Mathematics & Statistics | Lancaster University

Corrected standard errors for two-stage residual inclusion estimators and a Stata package for MR-Egger regression type analyses

Leibniz Institute for Prevention Research and Epidemiology – BIPS

20th April 2017

Dr Tom Palmer

Outline

Mathematics & Statistics | Lancaster University

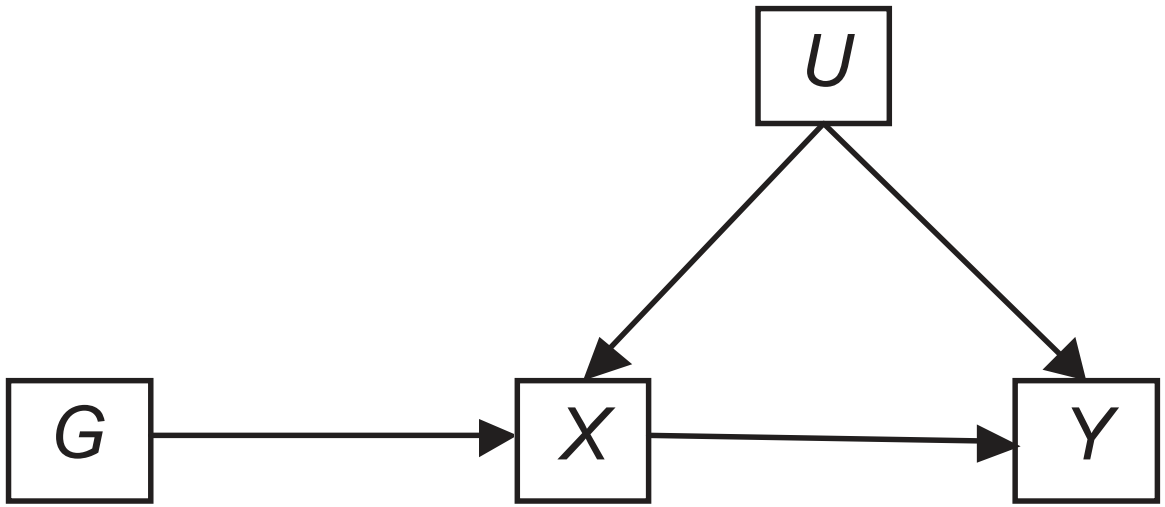
- Corrected SEs for two-stage residual inclusion (TSRI) estimators (individual patient data)
- `mrrobust`: A Stata package for MR-Egger type regression models (summary data)

Part One

Correcting standard errors for two-stage residual inclusion estimators

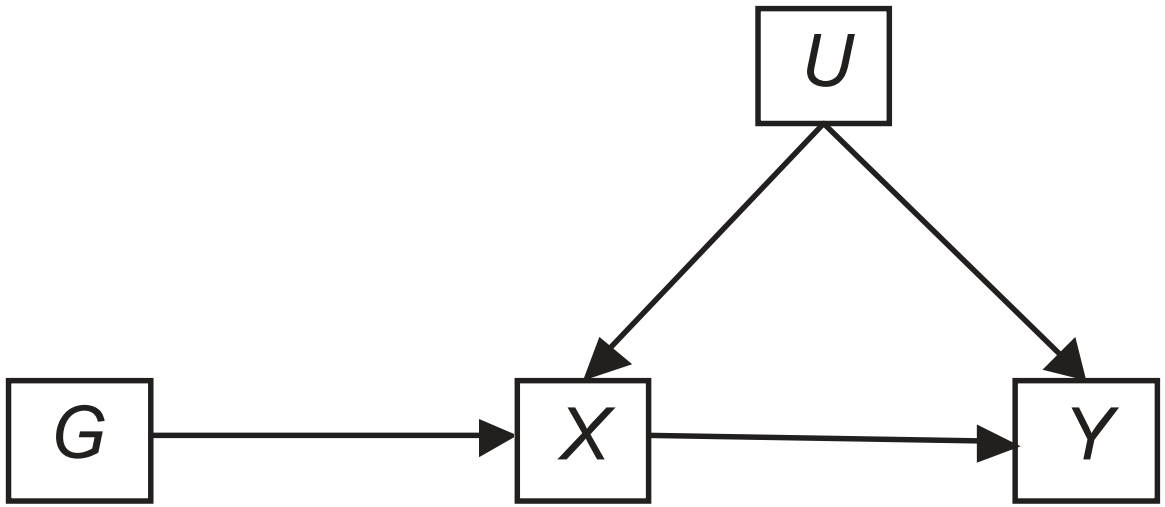
- Mendelian randomization – genotypes as instrumental variables.

DAG

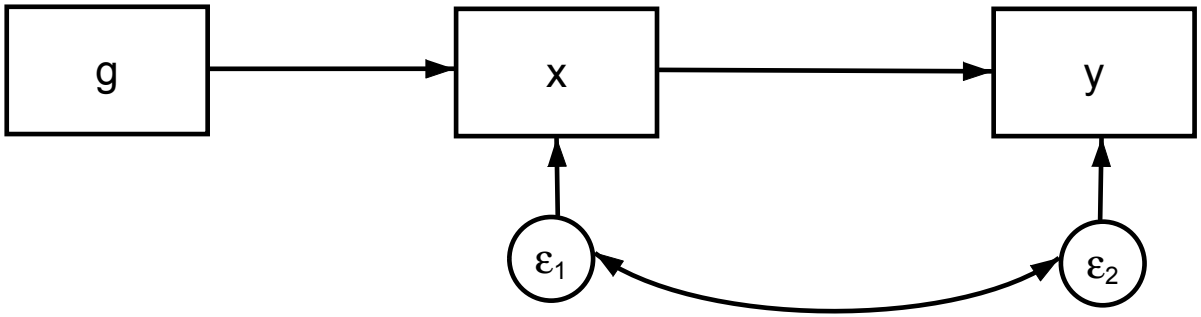


- Mendelian randomization – genotypes as instrumental variables.

DAG



Structural Equation Model Path Diagram



$$\text{Stage 1:} \quad X = \alpha_0 + \alpha_1 Z + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2)$$

$$\text{Stage 2:} \quad h(E[Y]) = \beta_0 + \beta_1 \widehat{X}$$

TSRI

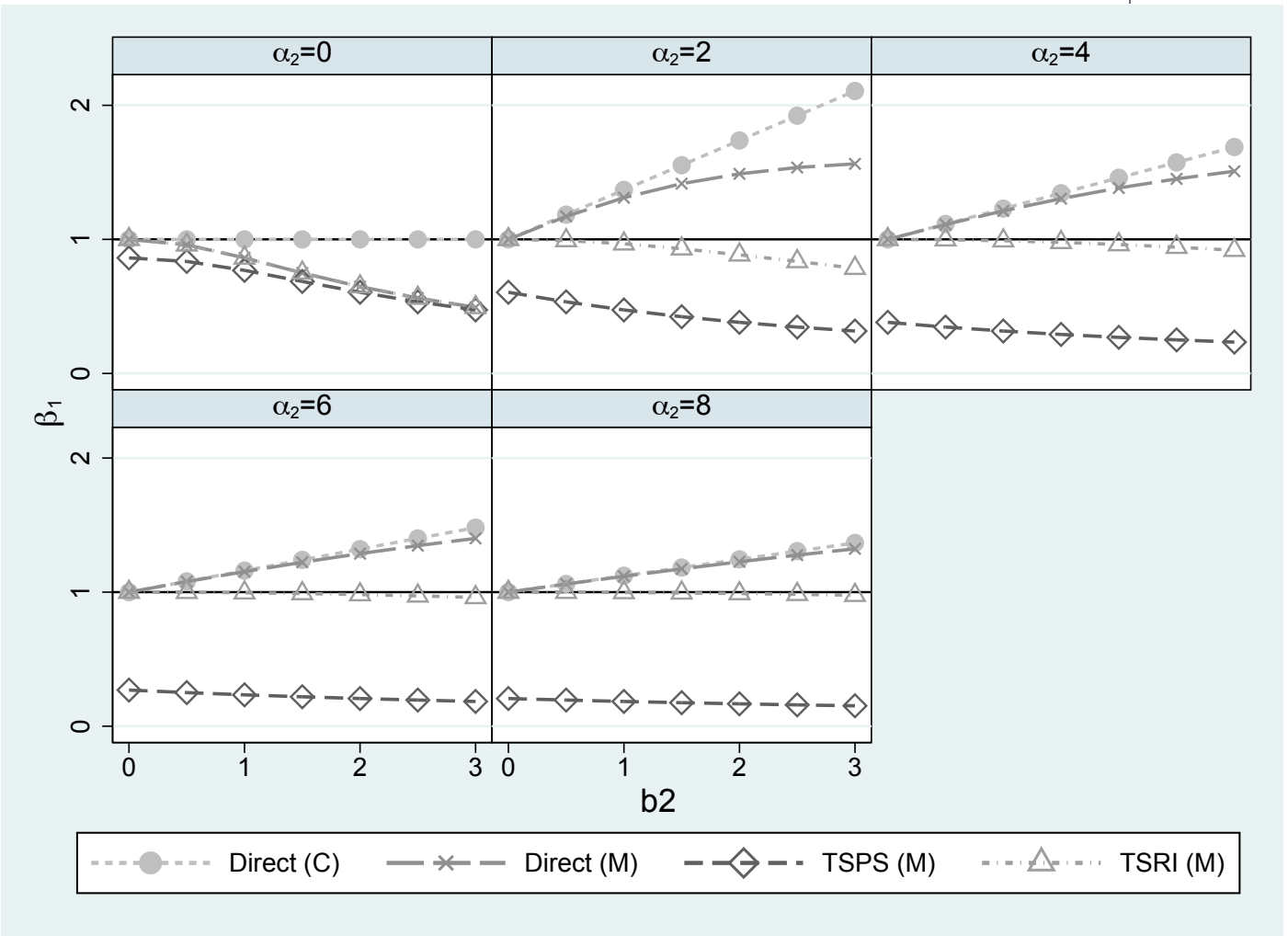
$$\text{Stage 2:} \quad h(E[Y]) = \beta_0 + \beta_1 X + \beta_2 \widehat{\varepsilon}_1.$$

