Lack of identification in structural mean models and multiple paired comparisons for investigating pleiotropy

Tom Palmer

Division of Health Sciences, Warwick Medical School,
University of Warwick, UK

9 September 2014

THE UNIVERSITY OF WARWICK

Warwick Medical School

Outline

1. Lack of identification in structural mean models (SMMs)
2. Mulitple paired comparisons for investigating pleiotropy
3. Summary

Lack of identification in SMMs

- Palmer TM, Sterne JAC, Harbord RM, Lawlor DA, Sheehan NA, Meng S, Granell R, Davey Smith G, Didelez V. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. American Journal of Epidemiology, 2011, 173 (12), 1392–1402.
- Clarke PS, Palmer TM, Windmeijer F. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. CMPO working paper 11/266.
- Burgess S, Granell R, Palmer TM, Sterne JAC, Didelez V. Lack of identification in semiparametric instrumental variable models with binary outcomes. American Journal of Epidemiology, 2014, 180 (1), 111–119.
- Granell R, Henderson AJ, Evans DM, Davey Smith G, Ness AR, Lewis S, Palmer TM, Sterne JAC. Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study, 2014, PLoS Medicine, 11 (7), e1001669.

Multiplicative SMM

Robins defined the multiplicative SMM as follows:

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

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

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

ψ : log causal risk ratio

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

GMM estimation of MSMM

Under the instrumental variable assumptions:

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

GMM estimation of MSMM

Under the instrumental variable assumptions:

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

Moment conditions (Clarke et al. Tech rep 2011)

Z=0,1

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

GMM estimation of MSMM

Under the instrumental variable assumptions:

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

Moment conditions (Clarke et al. Tech rep 2011)

Z=0,1,2 Over-identified

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

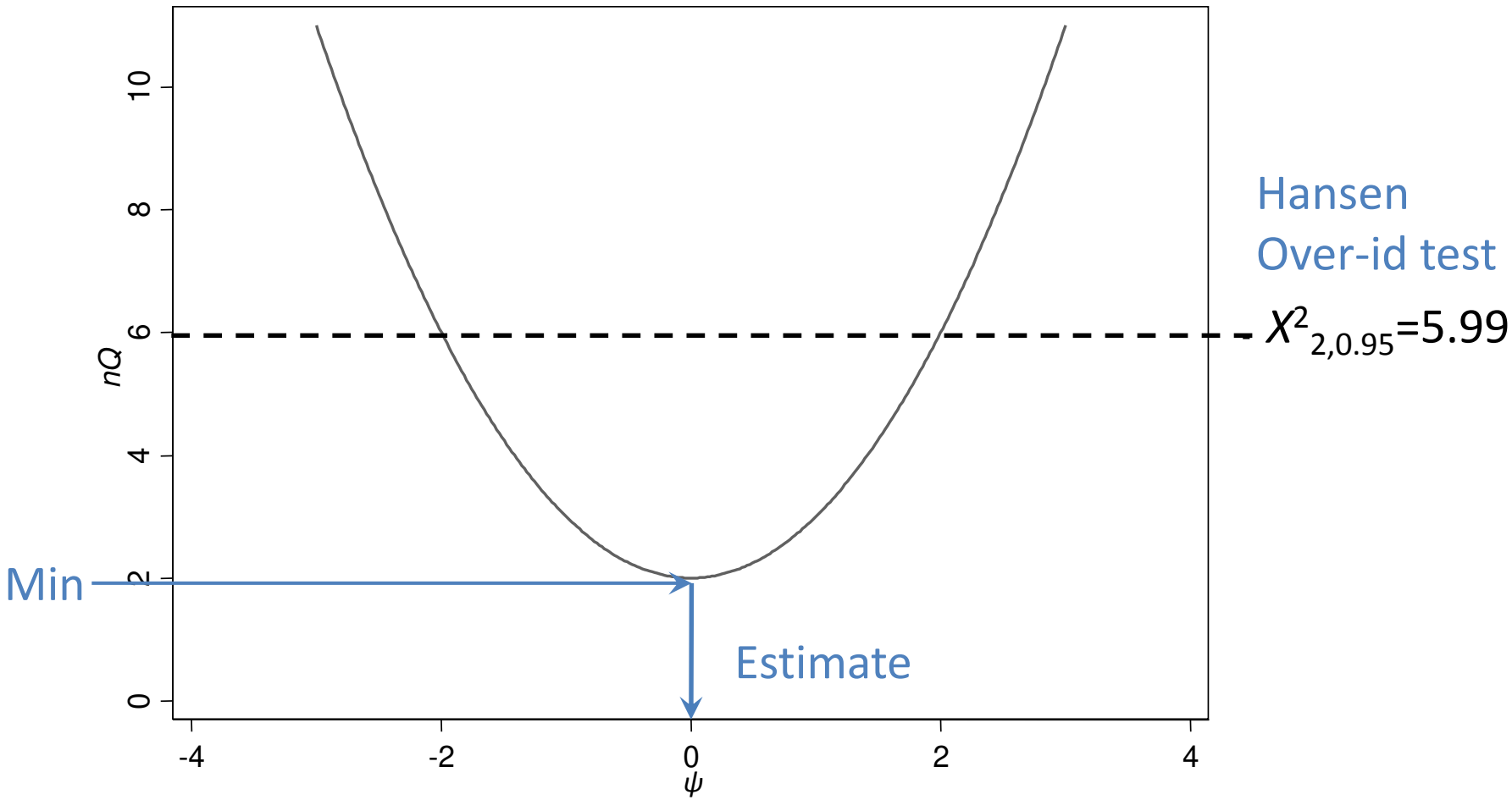
GMM estimation of MSMM

MSMM Stata gmm syntax

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

What is GMM?

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



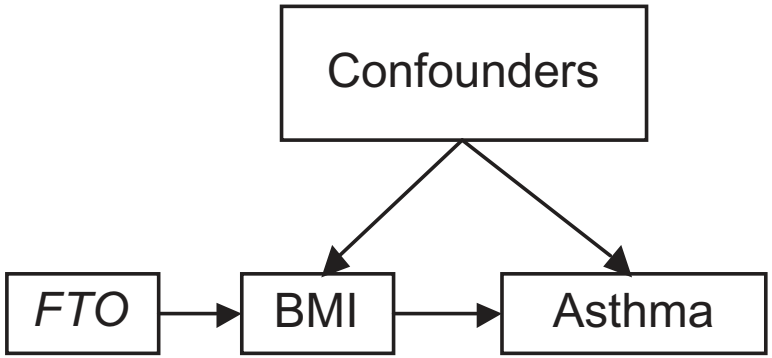
Alternative estimation approach

Bowden and Vansteelandt, Stats Med, 2010.

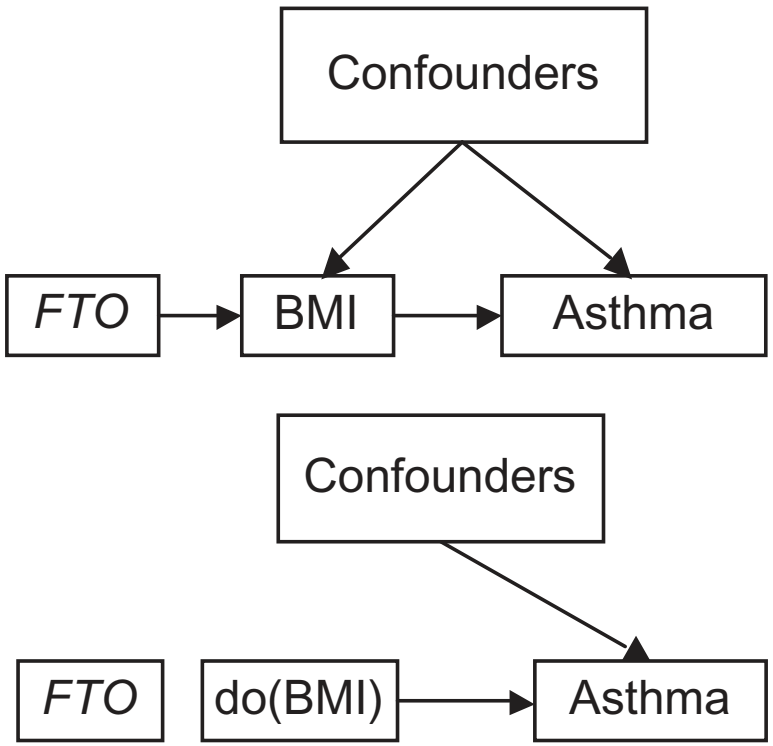
Solve estimating equation for ψ

$$\sum_{i=1}^{N} Y_i \exp(-\psi X_i)(Z_i - \overline{Z}) = 0$$

Asthma data example



Asthma data example



Asthma data example

Table 2. Distribution of Asthma and Possible Confounders by Fat Mass and Obesity-Associated (*FTO*) Genotype (rs9939609) in Children Aged 7 Years, Avon Longitudinal Study of Parents and Children, 1991–1992

	Total No.	TT		AT		AA		*P* Value From χ^2 Test
		No.	%	No.	%	No.	%	
No. and % of participants	4,647	1,699	37	2,220	48	728	16	0.95[a]
Asthma (yes)	4,647	234	13.8	302	13.6	113	15.5	0.41
Female sex	4,647	832	49	1,070	48	386	53	0.08
Low birth weight	4,594	75	4	80	4	36	5	0.21
Parental education (less than university degree)	4,593	893	54	1,214	56	390	55	0.44
Prenatal smoking	4,579	404	24	562	26	167	23	0.30
Postnatal smoking	4,407	270	17	390	19	115	17	0.23
Low parental social class	3,974	211	15	295	15	82	13	0.41

[a] Test for Hardy-Weinberg equilibrium.

