Lancaster University

`mrrobust`: a Stata package for MR-Egger regression type analyses
London Stata User Group Meeting 2017

8th September 2017
Tom Palmer   Wesley Spiller   Neil Davies

Outline

Lancaster University

- Introduction
- GitHub and installation
- Worked example
- Stata wishes
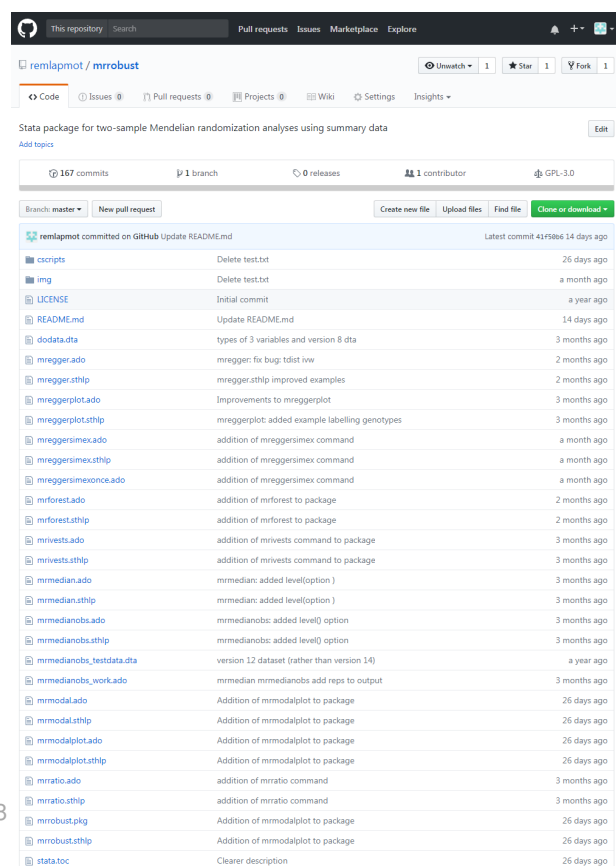- Discussion

Introduction

- Mendelian randomization: instrumental variable analysis using genotypes as instruments in epidemiology (Davey Smith, 2003)
- Researchers do still work on individual level data (`ivreg2`)
- However so much summary data now available from GWAS that researchers mainly fitting summary data estimators (IVW, MR-Egger, median, modal)
- This package implements several of these methods.
- R packages:
  - `MendelianRandomization` package (Yavorska & Burgess, 2017)
  - `TwoSampleMR` package, companion to MR-Base
    https://mrcieu.github.io/TwoSampleMR
    http://www.mrbase.org

GitHub repository

https://github.com/remlapmot/mrrobust

- `parallel` package
- Based on `git` (Linus Torvalds)
- GitHub – excellent for projects with a small no. collaborators
- `master` branch; make new feature in a new branch - merge into master when ready
- To help someone else: fork repo - new feature in new branch - send pull request

GitHub `README.md`

- Every repo has a README.md - can do alot with this
- I include installation instructions and link to a short video

The screenshot shows the mrrobust README:

mrrobust: Stata package for two-sample Mendelian randomization analyses using summary data

Short video introduction

Click here for a short video demonstrating the use of the package.

Overview

The `mrrobust` package is a collection of commands for performing two-sample Mendelian randomization analyses using summary data of genotype-phenotype and genotype-outcome associations.

Such data can be obtained from repositories such as MR-Base http://www.mrbase.org (Hemani et al. 2016).

The package contains the following commands:

- `mrratio` implements the standard instrumental variable ratio (Wald) estimate with a choice of standard errors/confidence intervals.
- `mrivests` automates calling `mrratio` on all the selected genotypes in your dataset.
- `mregger` implements the IVW and MR-Egger regression approaches introduced in Bowden et al. 2015.
- `mreggersimex` implements the simulation extrapolation algorithm for the MR-Egger model.
- `mreggerplot` implements a scatter plot with fitted line (either from IVW, MR-Egger, or weighted median estimators) and confidence interval.
- `mrmedian` and `mrmedianobs` implement the unweighted, weighted, and penalized weighted median IV estimators robust to 50% invalid instruments in Bowden et al. 2016.
- `mrmodal` implements the zero modal estimator of Hartwig et al. 2017.
- `mrmodalplot` plot of density used in modal estimator.
- `mrforest` implements a forest plot of genotype specific IV estimates and estimates from models (e.g. IVW and MR-Egger).