- If model holds test of $\widehat{\beta}_2 = 0$ is Hausman endogenity test.
- If h() is identity link then $\widehat{\beta}_{\text{TSPS}} = \widehat{\beta}_{\text{TSRI}}$
- When h() is a non-collapsible link function $\widehat{\beta}_{\text{TSPS}} \neq \widehat{\beta}_{\text{TSRI}}$
- When h() is a non-collapsible link function marginal and conditional parameter estimates not equal

Marginal parameter values for the logistic TSPS and TSRI estimators

$$\beta_{1m} = \beta_1 \frac{1}{\sqrt{1 + c^2 V}}, \quad \text{where } c = \frac{16\sqrt{3}}{15\pi}.$$

Different estimators require different Vs.



Web Figure 4: Values of the marginal (M) and conditional (C) parameters of the direct logistic regression of Y on X, logistic two-stage predictor substitution (TSPS), and logistic two-stage residual inclusion (TSRI) estimators used in the simulations.

- If fitting TSPS and TSRI manually we need to be aware that the second stage SEs will not be correct.
- Intuitively we need to incorporate uncertainty from both stages of estimation

The difference between unadjusted and corrected standard errors for TSRI estimators – linear outcome model

$$
\begin{aligned}
g_i &\sim \text{Binomial}(2,\ p_g) \\
x_i &= \alpha_0 + \alpha_1 g_i + \varepsilon_{1i}, \\
y_i &= \beta_0 + \beta_1 x_i + \varepsilon_{2i},
\end{aligned}
\qquad
\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \text{MVN}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)
$$

$$
\begin{aligned}
x_i &= \alpha_0 + \alpha_1 g_i + \varepsilon_{3i} \\
y_i &= \beta_0 + \beta_1 \widehat{x}_i + \varepsilon_{4i}
\end{aligned}
$$

After the second stage of manual estimation we have

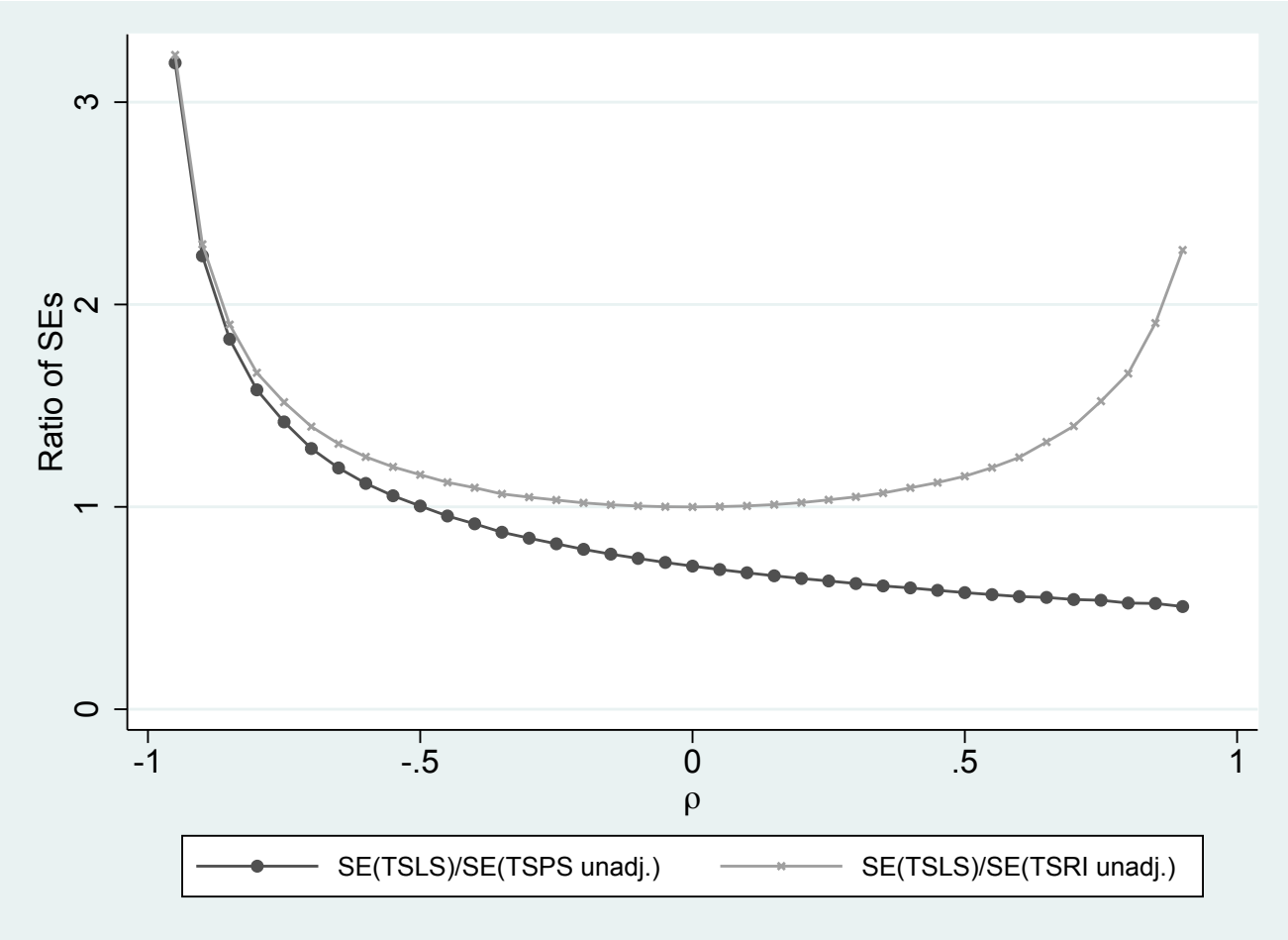
$$\text{var}(\widehat{\beta}) = s^2(\widehat{X}'\widehat{X})^{-1}$$
$$\text{where } s^2 = \frac{\sum_{i=1}^{N}(Y - \widehat{X}\widehat{\beta})^2}{(N-k)}.$$