Asthma data example

Table 3. Instrumental Variable Estimates of the Causal Odds Ratio and Causal Risk Ratio for the Effect of Body Mass Index on Asthma Risk, Avon Longitudinal Study of Parents and Children, 1991–1992

	COR or CRR	95% CI
Standard logistic regression analysis		
Unadjusted odds ratio	1.06	1.02, 1.10
Adjusted[a] odds ratio	1.08	1.03, 1.13
Wald/ratio estimator[b]		
CRR	1.37	0.64, 2.96
COR	1.45	0.65, 3.43
2-stage estimator[c]		
CRR	1.37	0.68, 2.78
COR	1.45	0.64, 3.29
Control function[c]		
CRR	1.37	0.68, 2.76
COR	1.44	0.63, 3.28
Logistic structural mean model[d]		
COR	1.64	0.29, 9.31
Multiplicative structural mean model[d]		
CRR	0.81	0.44, 1.48

Asthma data example

- Possible explanation for MSMM point estimate < 1
- Interaction between BMI and FTO genotype ($p = 0.038$)

Asthma data example

- Possible explanation for MSMM point estimate < 1
- Interaction between BMI and FTO genotype ($p = 0.038$)

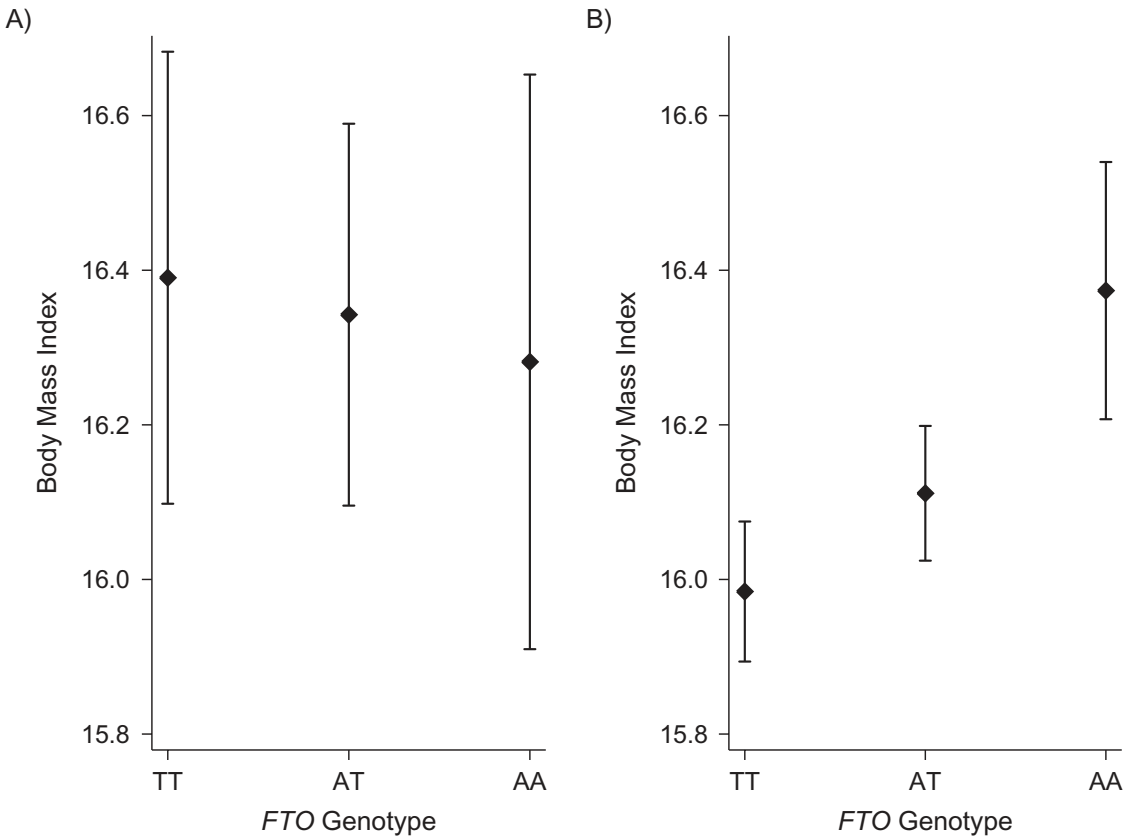


Figure 4. Mean body mass index (weight (kg)/height (m)^2), denoted by diamonds, according to fat mass and obesity-associated (*FTO*) genotype (rs9939609) for A) asthmatic and B) nonasthmatic children aged 7 years, Avon Longitudinal Study of Parents and Children, 1991–1992. Bars, 95% confidence interval.

Asthma data example

- This associational interaction could result from an interaction between FTO and unobserved confounders
- or could be a chance finding

Table 4. Results of Simulations Comparing the Multiplicative Generalized Method of Moments and 2-Stage Estimators of the Causal Risk Ratio

	2-Stage Estimate for Log CRR (MCE)	MGMM Estimate for Log CRR (MCE)
Scenario 1: no causal effect with interaction		
Mean bias	−0.007 (0.0046)	0.009 (0.0094)
MSE	0.021 (0.0010)	0.088 (0.0042)
Coverage	0.952 (0.0068)	0.964 (0.0059)
Correlation between estimates	−0.23	
% of estimates on opposite sides of the CRR of 1	64.1	

Asthma data example

- This associational interaction could result from an interaction between FTO and unobserved confounders
- or could be a chance finding

Scenario 2: causal effect with interaction		
Mean bias	−0.206 (0.0042)	−0.146 (0.0100)
MSE	0.060 (0.0019)	0.120 (0.0055)
Coverage	0.674 (0.0148)	0.919 (0.0086)
Correlation between estimates	−0.12	
% of estimates on opposite sides of the CRR of 1.2	35.9	

Asthma data example

- This associational interaction could result from an interaction between FTO and unobserved confounders
- or could be a chance finding

Scenario 3: no causal effect with no interaction			
Mean bias	−0.005 (0.0049)		−0.001 (0.0053)
MSE	0.024 (0.0010)		0.029 (0.0018)
Coverage	0.942 (0.0074)		0.964 (0.0059)
Correlation between estimates		0.88	
% of estimates on opposite sides of the CRR of 1		7.3	

Asthma data example

- This associational interaction could result from an interaction between FTO and unobserved confounders
- or could be a chance finding

Scenario 4: causal effect with no interaction			
Mean bias	0.003 (0.0043)		0.003 (0.0049)
MSE	0.018 (0.0009)		0.024 (0.0014)
Coverage	0.954 (0.0066)		0.964 (0.0059)
Correlation between estimates		0.82	
% of estimates on opposite sides of the CRR of 1.2		15	

