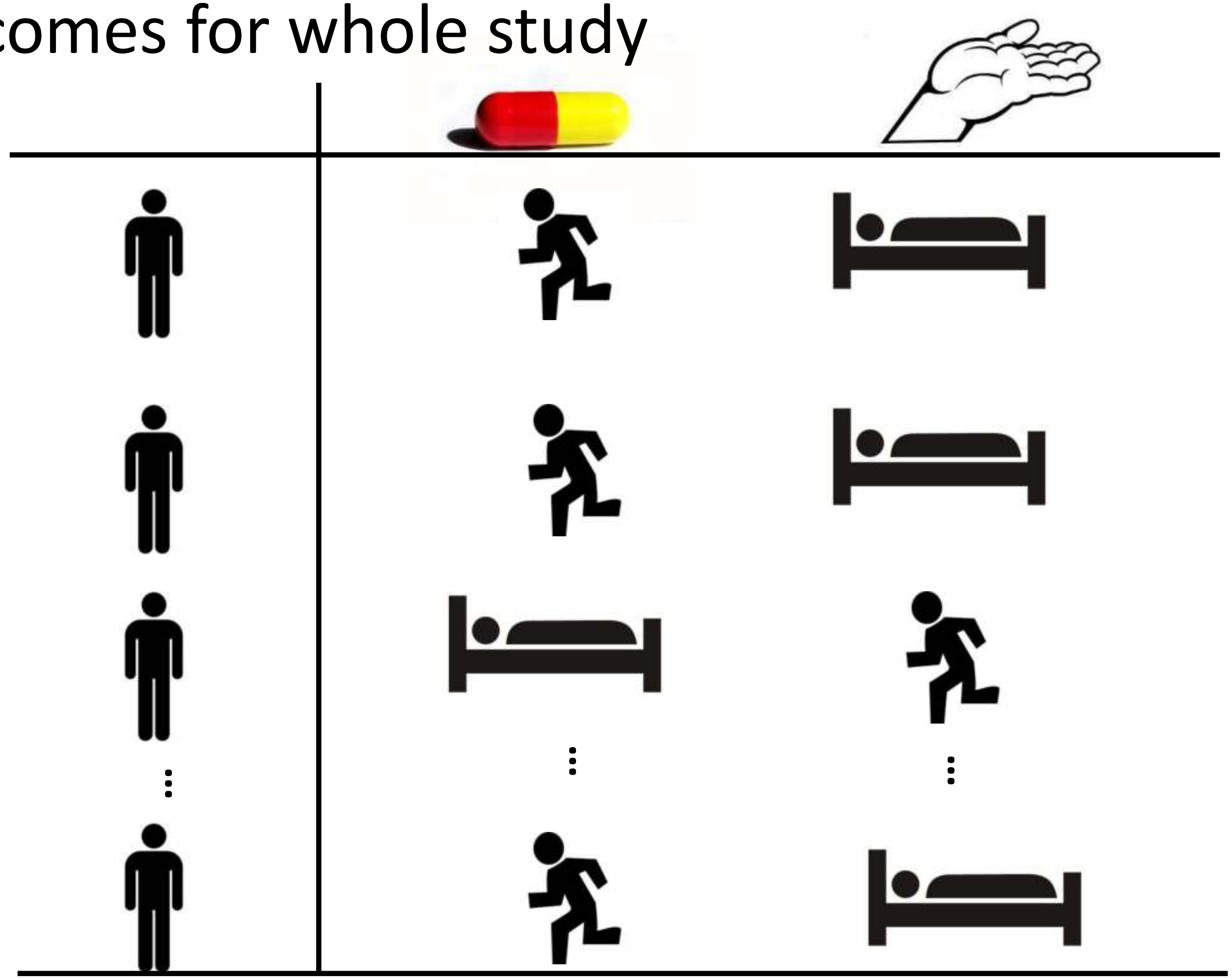
- SMMs defined in terms of potential outcomes Hernan & Robins 2006
- X: exposure/treatment, Y: outcome, Z: IV
- $Y(X=1)$ outcome subject would experience if they were given treatment/exposure under intervention