However the causal model is in terms of X not $\widehat{X}$, and so we need

$$s^2 = \frac{\sum_{i=1}^{N}(Y - X\widehat{\beta})^2}{N}.$$

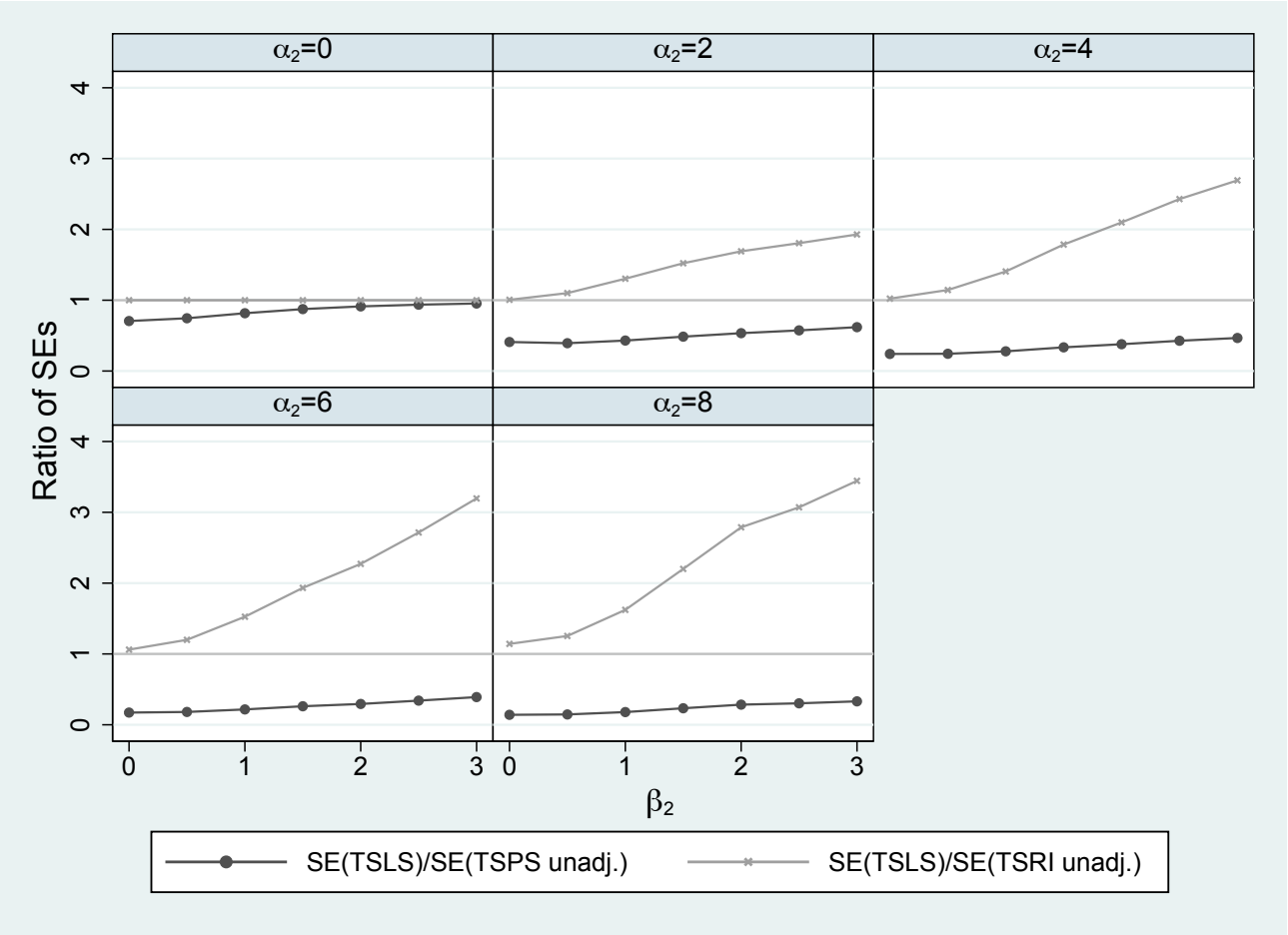
Correction implemented in `tsls()` function in R `sem` package

Ratio of TSLS SEs to unadjusted TSPS and TSRI SEs



(a) Average SEs in simulations with N=1 000 using 50 replications.

Ratio of TSLS SEs to unadjusted TSPS and TSRI SEs



(b) Average SEs in the linear simulations with N=1 000.

Will consider several methods of correcting the SE

- BS 1: Bootstrapping second stage
- BS 2: Bootstrapping both first and second stages jointly
- Newey: method proposed in context of probit regression
- Terza 1 & 2: Recently proposed analytical correction
- Researchers currently often specify heteroskedasticity robust SEs at second stage - as think this may be an informal correction.

Applied example: effect of BMI on SBP and Diabetes

- 17057 participants from 6 propestive cohorts
- Externally weighted allele score was constructed out of the variants for BMI
- Continuous outcome: Systolic blood pressure (SBP; mmHg)
- Binary outcome: Diabetes (sample prevalence 14%)

Odds ratios for Diabetes

Table 1: Estimates of the causal odds ratios for diabetes for a one unit increase in body mass index across 6 cohorts ARIC, CHS, CARDIA, FHS, MEDAL, and MESA (All N=17 057).

Estimator	SE (log OR scale)	z	OR	95% CI
Direct logistic	0.004	29.6	1.14	1.13, 1.15
Logistic TSPS (Stage 1: F=119, R^2=0.007)	0.056	4.96	1.32	1.19, 1.48
Logistic TSRI (unadjusted SE)	0.058	4.79	1.32	1.18, 1.48
Logistic TSRI (robust SE)	0.057	4.86	1.32	1.18, 1.47
Logistic TSRI (TSPS unadjusted SE)	0.056	4.96	1.32	1.18, 1.47
Logistic TSRI (BS 1)	0.057	4.80	1.32	1.18, 1.48
Logistic TSRI (BS 2)	0.061	4.50	1.32	1.17, 1.49
Logistic TSRI (Newey SE)	0.059	4.71	1.32	1.17, 1.48
Logistic TSRI (Terza SE 1)	0.057	4.83	1.32	1.18, 1.47
Logistic TSRI (Terza SE 2)	0.059	4.77	1.32	1.18, 1.48
Logistic SMM	0.101	3.26	1.39	1.19, 1.59
Probit TSRI (on OR scale)	0.090	4.74	1.28	1.15, 1.42

SEs given on log odds ratio scale. Bootstrapping using 500 replications (BS: bootstrap, CI: confidence interval, IV: instrumental variable, OR: odds ratio, SE: standard error, SMM: structural mean model, TSPS: two-stage predictor substitution, TSRI: two-stage residual inclusion).

Estimates for SBP

Table 2: Estimates of the causal effect of a one unit increase in body mass index on systolic blood pressure (mmHg) across 6 cohorts ARIC, CHS, CARDIA, FHS, MEDAL, and MESA (All N=17 057).

Estimator	SE	Estimate	95% CI
Direct linear	0.031	0.76	0.70, 0.82
TSLS (Stage 1: F=119, R^2=0.007)	0.374	0.36	-0.37, 1.10
TSPS (unadjusted SE)	0.378	0.36	-0.38, 1.11
Linear TSRI (unadjusted SE)	0.372	0.36	-0.37, 1.09
Linear TSRI (robust SE)	0.370	0.36	-0.36, 1.09
Linear TSRI (TSPS unadjusted SE)	0.378	0.36	-0.38, 1.11
Linear TSRI (BS 1 SE)	0.376	0.36	-0.37, 1.10
Linear TSRI (BS 2 SE)	0.384	0.36	-0.39, 1.12
Linear TSRI (Newey SE)	0.374	0.36	-0.37, 1.10
Linear TSRI (Terza SE 1)	0.370	0.36	-0.36, 1.09
Linear TSRI (Terza SE 2)	0.372	0.36	-0.37, 1.09

Bootstrapping using 500 replications (BS: bootstrap, CI: confidence interval, SE: standard error, TSLS: two-stage least squares, TSRI: two-stage residual inclusion).

Binary outcome simulations

$$g_i \sim \text{Binomial}(2,\ 0.3)$$

$$u_i \sim N(0,1) \text{ – representing the unmeasured confounding,}$$

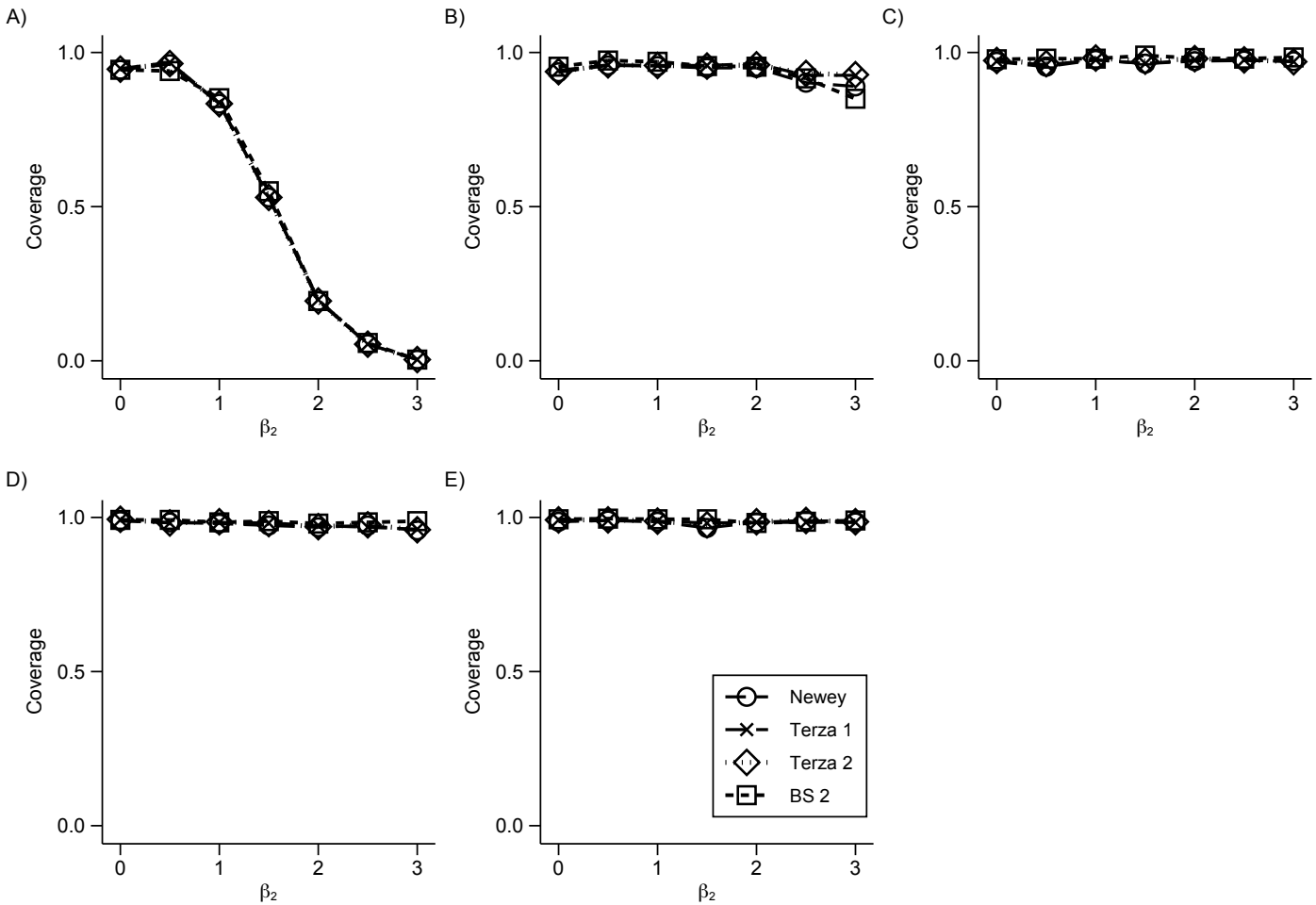
$$x_i \sim \alpha_0 + \alpha_1 g_i + \alpha_2 u_i + \varepsilon_{1i},\ \varepsilon_{1i} \sim N(0,1)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 u_i$$