Asthma data example – 2 solutions to estimating equation

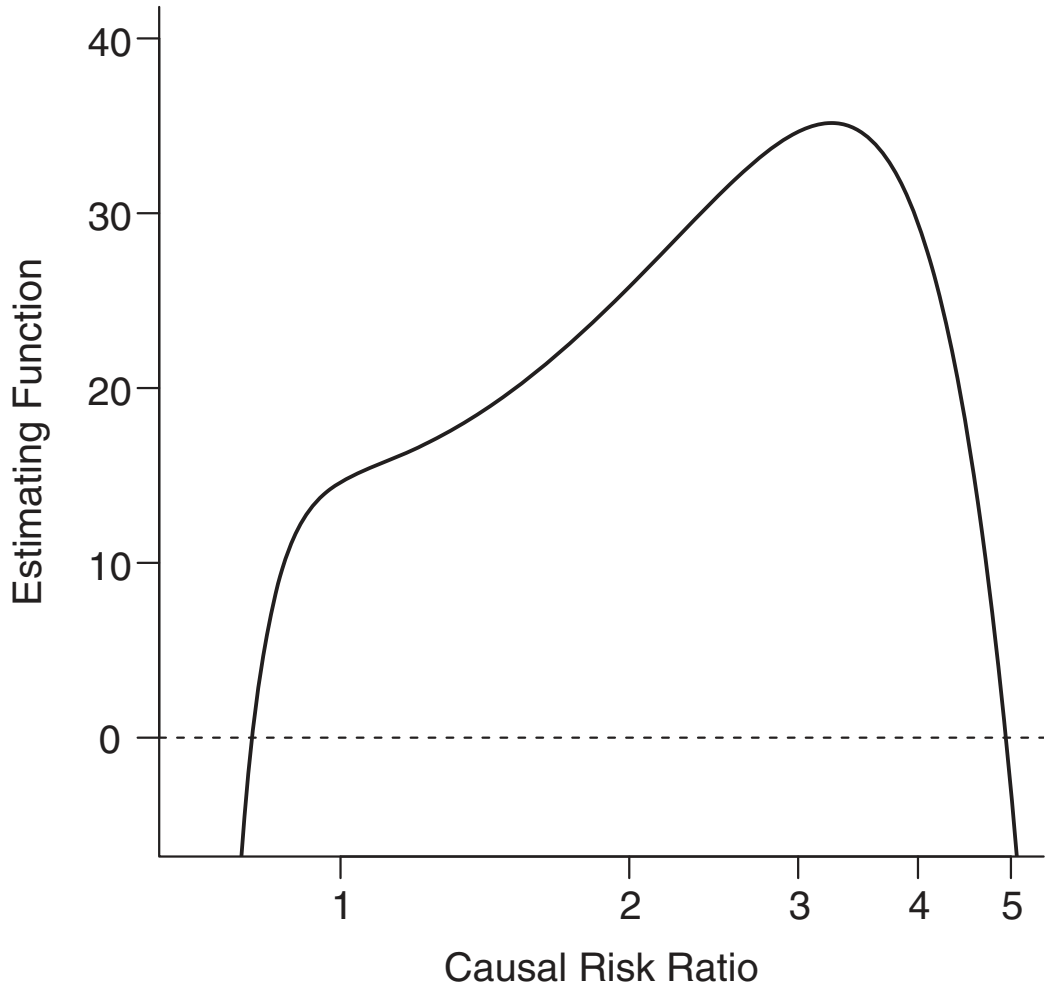


Figure 1. Estimating function for the example from Palmer et al. (20) demonstrating lack of identification. Two distinct parameter values for the causal risk ratio (0.81 and 4.95) satisfy the estimating equation $\sum_i y_i \exp(-\beta_1 x_i)(g_i - \bar{g}) = 0$, where $\bar{g}$ is the average value of G in the population.

Asthma data example

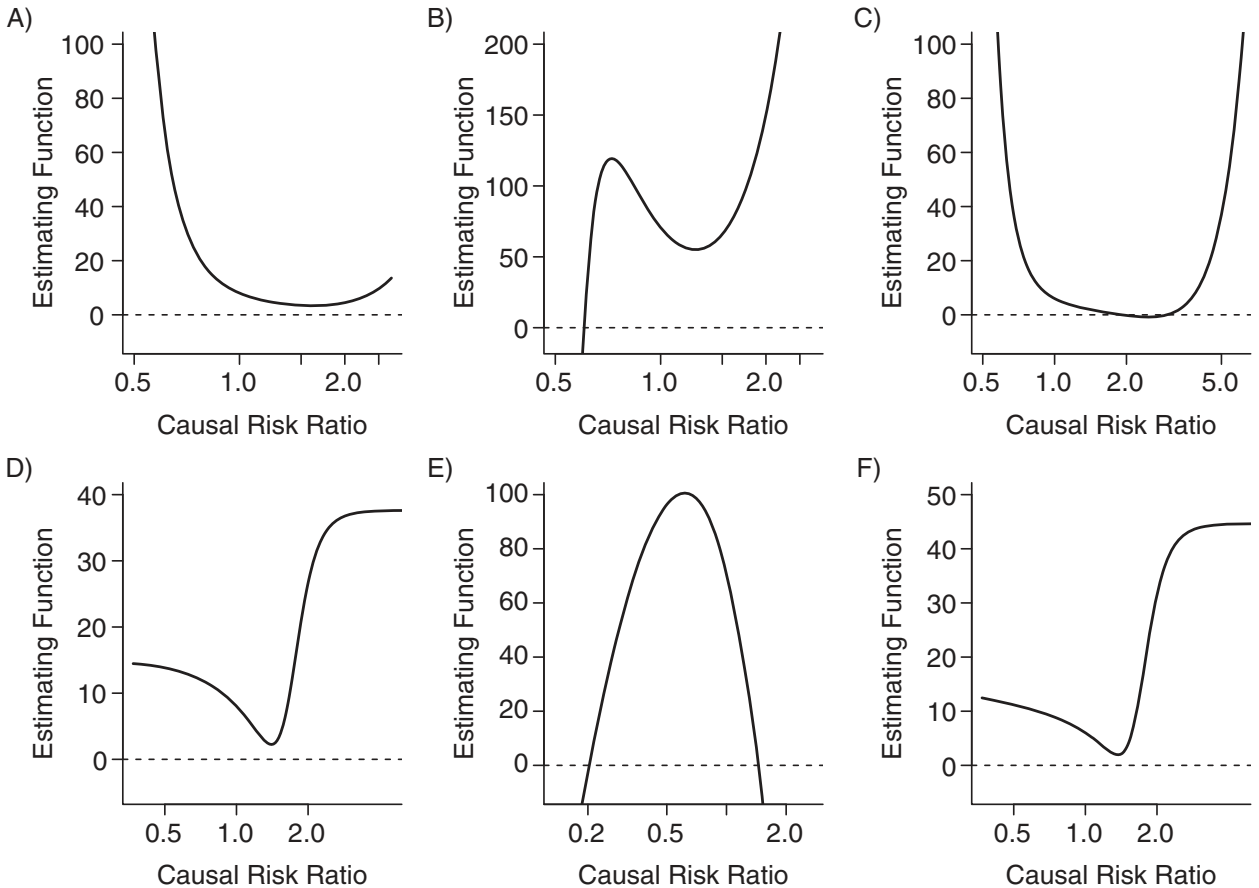


Figure 3. Estimating functions for the applied example from the multiplicative generalized method of moments method (in A, B, and C), and the linear generalized method of moments method (in D, E, and F) for the following 3 instruments: in A and D, a variant from the fat mass and obesity associated (*FTO*) gene; in B and E, the Speliotes score; and in C and F, the Speliotes score with the *FTO* genetic variant omitted. Avon Longitudinal Study of Parents and Children, 1991–1997.

Steve's simulations

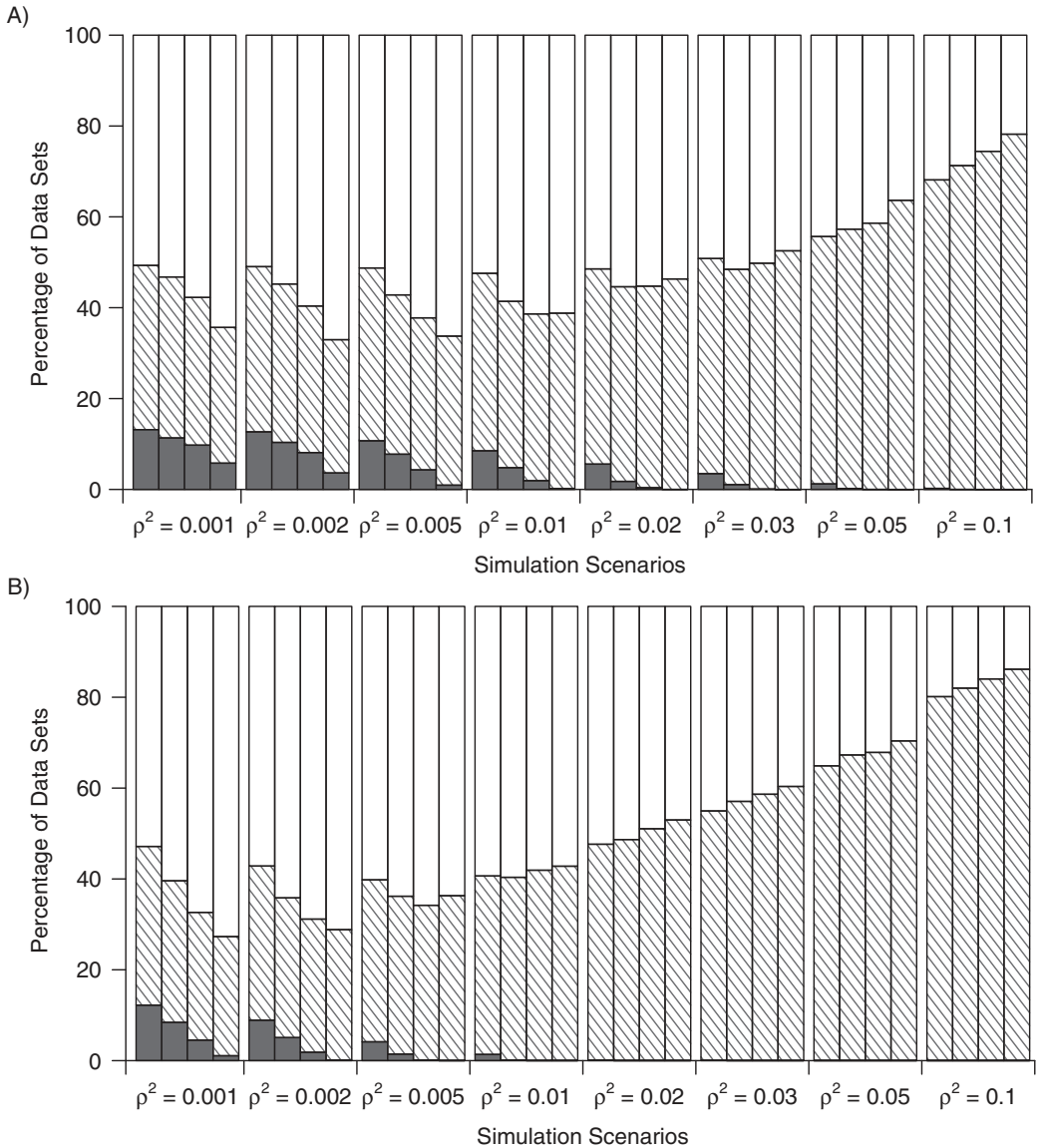


Figure 2. Percentage of simulated data sets with no solution (solid color), 1 solution (shaded), and multiple solutions (no color) from A) multiplicative generalized method of moments, and B) linear generalized method of moments methods with different strengths of instrument as measured by the squared correlation between the instrument and exposure (ρ^2) and different sample sizes (n). For each value of ρ^2, the first column is n = 5,000, the second column is n = 10,000, the third column is n = 20,000, and the fourth column is n = 50,000.