Potential outcomes for an individual



Potential outcomes for whole study

Recent discussion of G-estimation: Snowden et al., AJE, 2011



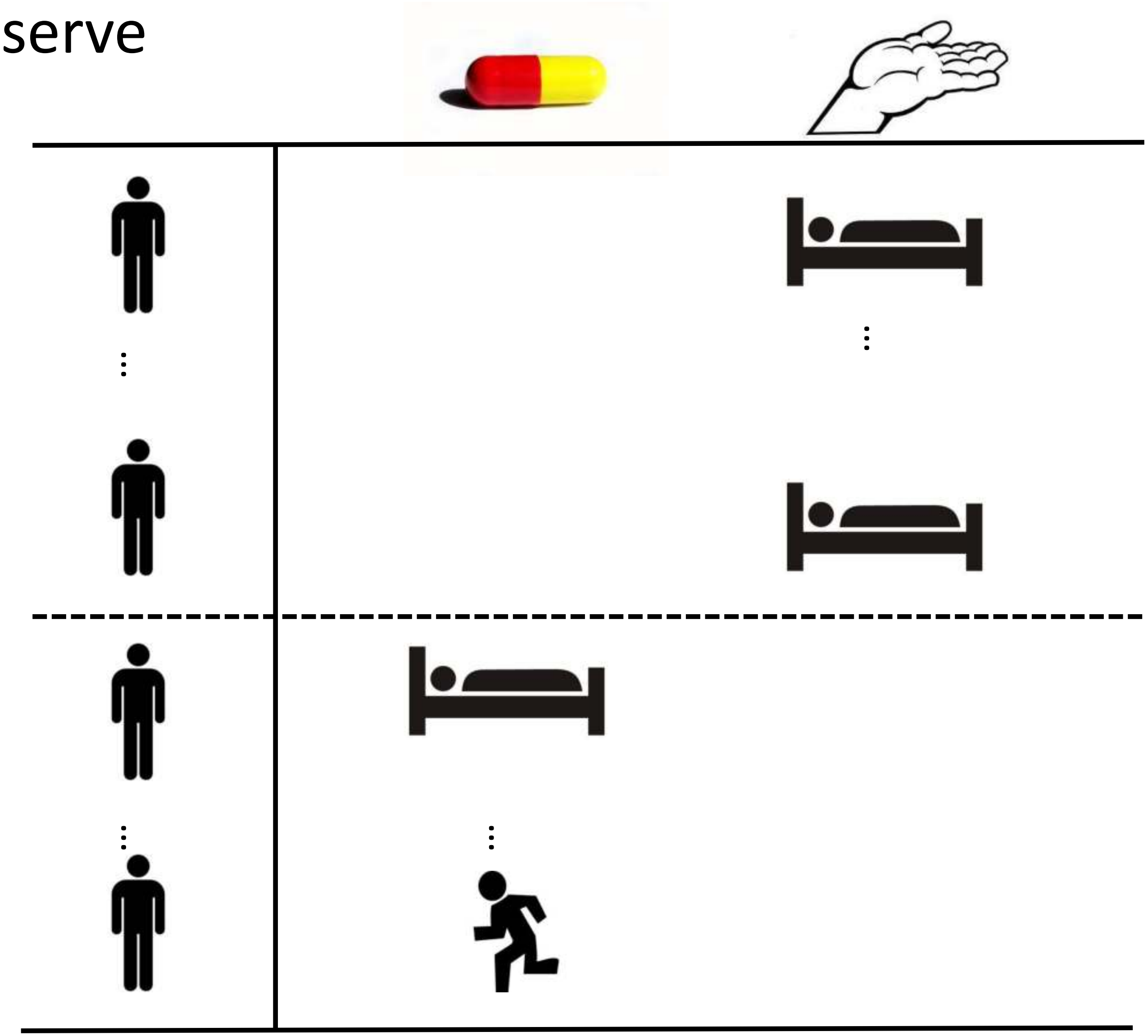
Average treatment effect = $E[Y(X=1)] - E[Y(X=0)]$

binary outcome: causal risk difference

Causal risk ratio = $E[Y(X=1)] \;/\; E[Y(X=0)]$

Causal odds ratio = $\text{odds}[Y(X=1)] \;/\; \text{odds}[Y(X=0)]$

What we observe



$$E[Y(1)|X=1] \quad - \quad E[Y(0)|X=1]$$

SMMs identify effect of treatment of treated

Multiplicative SMM

Z is instrumental variable X is exposure Y is outcome

Y, X and Z are binary

$$\frac{E[Y|X,Z]}{E[Y(0)|X,Z]} = \exp\{(\theta_0 + \theta_1 Z) X\}$$

$Y(0)$ is the exposure- or treatment-free potential outcome

…so far … model non-identified: 2 parameters, 1 equation

No effect modification by Z (NEM): $\theta_1 = 0$

θ_0: log causal risk ratio

Conditional mean independence (CMI) from IV assumptions:

$$E[Y(0)|Z=1] = E[Y(0)|Z=0] = E[Y(0)]$$

Moment conditions

$$\alpha_0 = E\left[Y(0)\right]$$

Multi-valued instrument/multiple instruments

$$\begin{aligned}
E\left[\{Y\exp(-X\theta_0)-\alpha_0\}\,|Z=2\right] &= 0\\
E\left[\{Y\exp(-X\theta_0)-\alpha_0\}\,|Z=1\right] &= 0\\
E\left[\{Y\exp(-X\theta_0)-\alpha_0\}\,|Z=0\right] &= 0
\end{aligned}$$

Over-identified:

3 moment conditions,

2 parameters …

… need GMM

Exactly identified:

2 moment conditions,

2 parameters

E[]=0 since *Z* independent of *Y* given *X*: exclusion restriction

If no E[*Y*(0)] – need to centre the instruments;

Vansteelandt & Goetghebeur, JRSS B, 2003

What is GMM?