$$y_i \sim \text{Binomial}(1,\ p_i)$$

$$\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = \{0,2,4,6,8\}, \beta_0 = \log(0.05/0.95), \beta_1 = 1, \beta_2 = [0,3]$$

Coverage of the logistic TSRI estimators for N = 1000 with respect to the conditional parameter, $\beta_1 = 1$. Panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8



Coverage of the logistic TSRI estimators for N = 1000 with respect to the marginal parameter. Panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

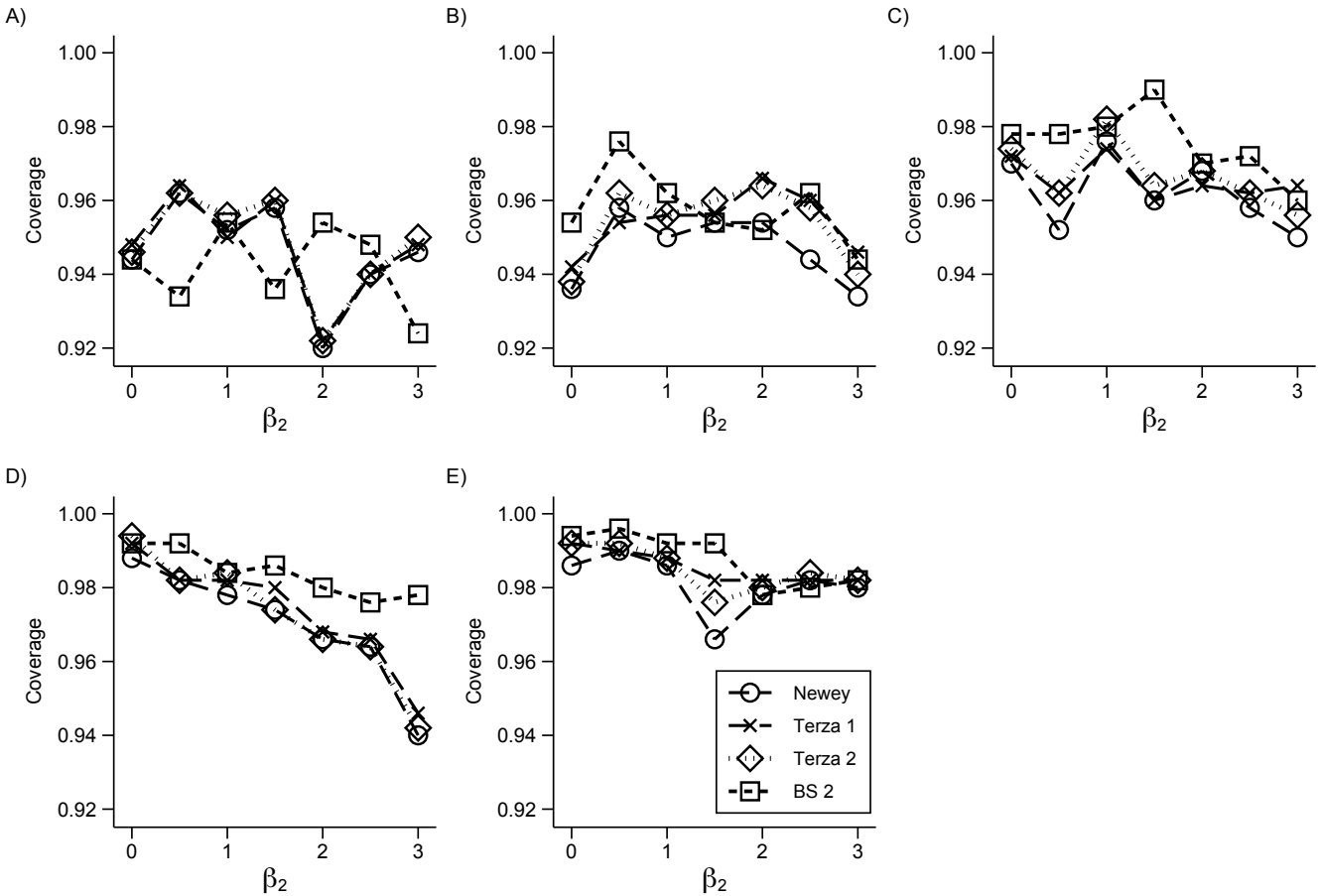


Figure 2

Type I error of the logistic TSRI estimators for N = 1000. Panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

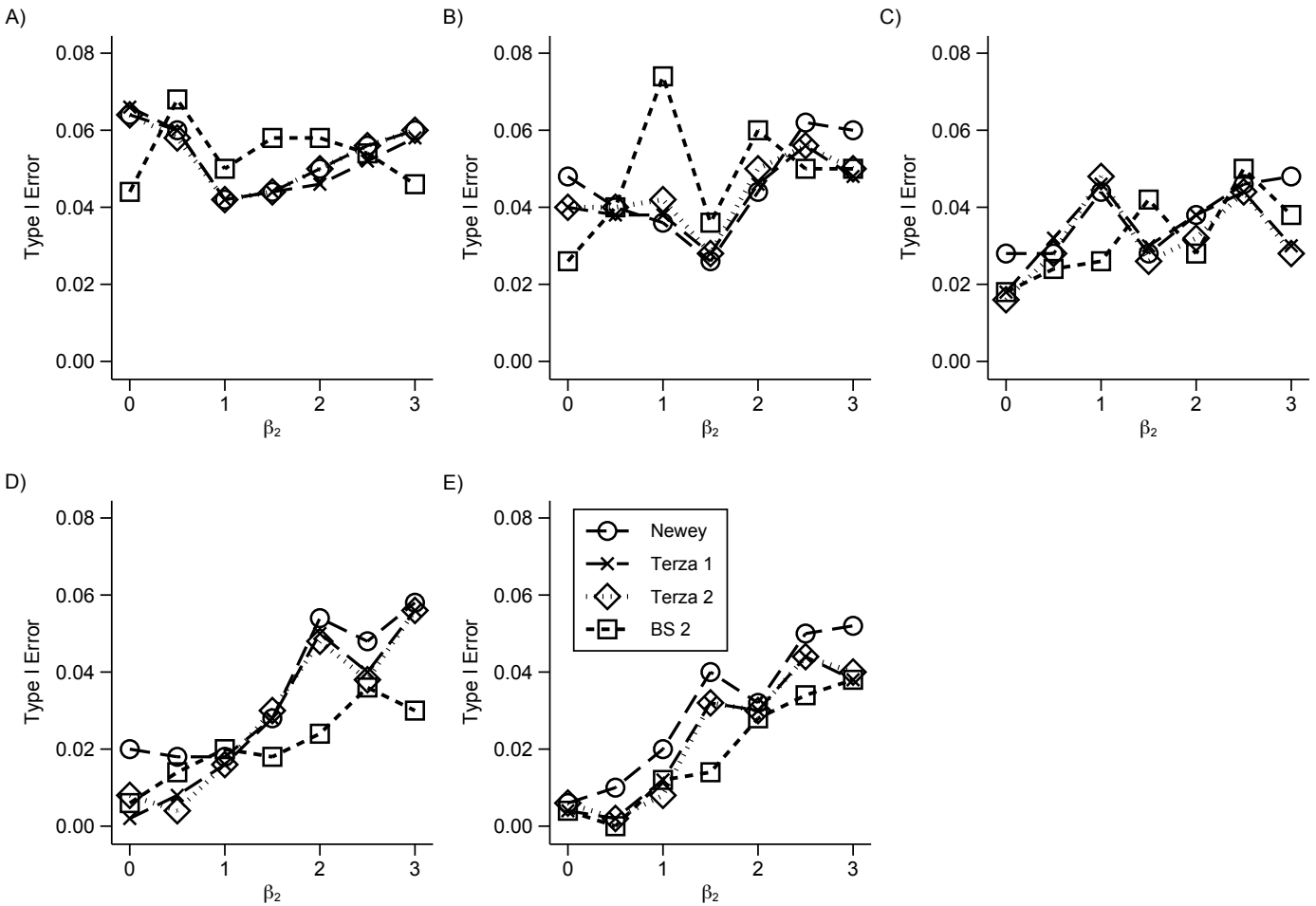


Figure 3

Continuous outcome simulations

$$y_i \sim \beta_0 + \beta_1 x_i + \beta_2 u_i + \varepsilon_{2i},\ \varepsilon_{2i} \sim N(0,1)$$

$$\alpha_0 = 0,\ \alpha_1 = 1,\ \alpha_2 = \{0,2,4,6,8\},\ \beta_0 = 0,\ \beta_1 = 1,\ \beta_2 = [0,3]$$