Related work: Brumback et al. SNMs 3-armed trial

- Brumback et al., Stats Med, 2014 "Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial"
- performed estimation using grid search
- 1 example of MSMM no solution (Appendix B)
- 3 examples of logistic SMM no solution (Appendix C)
- No examples of SMM with more than 1 solution

Brumback Appendix B example

Z_i	A_i	Y_i	freq/n
0	0	0	0.13
0	0	1	0.12
0	1	0	0.07
0	1	1	0.18
1	0	0	0.1
1	0	1	0.09
1	1	0	0.21
1	1	1	0.10

Brumback Appendix B example

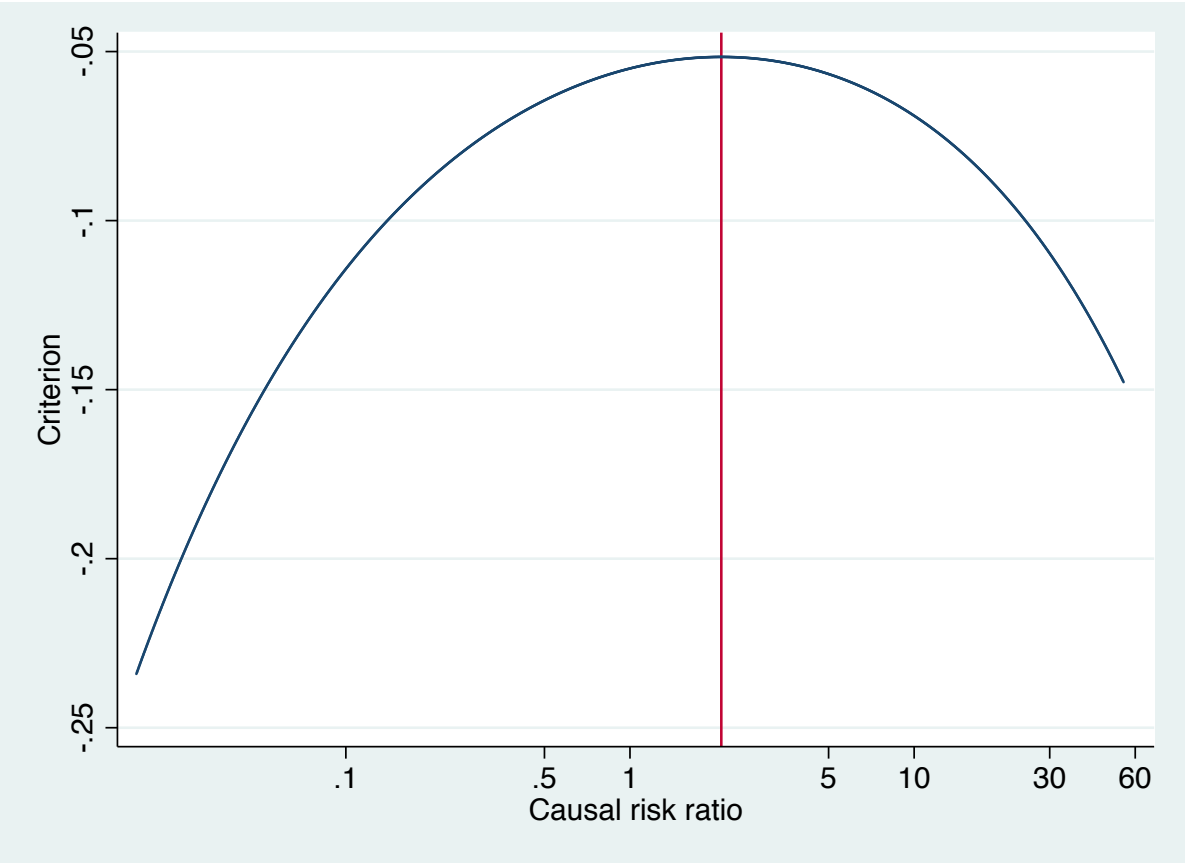
```
. tab a z, chi2

           |           z
         a |         0           1 |     Total
-----------+----------------------+----------
         0 |       125          95 |       220
         1 |       125         155 |       280
-----------+----------------------+----------
     Total |       250         250 |       500

          Pearson chi2(1) =   7.3052   Pr = 0.007

. tab y z, chi2

           |           z
         y |         0           1 |     Total
-----------+----------------------+----------
         0 |       100         155 |       255
         1 |       150          95 |       245
-----------+----------------------+----------
     Total |       250         250 |       500

          Pearson chi2(1) =  24.2097   Pr = 0.000
```

Brumback Appendix B example

```
. regress a z

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =    7.38
       Model |         1.8     1         1.8           Prob > F      =  0.0068
    Residual |       121.4   498    .2437751           R-squared     =  0.0146
-------------+------------------------------           Adj R-squared =  0.0126
       Total |       123.2   499  .246893788           Root MSE      =  .49374

------------------------------------------------------------------------------
           a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           z |        .12   .0441611     2.72   0.007      .033235     .206765
       _cons |         .5   .0312266    16.01   0.000     .4386479    .5613521
------------------------------------------------------------------------------
```

Brumback Appendix B example

- ASMM risk difference = -1.83 (95% CI -3.36, -0.30)
- MSMM estimating equation plot
  (with centred X and Z; closest to 0 at CRR=2.10)



Brumback Appendix B example

```
. gmm (y*exp(-1*a*{psi}) - {ey0}), instruments(z) ///
>       conv_maxiter(500) nolog
convergence not achieved
```

Brumback Appendix B example

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

Final GMM criterion Q(b) =  .2518336

GMM estimation

Number of parameters =   1
Number of moments    =   2
Initial weight matrix: Unadjusted                     Number of obs   =      500

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |   .0774102   1.764262     0.04   0.965    -3.380479    3.535299
------------------------------------------------------------------------------
Instruments for equation 1: c_z _cons

. lincom [psi]_cons, eform

 ( 1)  [psi]_cons = 0

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.080485   1.906258     0.04   0.965     .0340312    34.30528
------------------------------------------------------------------------------
```

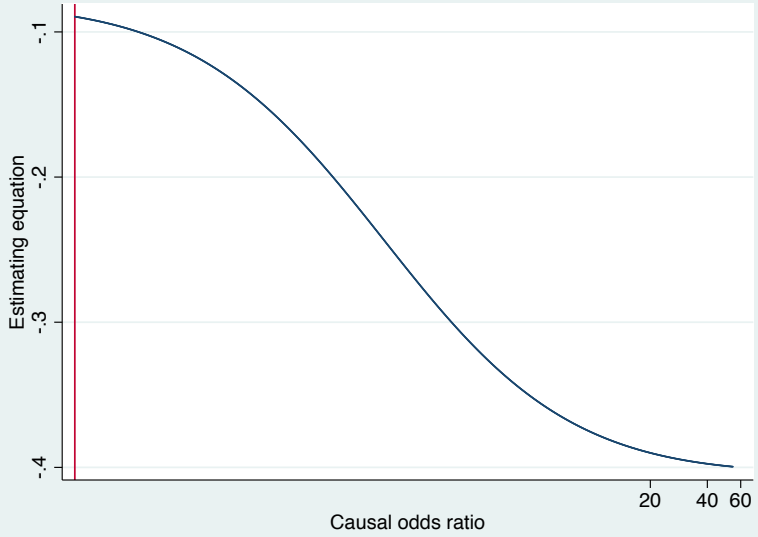
Brumback Appendix B example

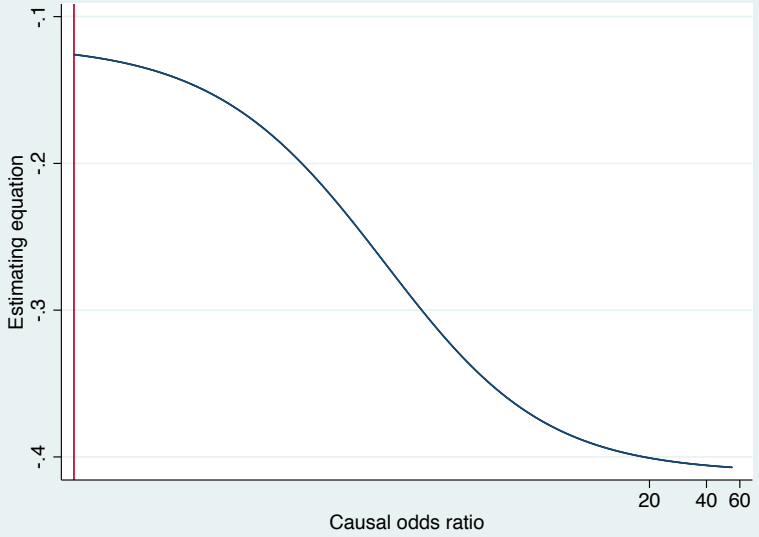
```
. gmm (y*exp(-1*c_a*{psi})), instruments(c_z) nolog