Designed to estimate over-identified models

GMM minimises quadratic form wrt parameters to be estimated

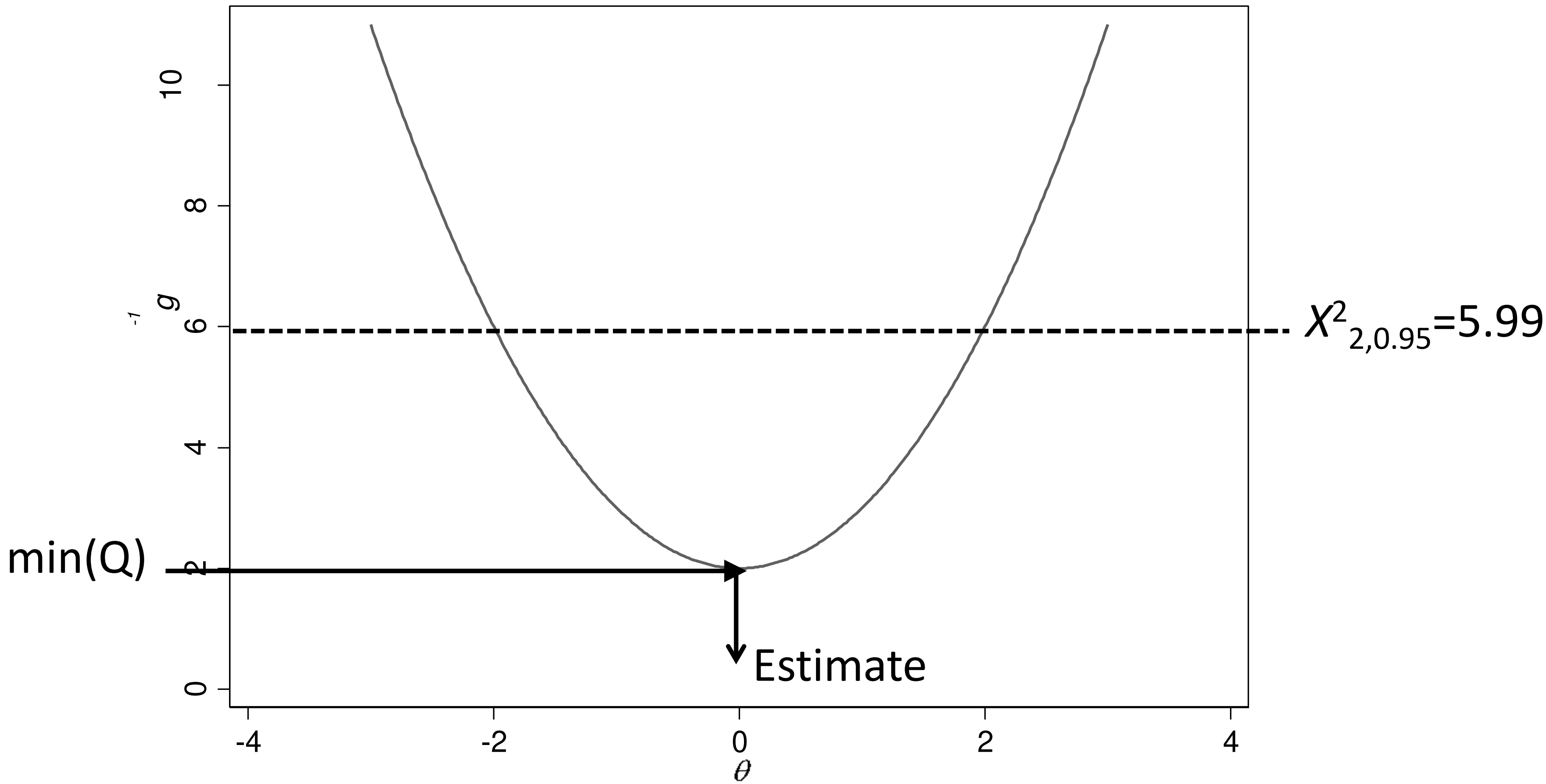
$$\widehat{\delta} = \arg\min_{\delta} \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)' W_n^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)$$

$$\begin{pmatrix} \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_0 & \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_1 & \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_2 \end{pmatrix} W_n^{-1} \begin{pmatrix} \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_0 \\ \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_1 \\ \{Y \exp(-X\theta_0) - \alpha_0\}\, Z_2 \end{pmatrix}$$

W^{-1} affects efficiency not consistency: one step/two step GMM

Over-identification test

Profiling over quadratic form (Q) for a single parameter



- Single instrument – exactly identified: min(Q)=0
- Multiple instruments – over identified: min(Q) should be close enough to 0 as given by Hansen over-id test statistic, ***Q ~ X^2_{m-p}*** when moments valid
- Not rejecting the over-id test ***doesn't*** mean the IV assumptions hold

Combining multiple instruments

How does GMM treat multiple instruments?

The instruments get combined into the projection $S\left(S'S\right)^{-1}S'D$, i.e. a constant 1 and the linear projection of $\frac{y_i}{\exp(x_i\theta)}x_i$ on s_i, the projection as proposed by Bowden and Vansteelandt (2010).

GMM satisfies $$D'S\left(S'S\right)^{-1}S'v = 0$$

$$\begin{aligned} D &= \{d_i'\};\ S = \{s_i'\};\ v = \{v_i\} \\ d_i &= \begin{pmatrix} 1 \\ \frac{y_i}{\exp(x_i\theta)}x_i \end{pmatrix};\ v_i = \frac{y_i}{\exp\left(x_i\theta\right)} - \alpha \end{aligned}$$

Two step GMM

Step 1: Estimate parameters and *W*
Step 2: repeat optimization starting from step 1 estimate of *W*

$$\widehat{\delta}_2 = \arg\min_{\delta} \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)' W_n^{-1}\left(\widehat{\delta}_1\right) \left(\frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)$$

Two-step GMM is efficient because it's Vcov matrix is the *smallest* (Chamberlain 1987)

One step: $\sqrt{n}\left(\widehat{\delta}_1 - \delta_0\right) \xrightarrow{d} N\left(0, \left(C_0' W C_0\right)^{-1} C_0 W \Omega_0 W C_0 \left(C_0' W C_0\right)^{-1}\right)$

Two step: $\sqrt{n}\left(\widehat{\delta}_2 - \delta_0\right) \xrightarrow{d} N\left(0, \left(C_0' \Omega_0 C_0\right)^{-1}\right)$

MSMM implementation in Stata

`gmm` command (Stata version 11)