Coverage of the linear TSRI estimators for N = 1 000. Panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

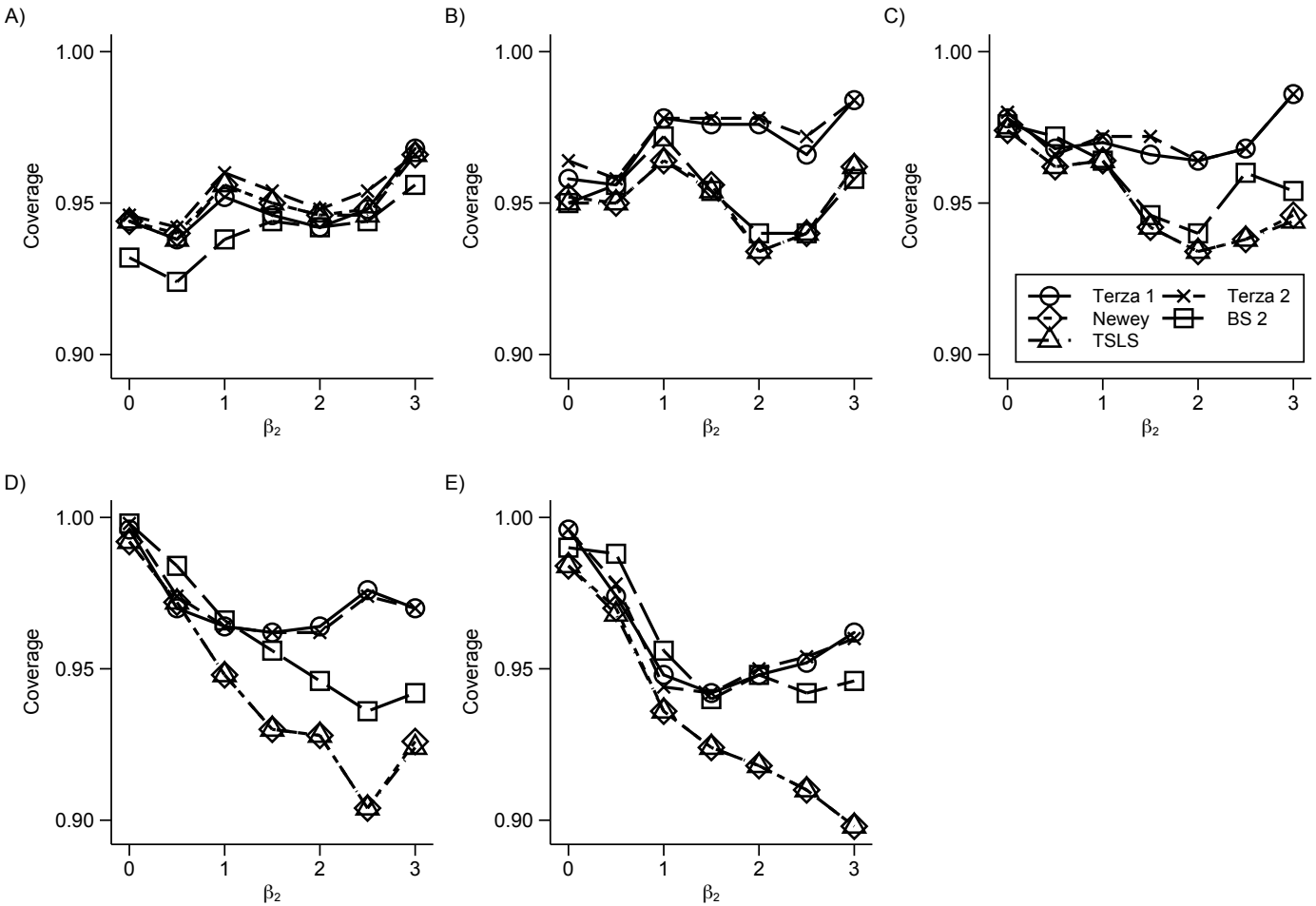


Figure 4

Type I error of the linear TSRI estimators for N = 1000. Panels correspond to α_2 being set to the following values A:0, B:2, C:4, D:6, and E:8.