Final GMM criterion Q(b) =  .4891671

GMM estimation

Number of parameters =   1
Number of moments    =   2
Initial weight matrix: Unadjusted                     Number of obs   =      500
GMM weight matrix:     Robust

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        /psi |   .0474277   1.770337     0.03   0.979     -3.42237    3.517225
------------------------------------------------------------------------------
Instruments for equation 1: c_z _cons

. lincom [psi]_cons, eform

 ( 1)  [psi]_cons = 0

------------------------------------------------------------------------------
             |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    1.04857   1.856323     0.03   0.979      .032635    33.69082
------------------------------------------------------------------------------
```

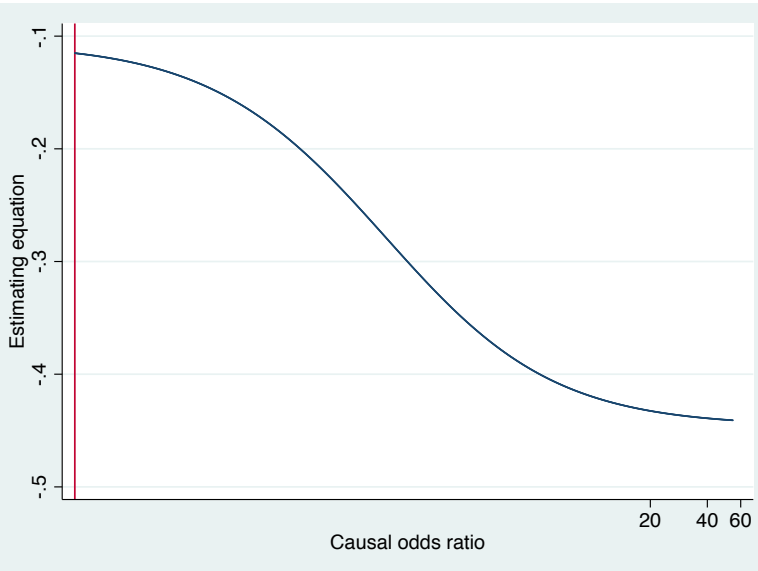
Brumback Appendix C example

Z_i	A_i	Y_i	E(freq)	freq 1	freq 2	freq 3
0	0	0	80	81	79	84
0	0	1	20	18	12	14
0	1	0	10	14	9	9
0	1	1	6.6667	7	8	6
0	2	0	5.5556	3	9	7
0	2	1	11.1111	3	8	7
1	0	0	12.5	17	9	11
1	0	1	4.17	4	8	3
1	1	0	66.6667	69	70	69
1	1	1	33.3333	36	25	27
1	2	0	5.5556	7	5	3
1	2	1	11.1111	6	18	11
2	0	0	11.1111	12	17	12
2	0	1	5.5556	17	6	13
2	1	0	10	5	17	10
2	1	1	6.6667	8	9	9
2	2	0	50	46	54	40
2	2	1	50	37	35	56

Brumback Appendix C example







Lack of identification of SMMs: Summary

- Don't just rely on `gmm` or whatever software you're using
- Plot the estimating equation for different values of ψ when fitting SMMs
- We found SMMs with 0, 1, and 2 solutions
- Future work: For logistic SMM alternative estimation strategy, PROC NLMIXED (Matsouaka & Tchetgen Tchetgen, Tech. Rep., 2014)

Multiple paired comparisons for investigating pleiotropy

Something to watch out for with ivregress/ivreg2 in Stata

Unusual results – simulations, TSLS, allele score as single IV

Something to watch out for with ivregress/ivreg2 in Stata

Unusual results – simulations, TSLS, allele score as single IV

```
. ivreg2 fvc (height = unwscore15), nocollin

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =     4216
                                                      F(  1,  4214) =  2.1e+05
                                                      Prob > F      =   0.0000
Total (centered) SS     =  423750161.1                Centered R2   =   0.2645
Total (uncentered) SS   =  1.60231e+10                Uncentered R2 =   0.9805
Residual SS             =    311685745                Root MSE      =    271.9

------------------------------------------------------------------------------
         fvc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   14.51982   .0316094   459.35   0.000     14.45787    14.58177
       _cons |          0  (omitted)
------------------------------------------------------------------------------
```

Something to watch out for with ivregress/ivreg2 in Stata

Unusual results – simulations, TSLS, allele score as single IV

```
. ivregress 2sls fvc (height = unwscore15)

Instrumental variables (2SLS) regression          Number of obs =    4216
                                                  Wald chi2(1)  =    0.00
                                                  Prob > chi2   =  0.9979
                                                  R-squared     =  0.1475
                                                  Root MSE      =  292.72

------------------------------------------------------------------------------
         fvc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   7.155347    2688.03     0.00   0.998    -5261.287    5275.598
       _cons |   975.6267   356103.4     0.00   0.998    -696974.3    698925.5
------------------------------------------------------------------------------
```

Something to watch out for with ivregress/ivreg2 in Stata

One solution is to center the intermediate:

```
. ivregress 2sls fvc (c_height = unwscore15)

Instrumental variables (2SLS) regression          Number of obs =    4216
                                                  Wald chi2(1)  =    0.00
                                                  Prob > chi2   =  0.9979
                                                  R-squared     =  0.1475
                                                  Root MSE      =  292.72

------------------------------------------------------------------------------
         fvc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    c_height |   7.155347    2688.03     0.00   0.998    -5261.287    5275.598
       _cons |   1923.549    4.50827   426.67   0.000     1914.713    1932.385
------------------------------------------------------------------------------
```

Something to watch out for with ivregress/ivreg2 in Stata

One solution is to center the intermediate:

```
. ivreg2 fvc (c_height = unwscore15), nocollin

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      4216
                                                      F(  1,  4214) =      0.00
                                                      Prob > F      =    0.9979
Total (centered) SS     =  423750161.1                Centered R2   =    0.1475
Total (uncentered) SS   =  1.60231e+10                Uncentered R2 =    0.9775
Residual SS             =  361258681.1                Root MSE      =     292.7