Vector of 1's automatically included

Moment condition

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

`lincom [theta]:_cons, eform` Causal risk ratio

`estat overid` Over-identification test

MSMM Stata output 1

```
.
. gmm (hyp*exp(-overw*{theta}) - {ey0}), instruments(Iz1 Iz2 Iz3)

Step 1
Iteration 0:   GMM criterion Q(b) =  .48211942
Iteration 1:   GMM criterion Q(b) =  .00021372
Iteration 2:   GMM criterion Q(b) =  6.662e-06
Iteration 3:   GMM criterion Q(b) =  6.572e-06

Step 2
Iteration 0:   GMM criterion Q(b) =  .00004253
Iteration 1:   GMM criterion Q(b) =  .00004253

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]
-------------+----------------------------------------------------------------
      /theta |   .3104495   .1192332     2.60   0.009     .0767568    .5441423
        /ey0 |   .5758842   .0388716    14.82   0.000     .4996973    .6520711
------------------------------------------------------------------------------
Instruments for equation 1: Iz1 Iz2 Iz3 _cons
```

Two step GMM

$E[Y(0)] = 0.58\ (0.50, 0.65)$

MSMM Stata output 2

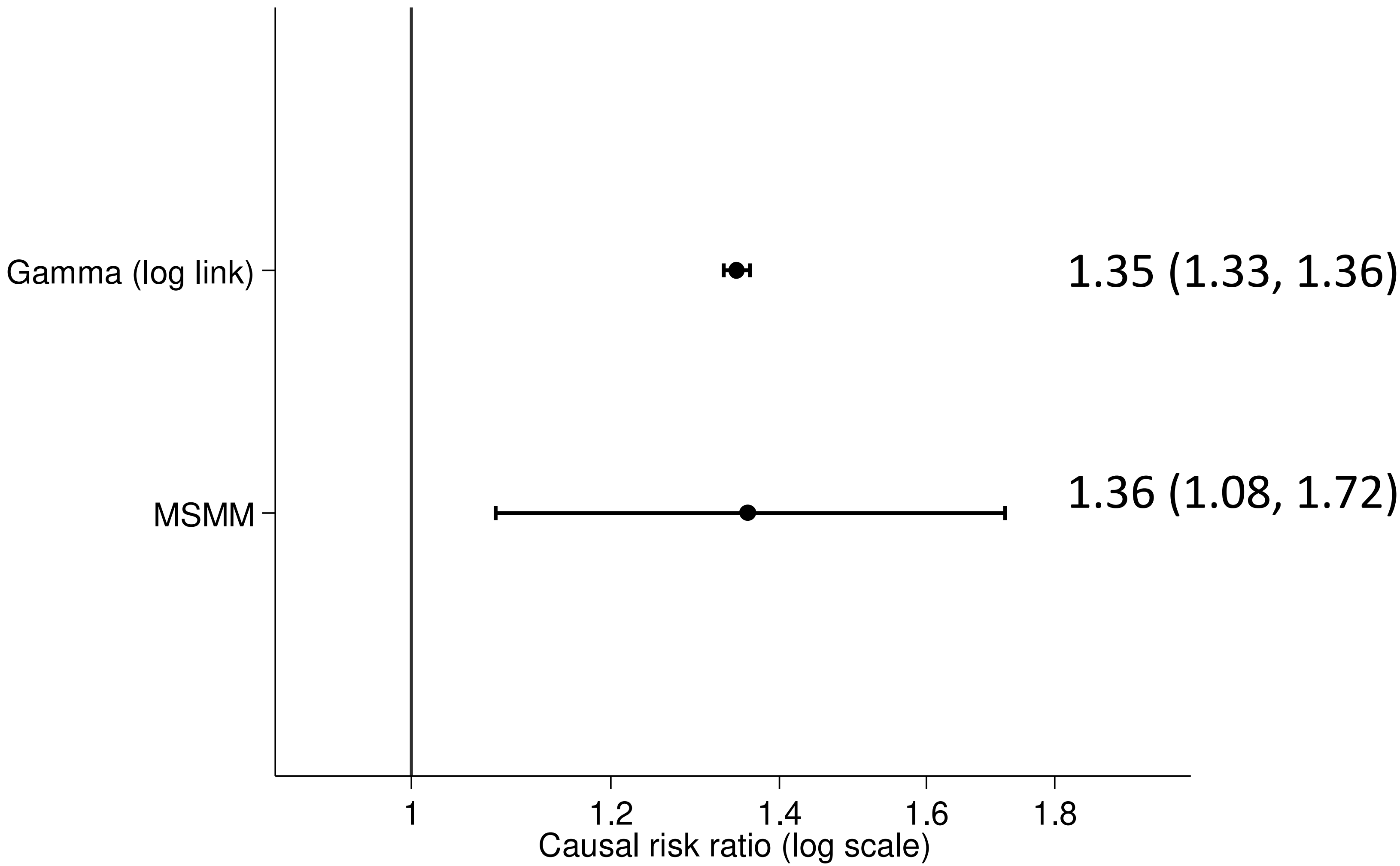
```
. lincom [theta]:_cons, eform

 ( 1)  [theta]_cons = 0
```

Causal risk ratio = 1.36 (1.08, 1.72)

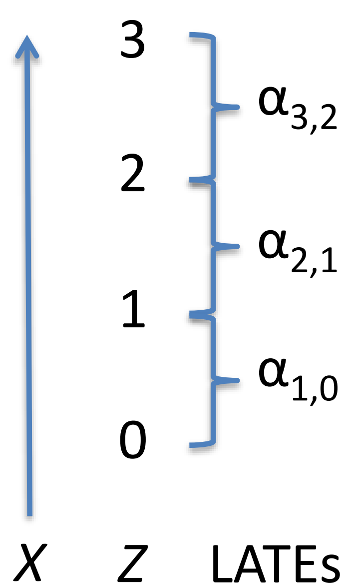
	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.364038	.1626386	2.60	0.009	1.079779	1.72313