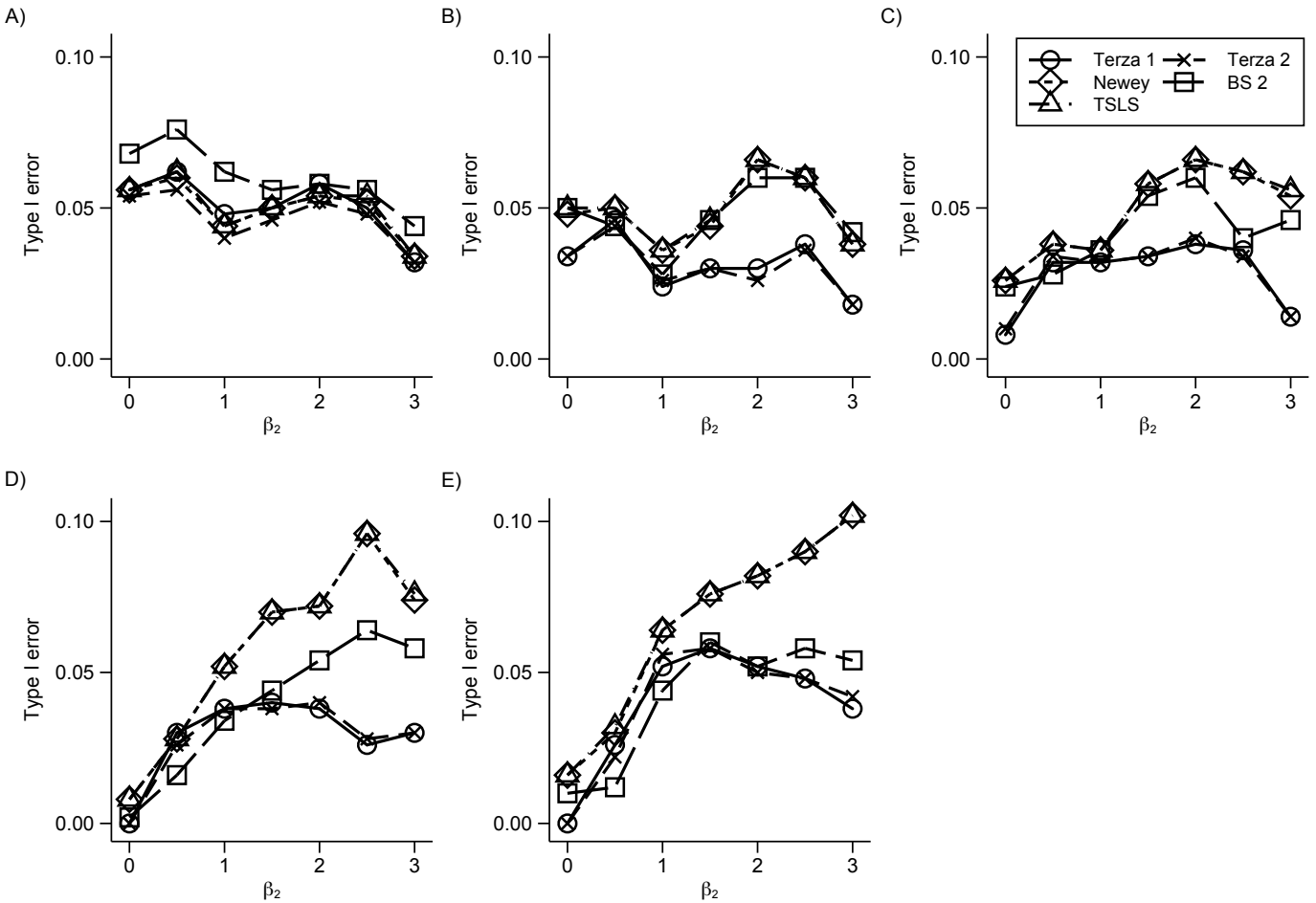


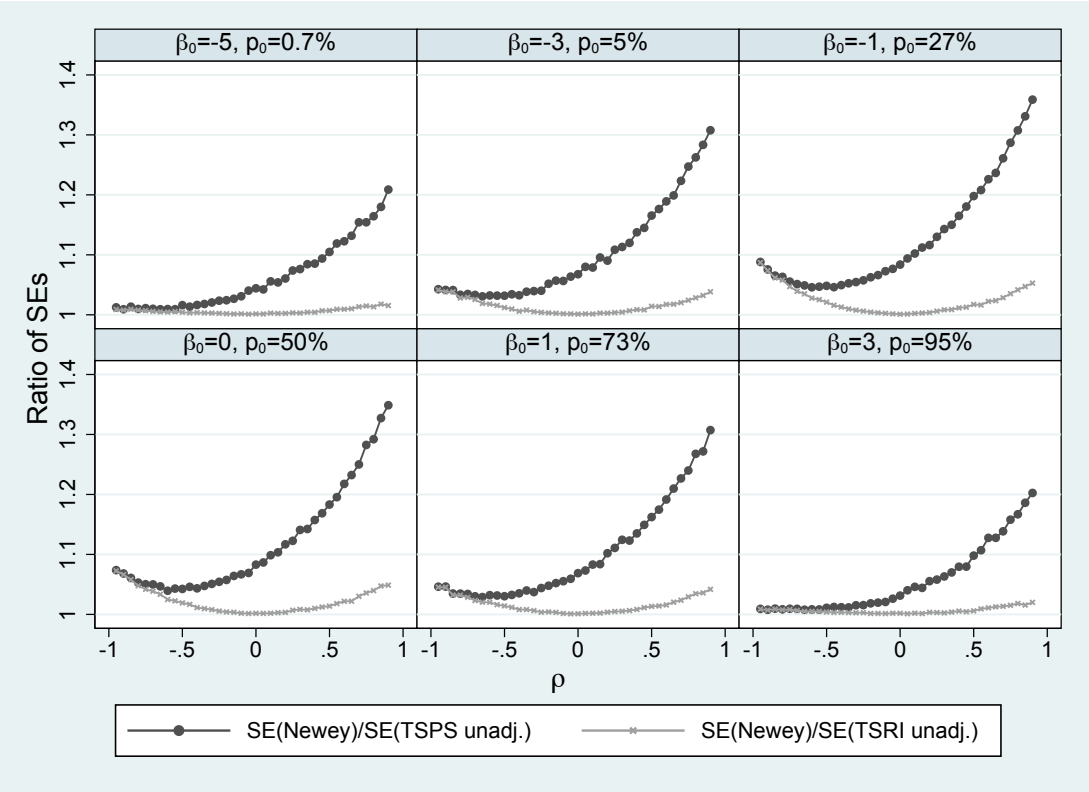
Figure 5

Comparing corrected and unadjusted SEs for logistic estimators

$$
\begin{aligned}
g_i &\sim \text{Binomial}(2,\ p_g) \\
x_i &= \alpha_0 + \alpha_1 g_i + \varepsilon_{1i}, \\
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i + \varepsilon_{2i}, \qquad \begin{pmatrix}\varepsilon_1 \\ \varepsilon_2\end{pmatrix} \sim \text{MVN}\left(0, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right) \\
y_i &\sim \text{Bernoulli}(p_i).
\end{aligned}
$$

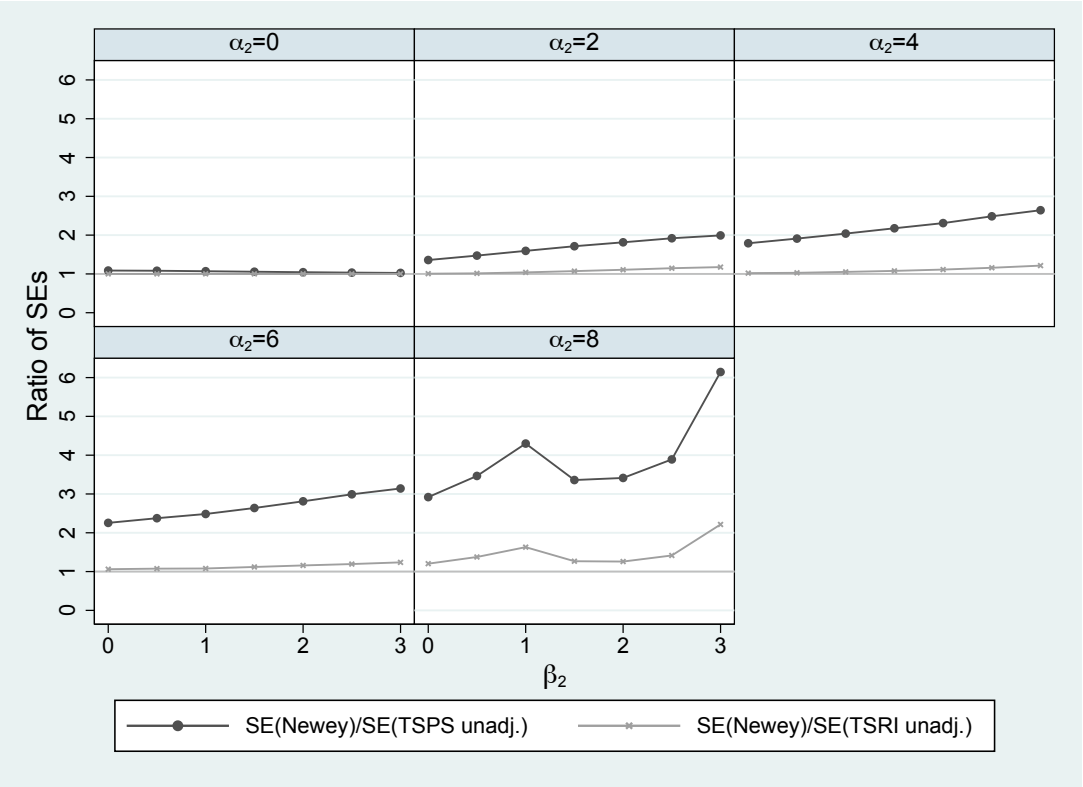
$$
\begin{aligned}
\text{var}(\widehat{\beta}) &= (X'VX)^{-1} \\
V &= I_N \circ \text{diag}(\widehat{p}(1-\widehat{p})').
\end{aligned}
$$

Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs. These are average SEs in simulations with N=1000 using 50 replications.



Web Figure 2: Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs. These are average SEs in simulations with N=1 000 using 50 replications (SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs in the logistic simulations with N=1000.



Web Figure 3: Ratio of logistic TSRI Newey SEs to unadjusted logistic TSPS and TSRI SEs in the logistic simulations with N=1 000 (SE: standard error; TSLS: two-stage least squares; TSPS: two-stage predictor substitution; TSRI: two-stage residual inclusion).

Part Two

`mrrobust`: A Stata package for MR-Egger type regression models

- `http://www.mrbase.org`
- Hemani et al. The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv.
- Two-sample Mendelian randomization
- Single genotype:

$$\beta = \frac{\text{genotype-disease: sample 1}}{\text{genotype-phenotype: sample 2}}$$

Multiple genotypes – Inverse variance weighted regression:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{J} w_j \hat{\beta}_j}{\sum_{j=1}^{J} w_j}, w_j = \frac{\hat{\gamma}_j^2}{\sigma_{Yj}^2}$$

Multiple genotypes – MR-Egger regression (Bowden et al., IJE, 2015)

Assumptions:

- INstrument Strength Independent of Direct Effect (InSIDE) – instrument-exposure and pleiotropic association parameters independent.
- Under InSIDE, estimates for variants with stronger instrument-exposure associations $\widehat{\gamma}_j$ will be closer to the true causal effect parameter than variants with weaker associations.
- NO Measurement Error (NOME) – requires no measurement error to be present in the instrument-exposure associations. This allows the variance in the set of variants J to be estimated as $\text{var}(\widehat{\beta}_j) = \frac{\sigma^2_{Yj}}{\widehat{\gamma}_j}$.