------------------------------------------------------------------------------
         fvc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    c_height |   7.155347    2688.03     0.00   0.998    -5261.287    5275.598
       _cons |   1923.549    4.50827   426.67   0.000     1914.713    1932.385
------------------------------------------------------------------------------
```

Multiple paired comparisons for investigating pleiotropy

Multiple paired comparisons for investigating pleiotropy



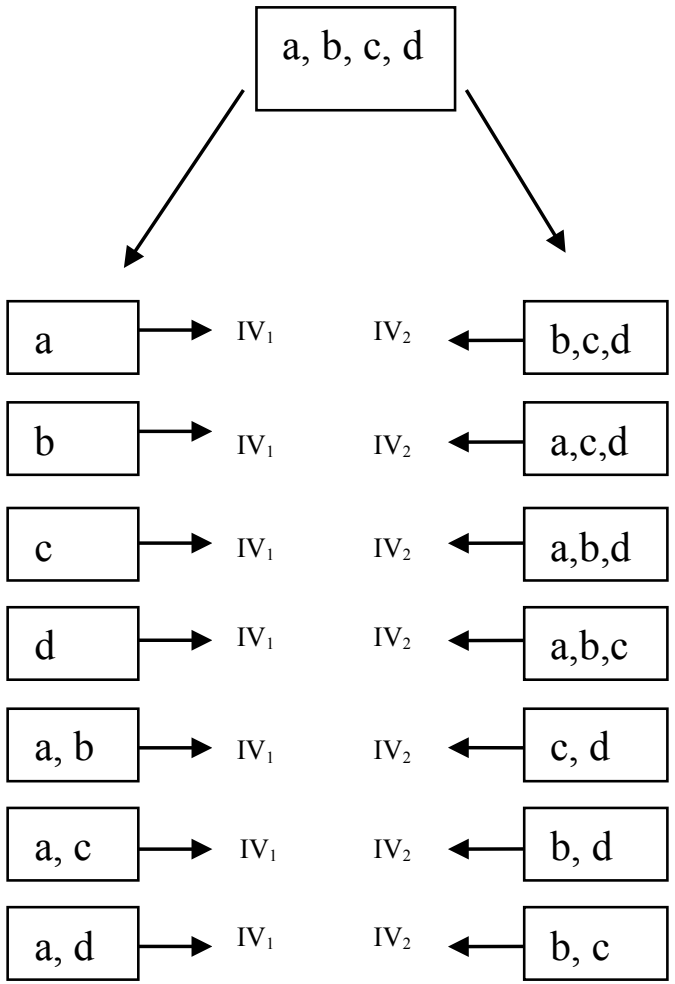
Multiple paired comparisons for investigating pleiotropy

Mendel

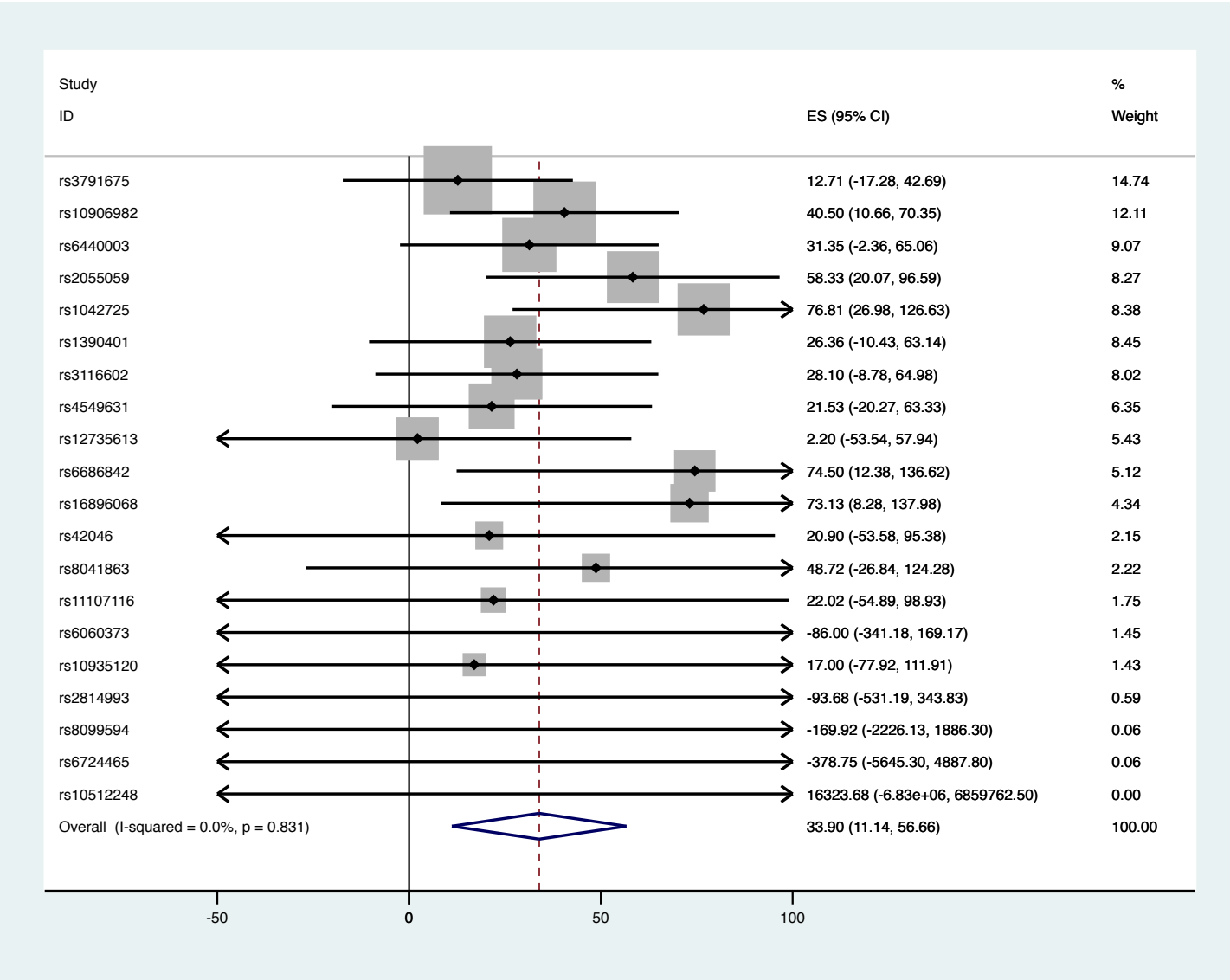
Sargan

Hansen

The idea

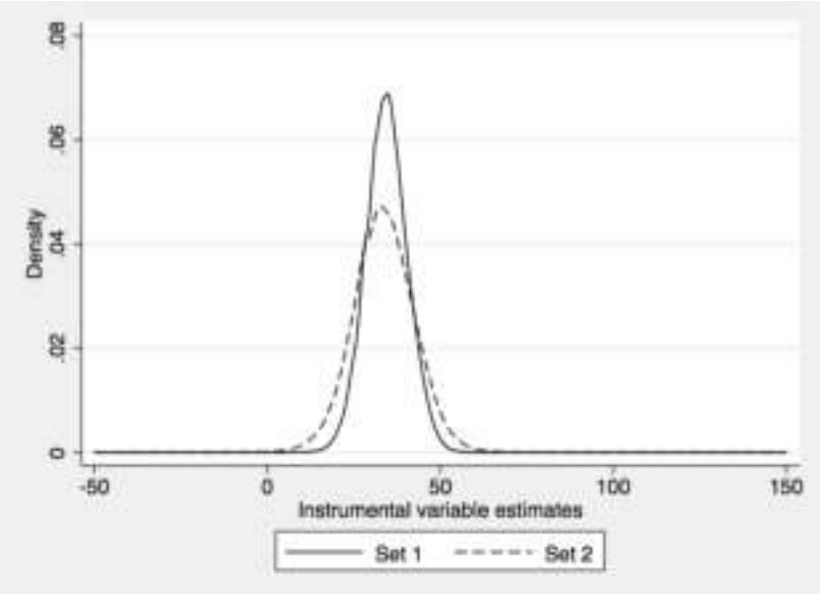


Example: effect of height on lung capacity (FVC) 20 SNPs

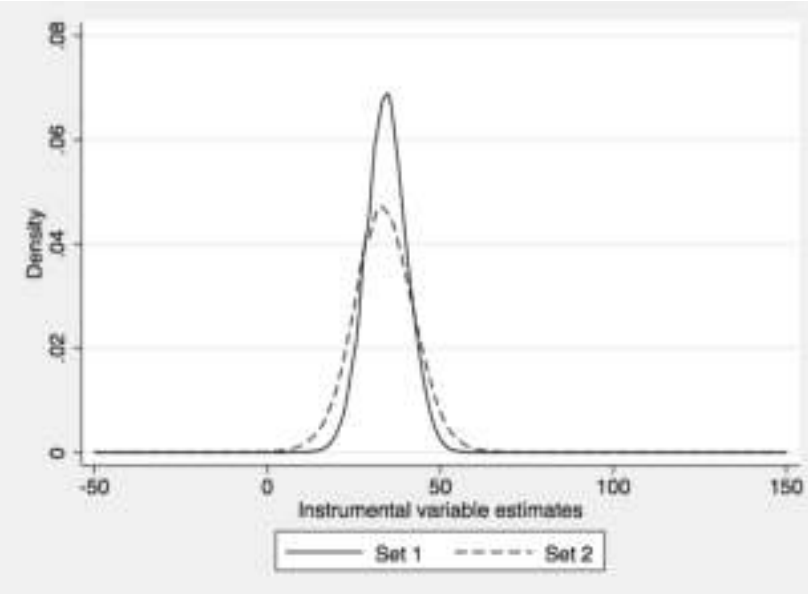


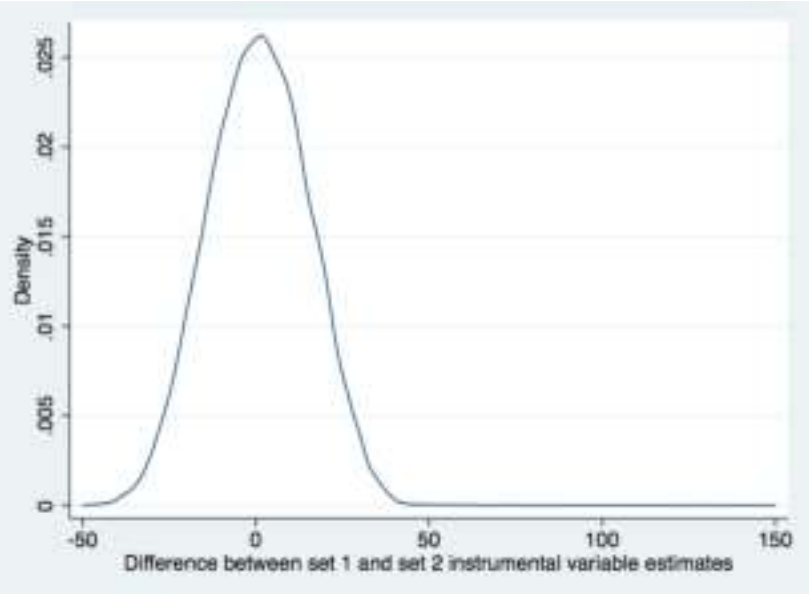
TSLS estimate 33.9 (23.6, 44.2), Sargan over-id test p=0.011