Observational and IV estimate in example



Local risk ratios

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



Linear IV: Imbens & Angrist 1994

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

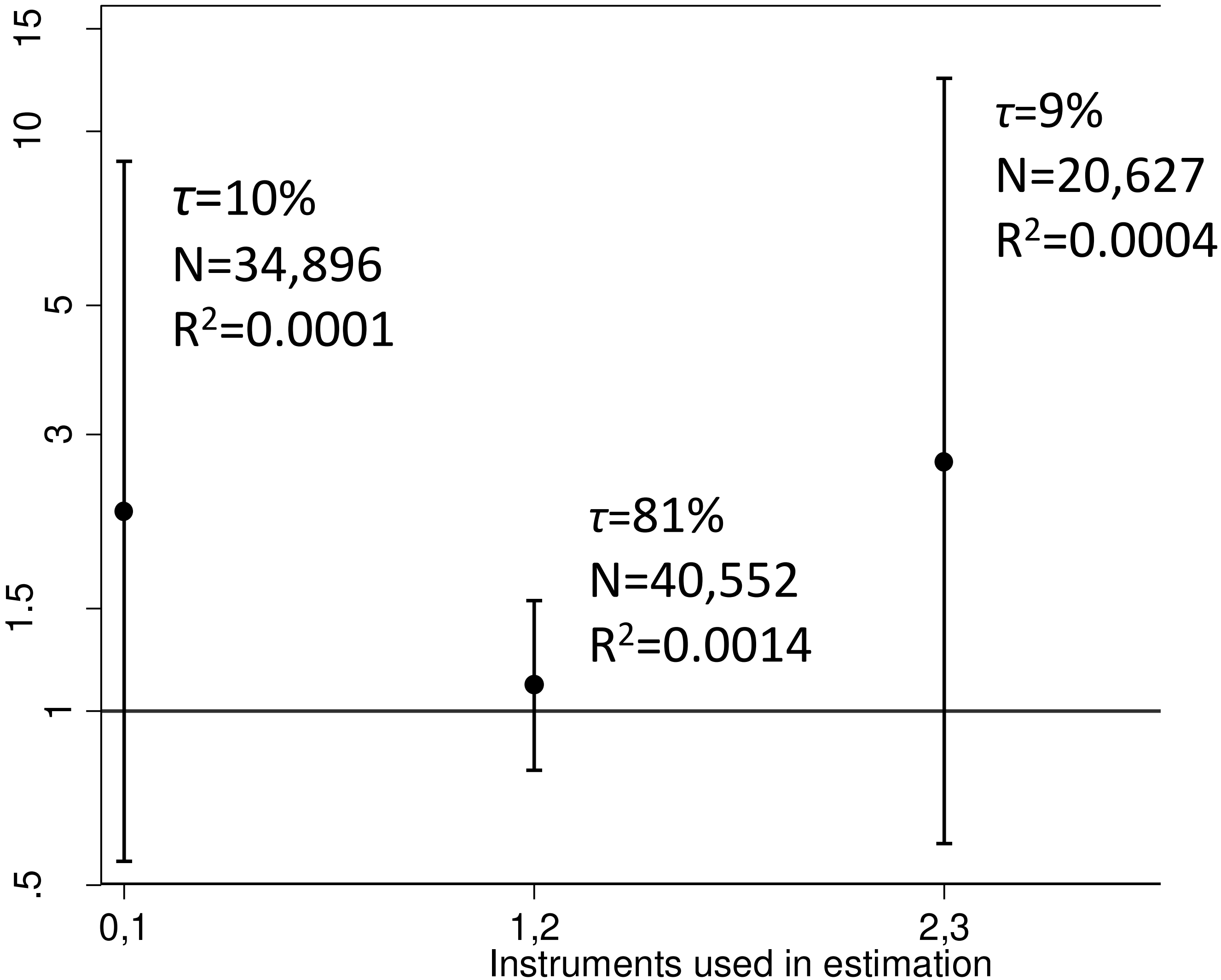
MSMM: We show a similar result holds for MSMM (*X, Y*: binary)

$$e_z^\theta = \sum_{k=1}^{K} \tau_k e_{k,k-1}^\theta$$

…weighted average of risk ratios

… rather than log risk ratios!