Multiple genotypes – MR-Egger regression (Bowden et al., IJE, 2015)

Γ_j: genotype-disease coefficients

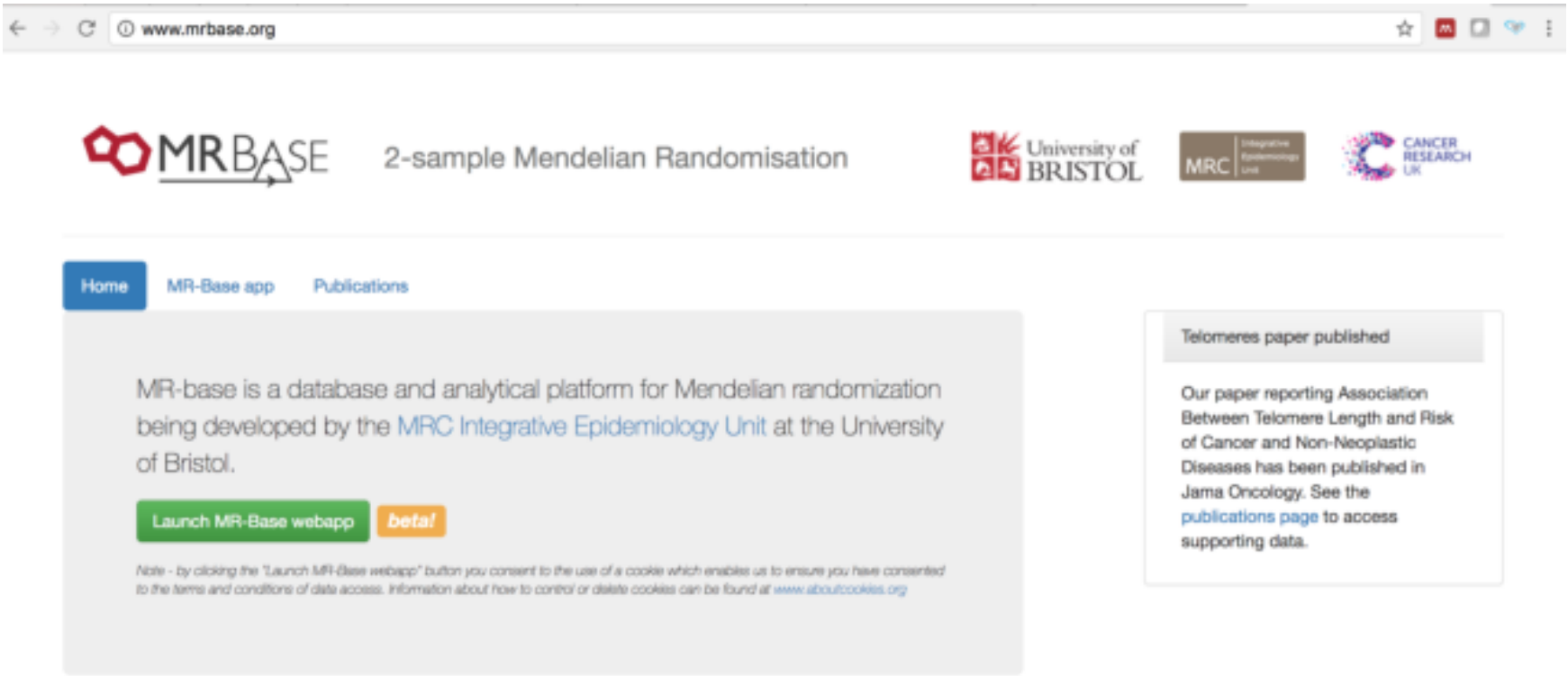
γ_j: genotype-phenotype coefficients

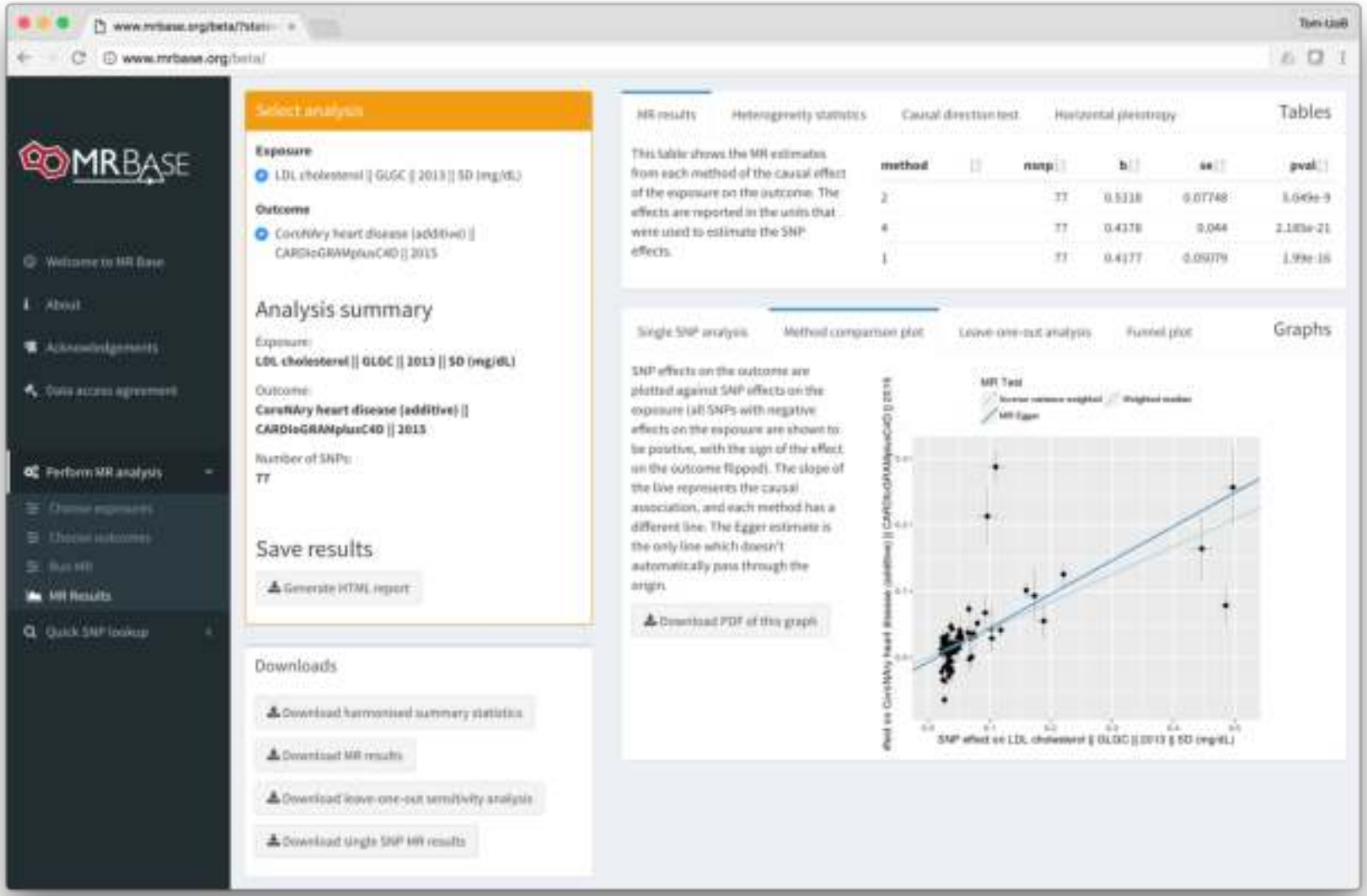
$$\widehat{\Gamma}_j = \beta_0 + \beta_1 \widehat{\gamma}_j + \varepsilon_j,\ \varepsilon_j \sim N(0,\ \sigma^2) \text{ weighted by } \sigma_{yj}^{-2}$$

- MR-Egger intercept: average directional pleiotropic effect across the set of variants
- MR-Egger slope: corrected causal effect estimate

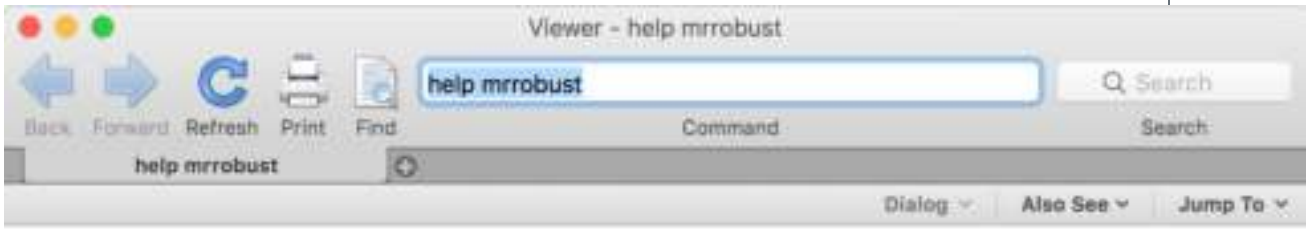
Software implementations

- `MendelianRandomization` R package on CRAN (Yavorska & Burgess, IJE, 2017)
- `TwoSampleMR` R package related to MR-Base (`https://mrcieu.github.io/TwoSampleMR`)


www.mrbase.org
MRBASE
2-sample Mendelian Randomisation
University of BRISTOL
MRC
CANCER RESEARCH UK
Home
MR-Base app
Publications
MR-base is a database and analytical platform for Mendelian randomization being developed by the MRC Integrative Epidemiology Unit at the University of Bristol.
Launch MR-Base webapp
beta!
Note - by clicking the "Launch MR-Base webapp" button you consent to the use of a cookie which enables us to ensure you have consented to the terms and conditions of data access. Information about how to control or delete cookies can be found at www.aboutcookies.org
Telomeres paper published
Our paper reporting Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases has been published in Jama Oncology. See the publications page to access supporting data.
Citation
Gibran Hemani, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Tan, James Yarmolinsky, Hashem A. Shihab, Nicholas Timpson, David M Evans, Caroline Relton, Richard M Martin, George Davey Smith, Tom R Gaunt, Philip C Haycock, The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv. doi: https://doi.org/10.1101/078972