Installation

First install the dependencies. The package uses Ben Jann's `addplot`, `kdens`, and `moremata` packages, the `heterogi` command (Orsini et al.), the `metan` command (Harris et al.), and the `grc1leg` command (Wiggins). Install those using the following commands:

```
. ssc install addplot
. ssc install kdens
. ssc install moremata
. ssc install heterogi
. ssc install metan
. net install grc1leg, from(http://www.stata.com/users/vwiggins)
```

To install `mrrobust` issue in Stata (in versions 13 and above):

```
. net install mrrobust, from(https://raw.github.com/remlapmot/mrrobust/master/)
```

Installation: from GitHub

- First install dependencies (thanks to Ben Jann for 3 of these):
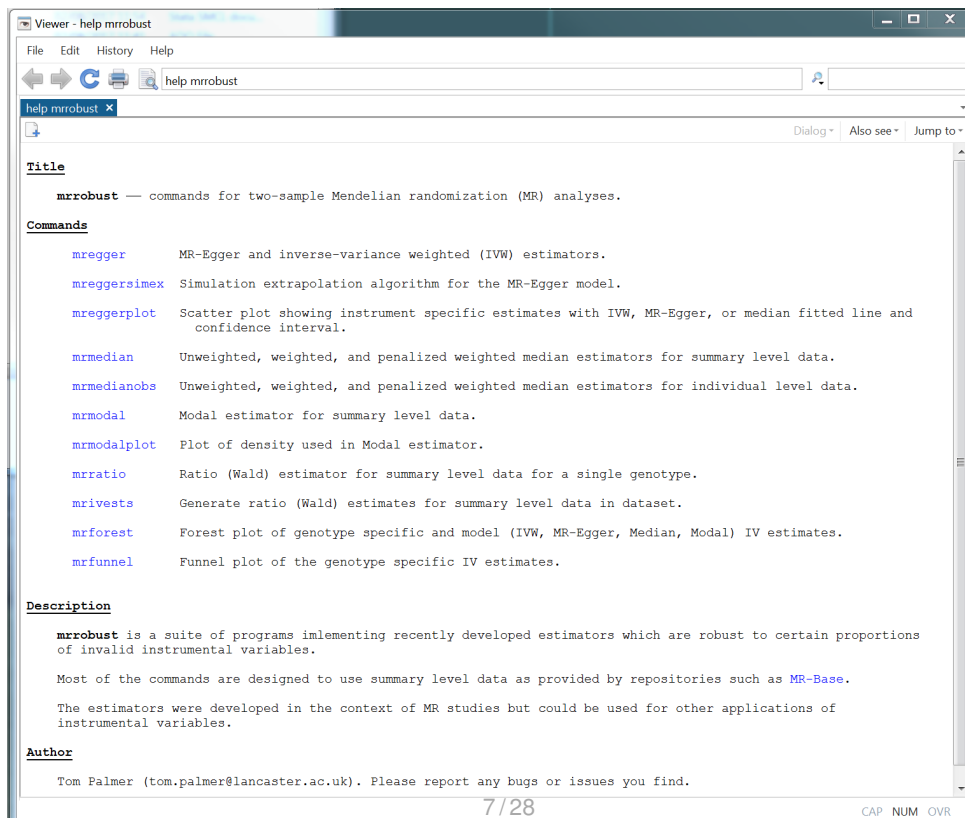
```
. ssc install addplot
. ssc install moremata
. ssc install heterogi
. ssc install kdens
. ssc install metan
```

- In Stata version 13 and above:

```
. net install mrrobust, from(https://raw.github.com/remlapmot/mrrobust/master/)
```

- Obtain updates with:

```
. adoupdate mrrobust, update
```

- In Stata version 12 and below (down to version 9) – install manually from zip archive of repository – save files in current working directory or on adopath.