Local risk ratios in the example



N=55,523

R^2=0.0022

Check: (0.10 × 2.21) + (0.81 × 1.11) + (0.09 × 2.6

MSMM and MGMM

MGMM: Mullahy 1997 – exponential mean model with multiplicative residual

Additive residual: $Y = \exp(X\theta) + U$

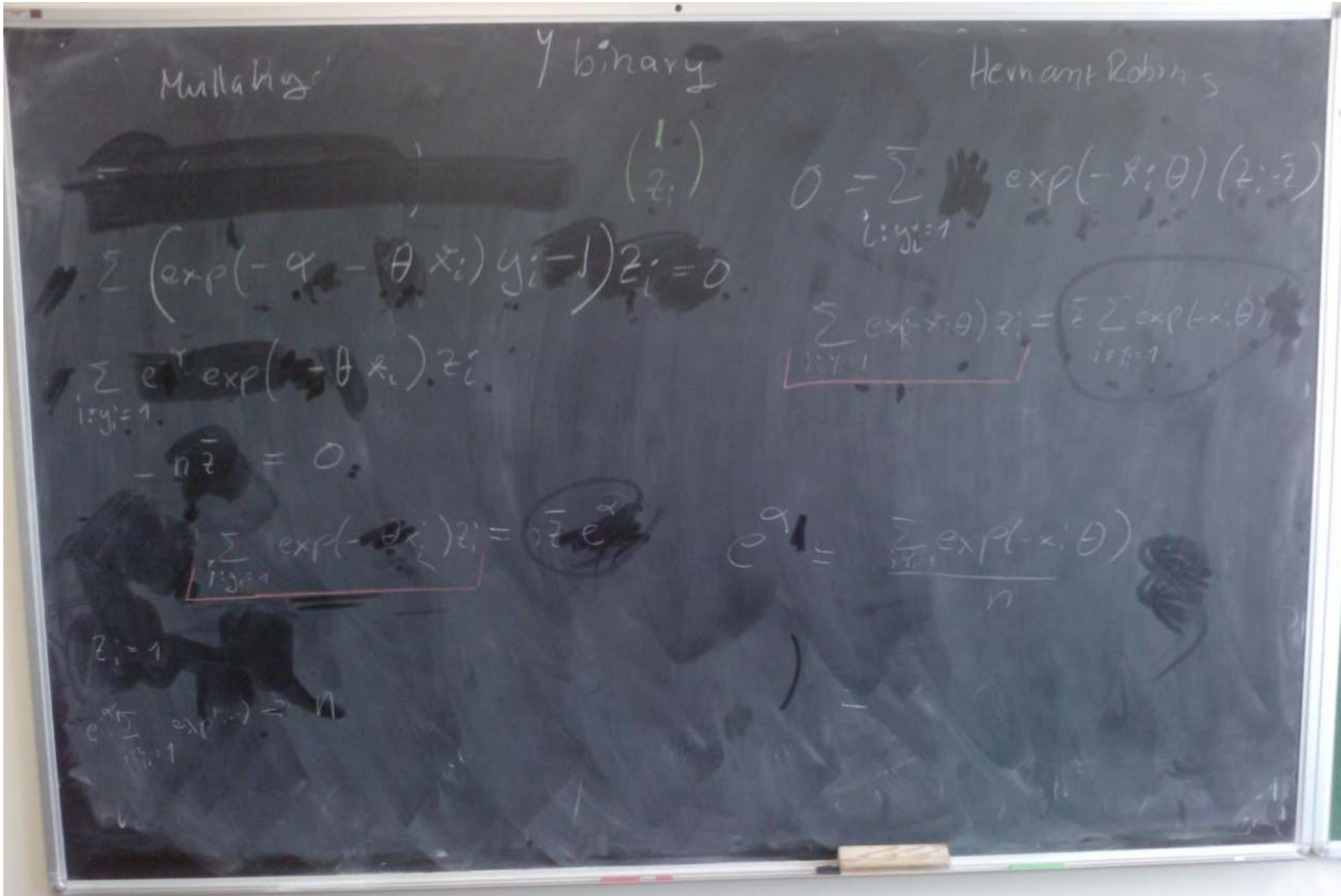
$E[Z\{Y - \exp(X\vartheta)\}] = 0$ Poisson regression

Multiplicative residual: $Y = \exp(X\theta + U)$

$$E\left[\frac{Y - \exp\left(\alpha_0^* + X\theta_0\right)}{\exp\left(\alpha_0^* + X\theta_0\right)} | S\right] = 0 \qquad S = (1, Z_1, Z_2)'$$

Discussed by Windmeijer 1997, 2002, 2006

Proof MSMM = MGMM



Clarke & Windmeijer 2010 ; Didelez, et al. 2010; Palmer et al., AJE, 2011
MGMM (one step GMM): `ivpois` for Stata (Nichols 2007)

Logistic SMM

- Implement joint estimation approach within GMM framework
- Vansteelandt & Goetghebeur (2003), Vansteelandt & Bowden (2010)

Two-stage estimation	Joint estimation
Stage 1 Association model: predict *Y* given *X*, *Z* **Stage 2** Causal model (MSMM/ASMM causal model only)	Estimate association model and causal model together
Need to correct SEs somehow	SEs automatically correct Gourieux 1996, Tan 2010

LSMM implementation in Stata

Two step estimation

```
logit y x z1 z2 xz1 xz2
matrix from = e(b)
predict xblog, xb
```

Association model: predict *Y* given *X*, *Z*

Causal model – incorrect SEs!

```
gmm (invlogit(xblog - x*{psi}) - {ey0}), instruments(z1 z2)
matrix from = (from,e(b))
```

Joint estimation – correct SEs!

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

lincom [psi]_cons, eform // causal odds ratio
estat overid
```

LSMM Stata output

```
. logit hyp overw Iz1 Iz2 Iz3 Iz1Xoverw Iz2Xoverw Iz3Xoverw

Iteration 0:   log likelihood =  -34179.76
Iteration 1:   log likelihood = -32895.818
Iteration 2:   log likelihood = -32885.846
Iteration 3:   log likelihood = -32885.845