Example: using SNPs as multiple instruments in TSLS

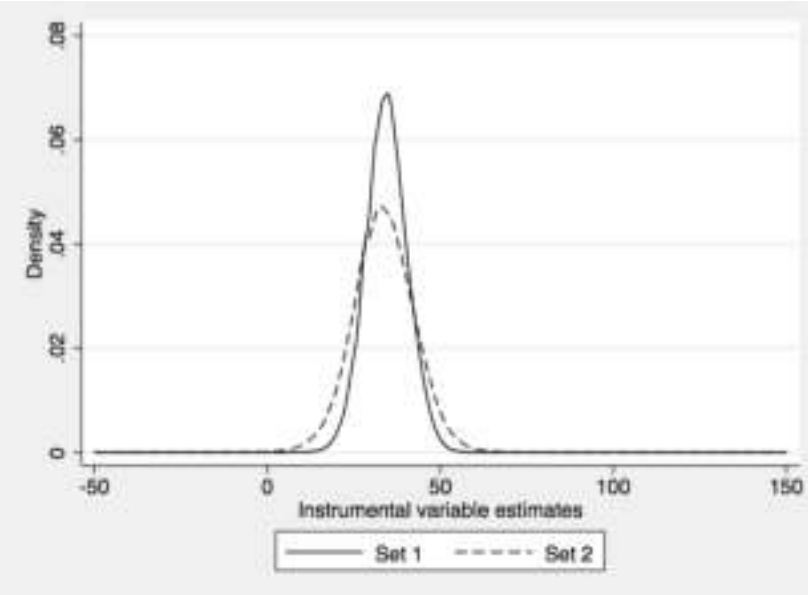


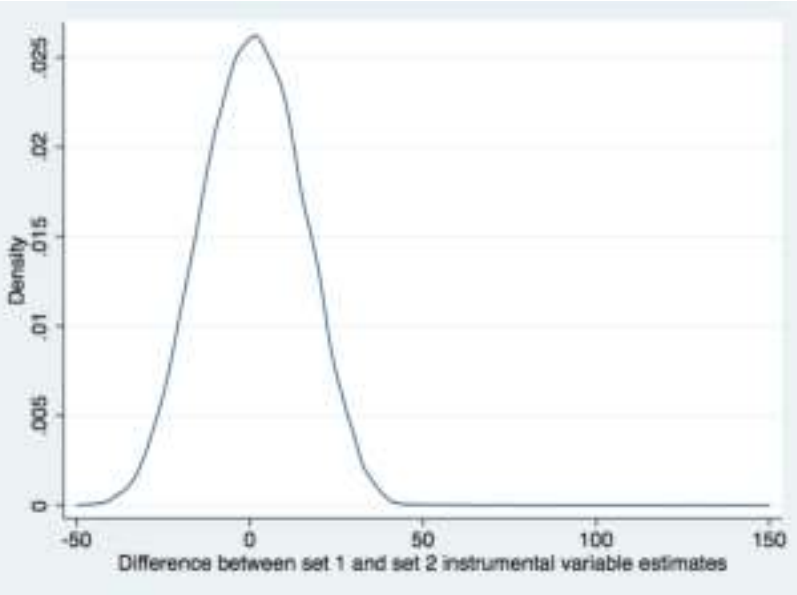
Example: using SNPs as multiple instruments in TSLS

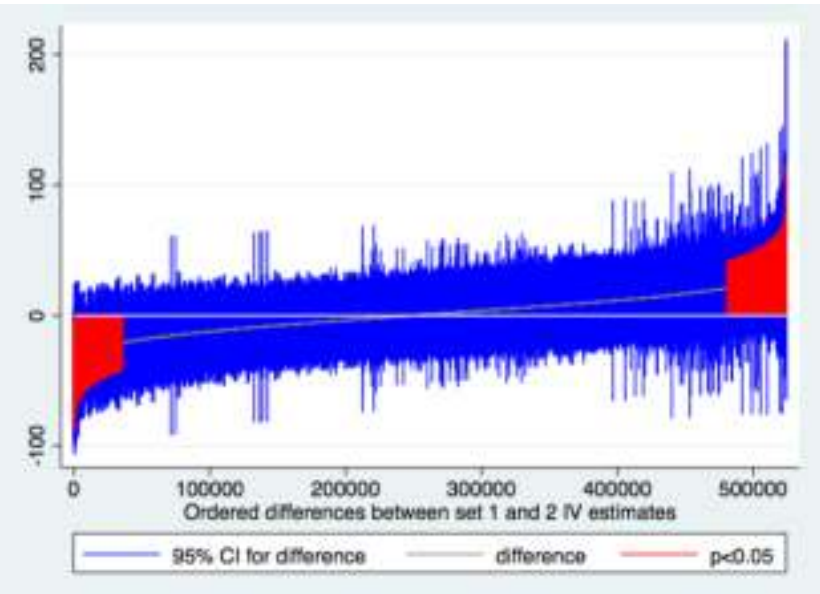




Example: using SNPs as multiple instruments in TSLS





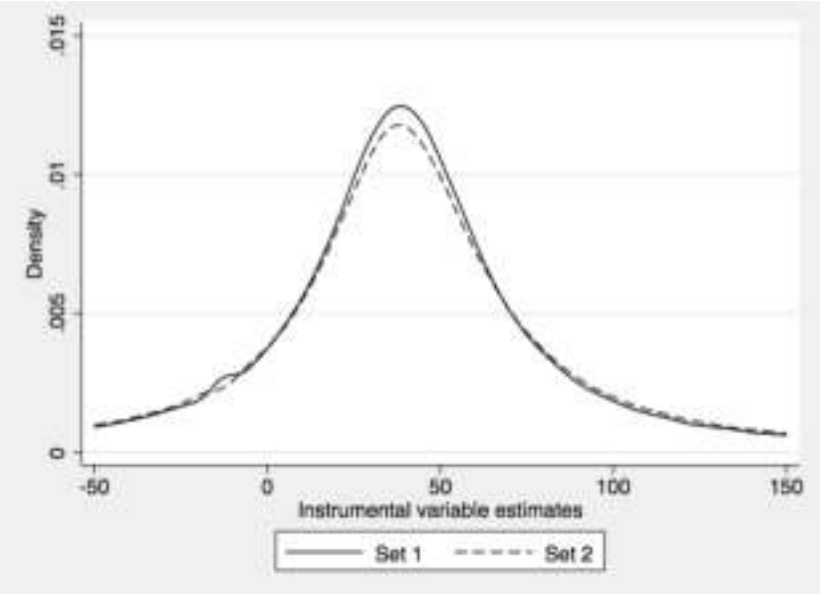


12% of 95% CIs exclude 0.

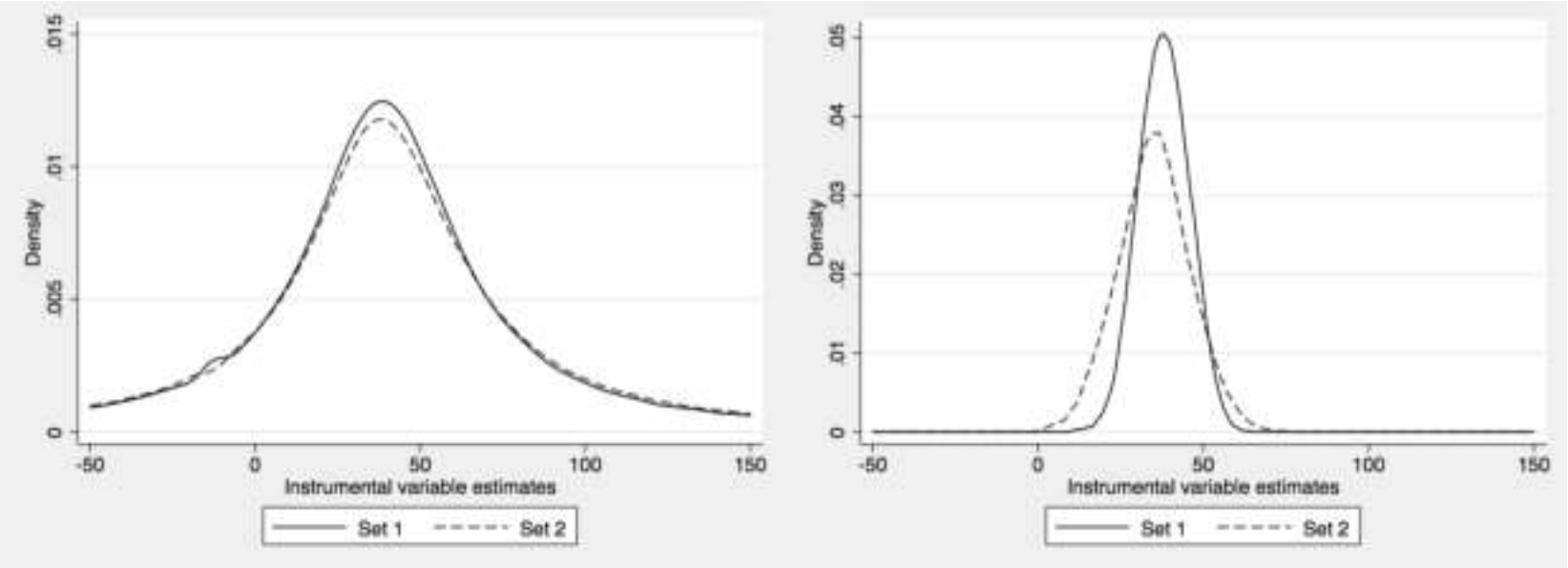
Example: using SNPs as multiple instruments in TSLS

- Sargan over-id test $p = 0.011$
- and 12% of paired differences exclude the null
- But paired differences centred on zero (2.5, 97.5 centiles: -27.0, 28.3)

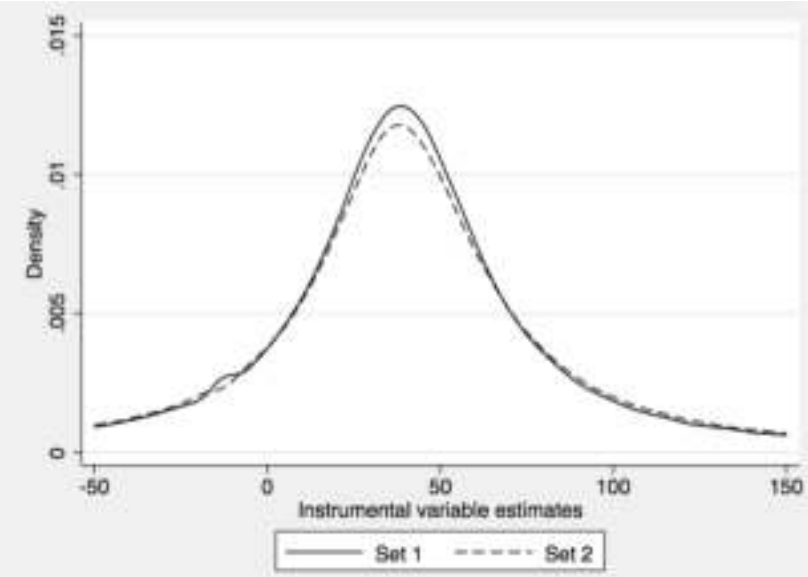
Example: using SNPs as allele count in TSLS

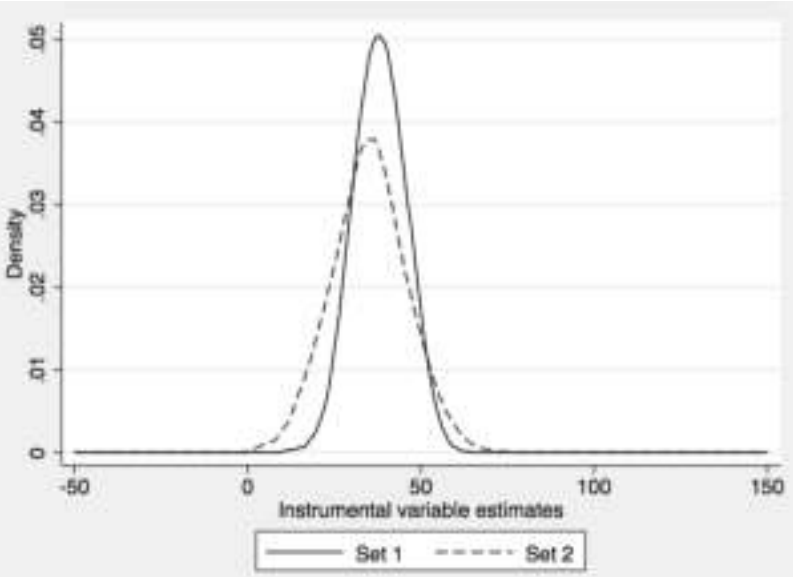


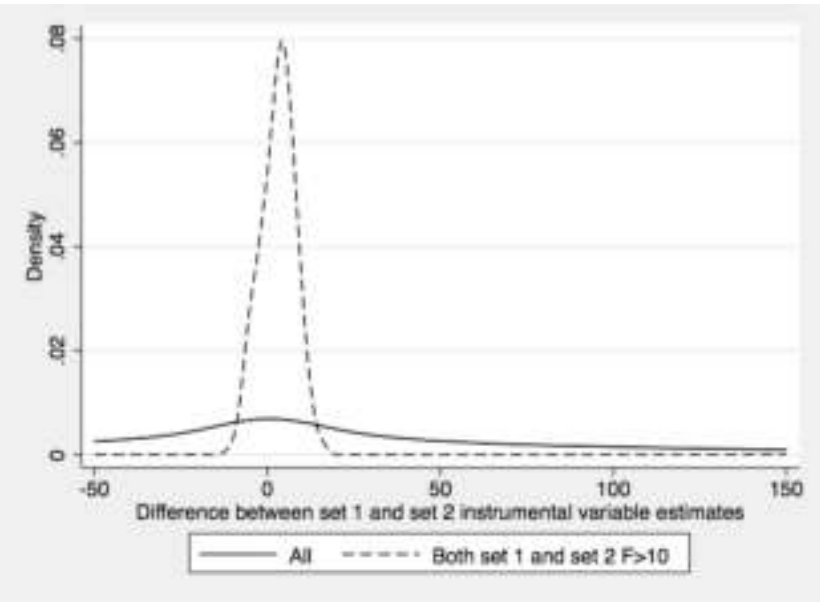
Example: using SNPs as allele count in TSLS



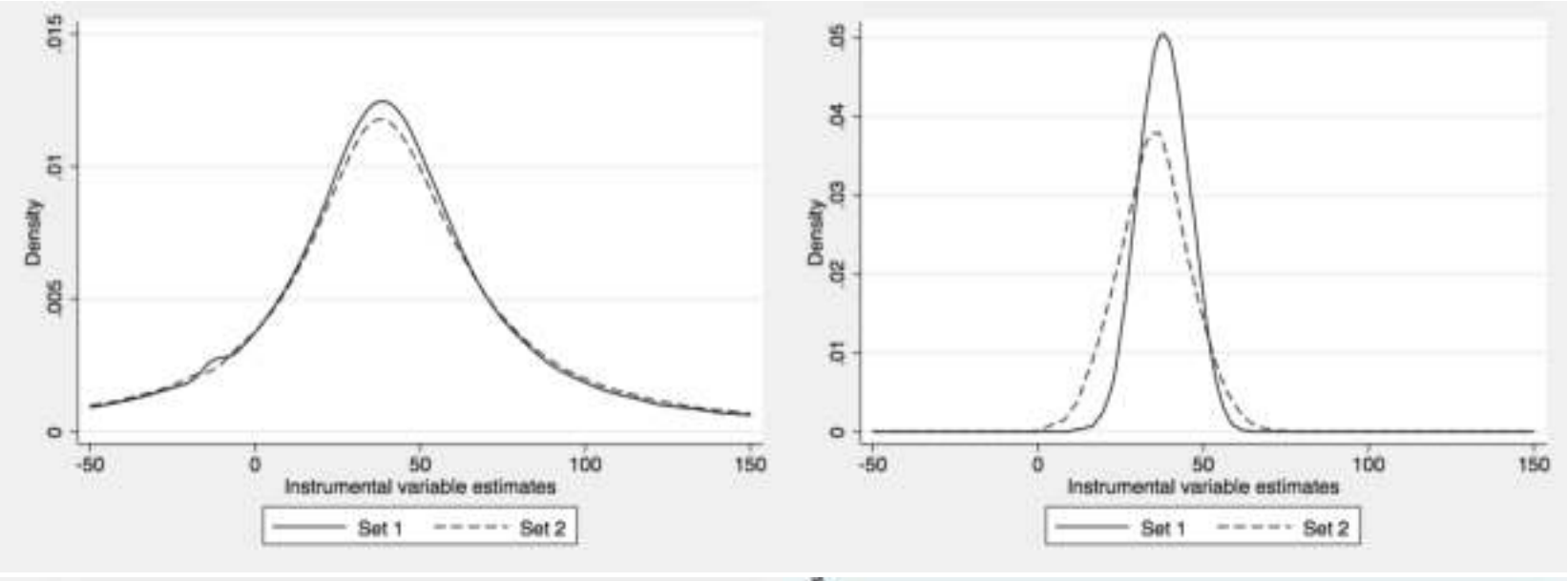
Example: using SNPs as allele count in TSLS







Example: using SNPs as allele count in TSLS



0.07% of 95% CIs exclude 0 (all differences), 0% exclude 0 (both set 1 & 2 $F > 10$).

Han's algorithm

- Han (2008) defined median of LATEs as robust L_1 GMM estimator
  Median of 20 separate instruments = 24.2
- Also proposed an algorithm to select instruments based on over-id test p-values
- Using $p = 0.05$ algorithm selects 15 of the 20 instruments; IV estimate = 36.8 (95% CI 26.3, 47.2); Sargan p=0.173

Summary

- Lack of identification in SMMs:
  - Don't just rely on `gmm` or whatever software you're using
  - Plot the estimating equation for different values of ψ when fitting SMMs
  - We found SMMs with 0, 1, and 2 solutions
- Multiple paired comparisons
  - Watch out for `ivregress`/`ivreg2` dropping the constant from 2^{nd} stage model
  - Dichotomy between over-id test results and distribution of the paired differences