www.mrbase.org/beta/
MRBASE
Welcome to MR Base
About
Acknowledgements
Data access agreement
Perform MR analysis
Choose exposures
Choose outcomes
Run MR
MR Results
Quick SNP lookup
Select analysis
Exposure
LDL cholesterol || GLGC || 2013 || SD (mg/dL)
Outcome
CoroNAry heart disease (additive) || CARDIoGRAMplusC4D || 2015
Analysis summary
Exposure:
LDL cholesterol || GLGC || 2013 || SD (mg/dL)
Outcome:
CoroNAry heart disease (additive) || CARDIoGRAMplusC4D || 2015
Number of SNPs:
77
Save results
Generate HTML report
Downloads
Download harmonised summary statistics
Download MR results
Download leave-one-out sensitivity analysis
Download single SNP MR results
MR results
Heterogeneity statistics
Causal direction test
Horizontal pleiotropy
Tables
This table shows the MR estimates from each method of the causal effect of the exposure on the outcome. The effects are reported in the units that were used to estimate the SNP effects.
method
nsnp
b
se
pval
Single SNP analysis
Method comparison plot
Leave-one-out analysis
Funnel plot
Graphs
SNP effects on the outcome are plotted against SNP effects on the exposure (all SNPs with negative effects on the exposure are shown to be positive, with the sign of the effect on the outcome flipped). The slope of the line represents the causal association, and each method has a different line. The Egger estimate is the only line which doesn't automatically pass through the origin.
Download PDF of this graph
MR Test
MR Egger

- `mrrobust` Stata package:
  - IVW and MR-Egger regression approaches, including fixed effects MR-Egger regression, standard error correction, and weighting options.
  - Unweighted, weighted and penalized weighted median IV estimators, providing pleiotropy robust estimates in cases where fewer than 50% of the genetic instruments are invalid.
  - Presentation of heterogeneity statistics, and statistics such as I^2_{GX} for use in assessing attenuation bias.
  - Plotting tools to visualise IVW, MR-Egger, and weighted median estimators.
  - Illustrative examples and documentation using data from Do et al. Nat Gen, 2013.



Title

mrrobust — Suite of commands implementing estimators robust to certain proportions of invalid instrumental variables which are becoming commonly applied in Mendelian randomization (MR) studies.

Commands

mregger MR-Egger and inverse-variance weighted (IVW) estimators.

mrmedian Unweighted, weighted, and penalized weighted median estimators for summary level data.

mrmedianobs Unweighted, weighted, and penalized weighted median estimators for individual level data.

mreggerplot Scatter plot showing instrument specific estimates with IVW, MR-Egger, or median fitted line and confidence interval.

Description

mrrobust is a suite of programs imlementing recently developed estimators which are robust to certain proportions of invalid instrumental variables.

The estimators were developed in the context of MR studies but could be used in other studies using instrumental variables.

I^2_{GX} statistic

- NOME violated - individual variants suffer from weak instrument bias – attenuation of MR Egger estimates to the null.
- Assess NOME assumption with I^2_{GX} statistic, Bowden et al., IJE, 2016. $I^2_{GX} = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + s^2}$; $\beta_E = \beta I^2_{GX}$

  σ^2_γ: variance of true genotype-exposure associations

  s^2: additional variability among $\widehat{\gamma}_j$ (due to ME)
- Degree of attenuation bias in the MR Egger estimate.
- I^2_{GX} of 0.7 represents an estimated relative bias of 30% towards the null.
- I^2_{GX} low use SIMEX or Bayesian error in variables methods.

Applied Example: Adiposity and Rheumatoid Arthritis

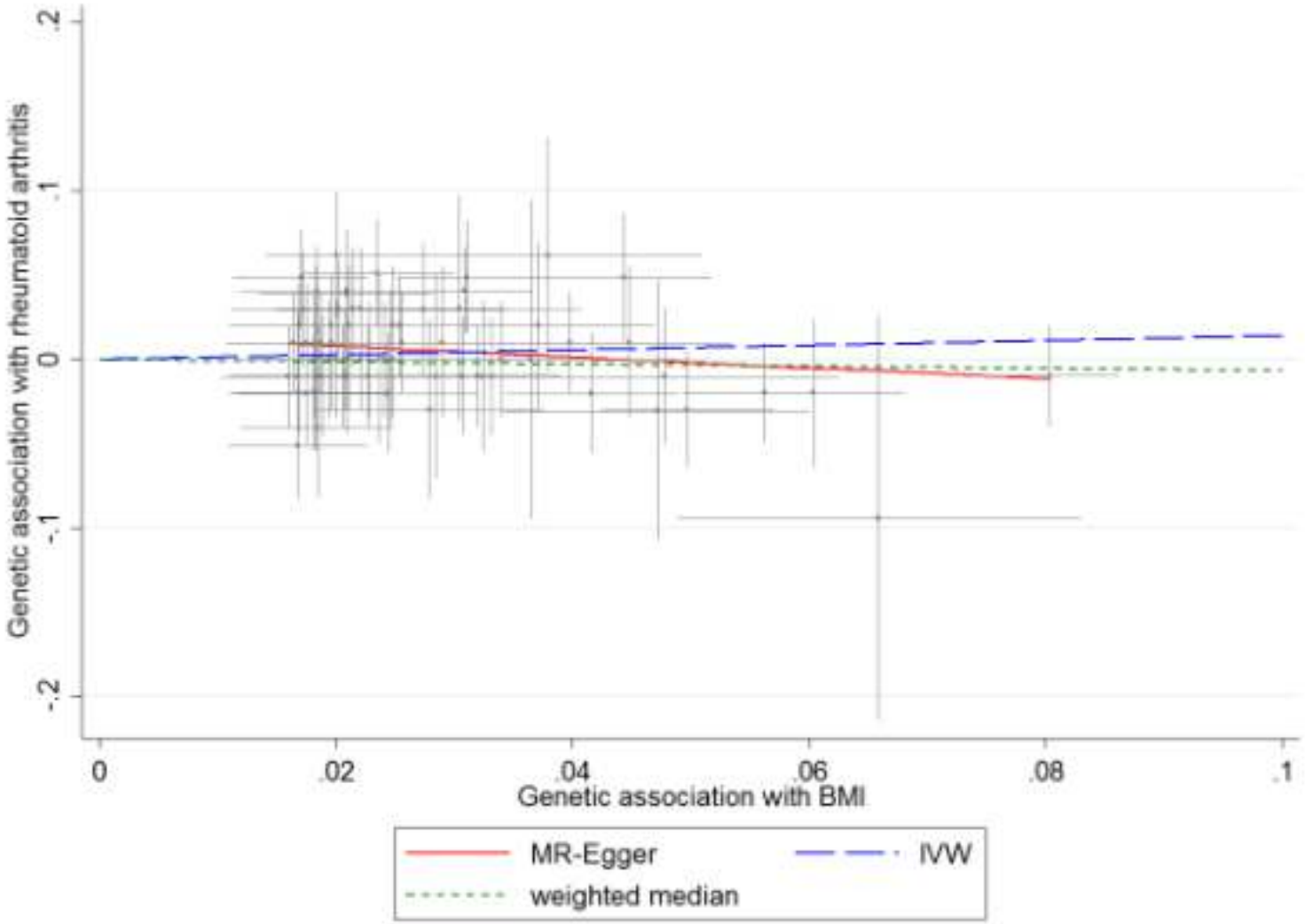
Table 1: Summary MR estimates for the effect of BMI upon rheumatoid arthritis on the odds ratio scale.

	Estimate	*SE*	*p-value*	*95% CI*
IVW				
Effect	1.107	0.111[a]	0.335	0.91, 1.35
MR Egger				
Intercept	1.011	0.006[a]	0.067	1.00, 1.27
Effect	0.773[b]	0.180	0.207	0.49, 1.22
Weighted Median				
Effect	0.916	0.121[c]	0.489	0.70, 1.17

a) Delta method *b)* $I^2_{GX} = 0.92$ *c) Bootstrap*

Median estimators are reliant upon the proportion of valid instruments being greater than 50%.

Figure 1: Scatterplot showing ratio estimates for each variant and MR estimates.



Summary

- Corrected SEs for TSRI estimators
  - Bootstrap both stages, Newey, or Terza SEs have best properties
  - Avoid using heteroskedasticity robust SEs second stage
- `mrrobust` Stata package:
  - IVW, MR-Egger, Median sensitivity analysis models

Acknowledgements
TSRI: Nuala Sheehan, Michael Holmes, and Brendan Keating.
`mrrobust`: Wesley Spiller and Neil Davies.

Thanks for your attention.

Any questions?