Logistic regression                               Number of obs   =      55523
                                                  LR chi2(7)      =    2587.83
                                                  Prob > chi2     =     0.0000
Log likelihood = -32885.845                       Pseudo R2       =     0.0379
```

Association model

hyp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overw	.9034696	.0419769	21.52	0.000	.8211964	.9857428
Iz1	.0023852	.0346439	0.07	0.945	-.0655155	.070286
Iz2	-.031613	.0375747	-0.84	0.400	-.105258	.042032
Iz3	.0285799	.0598671	0.48	0.633	-.0887574	.1459173
Iz1Xoverw	.0500117	.0509504	0.98	0.326	-.0498493	.1498727
Iz2Xoverw	.06952	.0543206	1.28	0.201	-.0369465	.1759864
Iz3Xoverw	.041216	.0837708	0.49	0.623	-.1229717	.2054037
_cons	.3295621	.0285043	11.56	0.000	.2736947	.3854295

```
. matrix from = e(b)

. predict xblog, xb
```

predicted values of outcome (on logit scale here)

```
. gmm (invlogit(xblog - overw*{psi}) - {ey0}), instruments(Iz1 Iz2 Iz3)

Step 1
Iteration 0:   GMM criterion Q(b) =  .48211941
Iteration 1:   GMM criterion Q(b) =  .00078422
Iteration 2:   GMM criterion Q(b) =  .00001363
Iteration 3:   GMM criterion Q(b) =  .00001362

Step 2
Iteration 0:   GMM criterion Q(b) =   .1911576
Iteration 1:   GMM criterion Q(b) =  .16822374
Iteration 2:   GMM criterion Q(b) =  .13183731
Iteration 3:   GMM criterion Q(b) =  .13181315
Iteration 4:   GMM criterion Q(b) =  .13181311

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 |   .6331413   .0362588    17.46   0.000     .5620754    .7042073
        /ey0 |   .6226167    .004652   133.84   0.000      .613499    .6317344
------------------------------------------------------------------------------
Instruments for equation 1: Iz1 Iz2 Iz3 _cons
```

Causal model

Incorrect SEs

```
. gmm (hyp - invlogit({logit:overw Iz1 Iz2 Iz3 Iz1Xoverw Iz2Xoverw Iz3Xoverw}
> + {logitconst})) ///
>         (invlogit({logit:} + {logitconst} - overw*{psi}) - {ey0}), ///
>         instruments(1:overw Iz1 Iz2 Iz3 Iz1Xoverw Iz2Xoverw Iz3Xoverw) ///
>         instruments(2:Iz1 Iz2 Iz3) ///
>         winitial(unadjusted,independent) from(from)
```

Joint estimation

```
Iteration 2:   GMM criterion Q(b) =  .00004429

GMM estimation

Number of parameters =  10
Number of moments    =  12
Initial weight matrix: Unadjusted                 Number of obs   =    55523
GMM weight matrix:     Robust
```

Corrected SEs: causal model SEs ×10

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
/logit_overw	.9091545	.0418464	21.73	0.000	.8271371	.9911719
/logit_Iz1	-.0207159	.0279367	-0.74	0.458	-.0754708	.034039
/logit_Iz2	-.0339566	.0343049	-0.99	0.322	-.1011929	.0332796
/logit_Iz3	-.0058356	.0550491	-0.11	0.916	-.1137299	.1020586
/logit_Iz1~w	.039923	.0502901	0.79	0.427	-.0586438	.1384898
/logit_Iz2~w	.0687247	.0542023	1.27	0.205	-.0375099	.1749592
/logit_Iz3~w	.0262868	.0826922	0.32	0.751	-.135787	.1883605
/logitconst	.3425951	.0253272	13.53	0.000	.2929548	.3922354
/psi	1.05276	.4217052	2.50	0.013	.2262333	1.879287
/ey0	.5656666	.0592066	9.55	0.000	.4496238	.6817094

```
Instruments for equation 1: overw Iz1 Iz2 Iz3 Iz1Xoverw Iz2Xoverw
    Iz3Xoverw _cons
Instruments for equation 2: Iz1 Iz2 Iz3 _cons
```

```
. lincom [psi]_cons, eform

 ( 1)  [psi]_cons = 0
```

Causal odds ratio = 2.87 (1.25, 6.55)

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	2.86555	1.208417	2.50	0.013	1.253868	6.548836

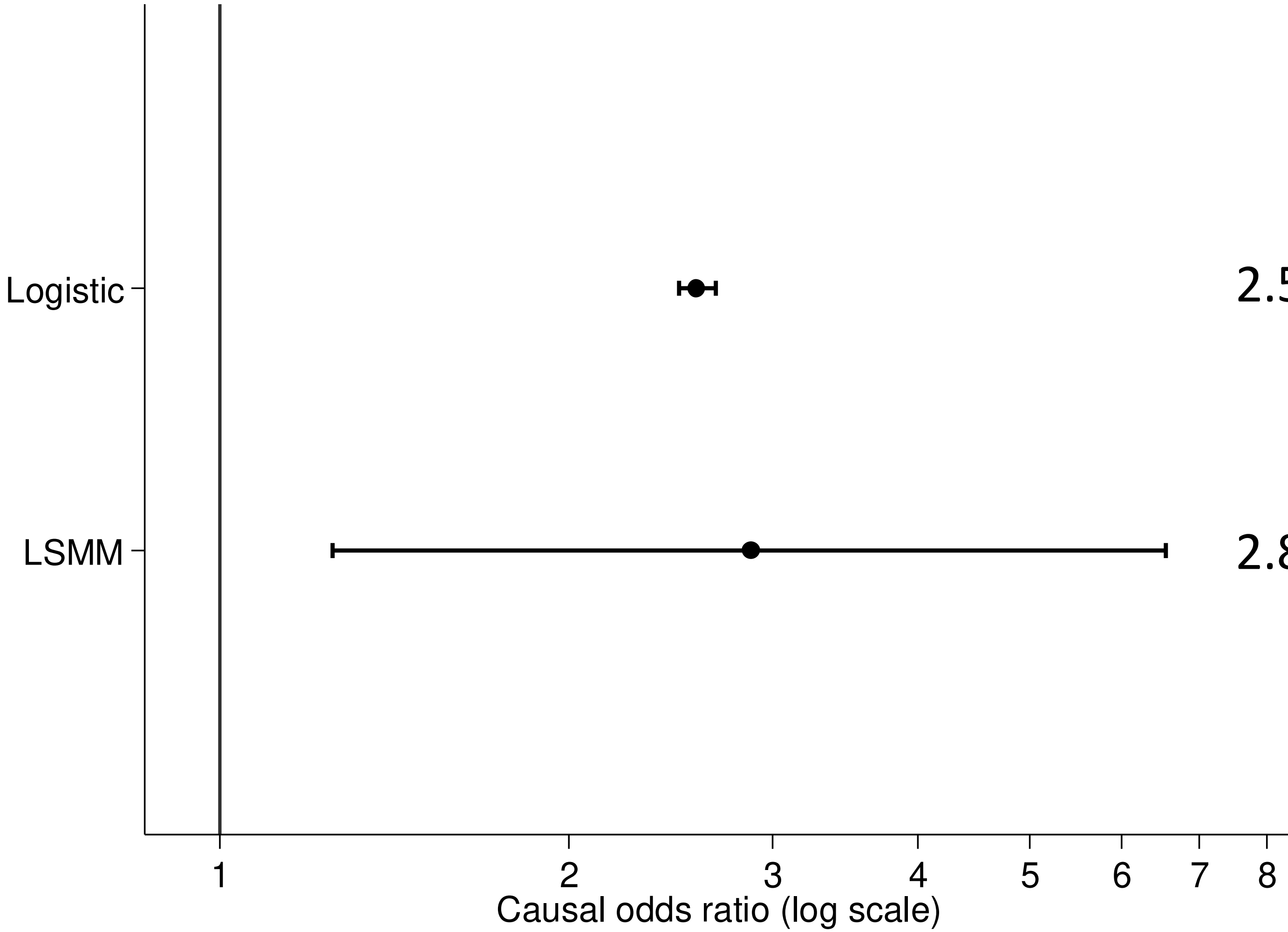
```
. estat overid

  Test of overidentifying restriction:

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

Degrees of freedom:
AM: exactly identified
CM: 4 moments – 2 pars

Observational and IV estimate in example



2.58 (2.49, 2.68)

2.87 (1.25, 6.55)

Summary

- Estimate SMMs within GMM framework
- GMM optimal combination of multiple instruments
- Two-step GMM is efficient
- Joint estimation for LSMM
- Hansen over-identification test
  - Joint validity of multiple instruments
  - Can help detect violations in NEM & CMI
- Straightforward implementation in Stata and R

References

Angrist, Imbens, Rubin, Identification of Causal Effects Using Instrumental Variables, JASA, 1996

Baum, Schaffer, Stillman, ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression, 2010. http://ideas.repec.org/c/boc/bocode/s425401.html

Bowden & Vansteelandt, Mendelian randomization analysis of case-control data using structural mean models, Stats Med, 2011

Chamberlain, Asymptotic efficiency in estimation with conditional moment restrictions, J Econ, 1987

Chausse, Computing Generalized Method of Moments and Generalized Empirical Likelihood with R, J Stat Soft, 2010

Clarke & Windmeijer, Identification of causal effects on binary outcomes using structural mean models, Biostatistics, 2010

Clarke & Windmeijer, Instrumental Variable Estimators for Binary Outcomes, CMPO Working Paper 09/209, 2010

Didelez & Sheehan, Mendelian randomization as an instrumental variable approach to causal inference, Stats Meth Med Res, 2007

Didelez et al., Assumptions of IV Methods for Observational Epidemiology, Stat Sci, 2010

Frayling et al., A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity, Science, 2007

Gourieroux et al., Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form, J Stat Plan Inf, 1996

Hernan & Robins, Instruments for Causal Inference: An Epidemiologist's Dream?, Epidemiol, 2006

Loos et al., Common variants near MC4R are associated with fat mass, weight and risk of obesity, Nature Genetics, 2008

Mullahy, Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior, The Review of Economics and Statistics, 1997

Nichols, ivpois: Stata Module to Estimate an Instrumental Variables Poisson Regression via GMM, 2007. http://ideas.repec.org/c/boc/bocode/s456890.html

Palmer et al., Instrumental Variable Estimation of Causal Risk Ratios and Causal Odds Ratios in Mendelian Randomization Analyses, AJE, 2011, in press

Robins, The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Health Service Research Methodology: A Focus on AIDS. Washington, 1989

Snowden et al., Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique, AJE, 2011

Tan, Marginal and Nested Structural Models Using Instrumental Variables, JASA, 2010

Windmeijer & Santos Silva, Endogeneity in Count Data Models: An Application to Demand for Health Care, J Appl Econ, 1997

Windmeijer, ExpEnd, A Gauss Programme for Non-Linear GMM Estimation of Exponential Models With Endogenous Regressors for Cross Section and Panel Data, CEMMAP Working Paper CWP14/02, 2002

Windmeijer, GMM for Panel Count Data Models, CEMMAP Working Paper CWP21/06, 2006

Vansteelandt & Goetghebeur, Causal Inference with Generalized Structural Mean Models, JRSS B, 2003

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, Roger Harbord, Sha Meng, Nic Timpson, Borge Nordestgaard, John Thompson, Martin Tobin.