help mrrobust


```
Viewer - help mrrobust
File  Edit  History  Help
help mrrobust

help mrrobust ×
                                                        Dialog ▾  Also see ▾  Jump to ▾

Title
    mrrobust — commands for two-sample Mendelian randomization (MR) analyses.
Commands
    mregger       MR-Egger and inverse-variance weighted (IVW) estimators.
    mreggersimex  Simulation extrapolation algorithm for the MR-Egger model.
    mreggerplot   Scatter plot showing instrument specific estimates with IVW, MR-Egger, or median fitted line and
                  confidence interval.
    mrmedian      Unweighted, weighted, and penalized weighted median estimators for summary level data.
    mrmedianobs   Unweighted, weighted, and penalized weighted median estimators for individual level data.
    mrmodal       Modal estimator for summary level data.
    mrmodalplot   Plot of density used in Modal estimator.
    mrratio       Ratio (Wald) estimator for summary level data for a single genotype.
    mrivests      Generate ratio (Wald) estimates for summary level data in dataset.
    mrforest      Forest plot of genotype specific and model (IVW, MR-Egger, Median, Modal) IV estimates.
    mrfunnel      Funnel plot of the genotype specific IV estimates.

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

    Most of the commands are designed to use summary level data as provided by repositories such as MR-Base.

    The estimators were developed in the context of MR studies but could be used for other applications of
    instrumental variables.
Author
    Tom Palmer (tom.palmer@lancaster.ac.uk). Please report any bugs or issues you find.
                                          7 / 28                              CAP  NUM  OVR
```

Two Sample MR

- With a single instrument IV estimator is:

$$\beta = \frac{\text{instrument-outcome association}}{\text{instrument-exposure association}}$$

- Can obtain such associations from published GWAS
- GWAS results also now available from online databases such as MR-Base
- Two-sample Mendelian randomization
- Single genotype:

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

Worked example

- Using data from Do et al., Nat Gen, 2013 and analysis in Bowden, Gen Epi, 2016
- Estimate effect of:
  - Exposure: LDL cholesterol (mean differences) on
  - Outcome: risk of coronary heart disease (log odds ratios)
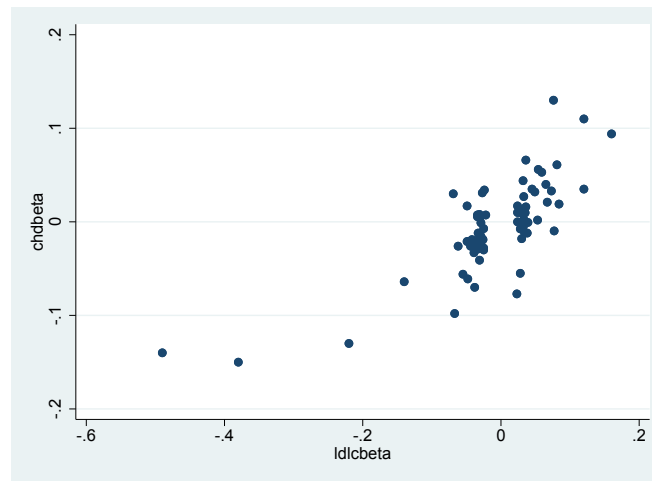
Genotype-specific IV estimates

`mrforest ..`

I_{GX}^2=98.5%

Funnel plot
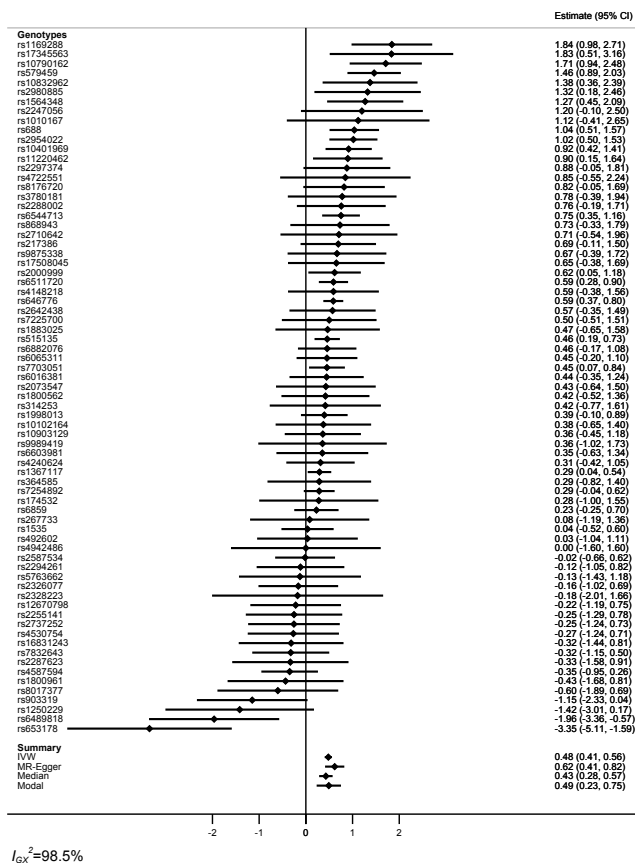
```
mrfunnel chdbeta chdse ldlcbeta ldlcse if sel1==1
```
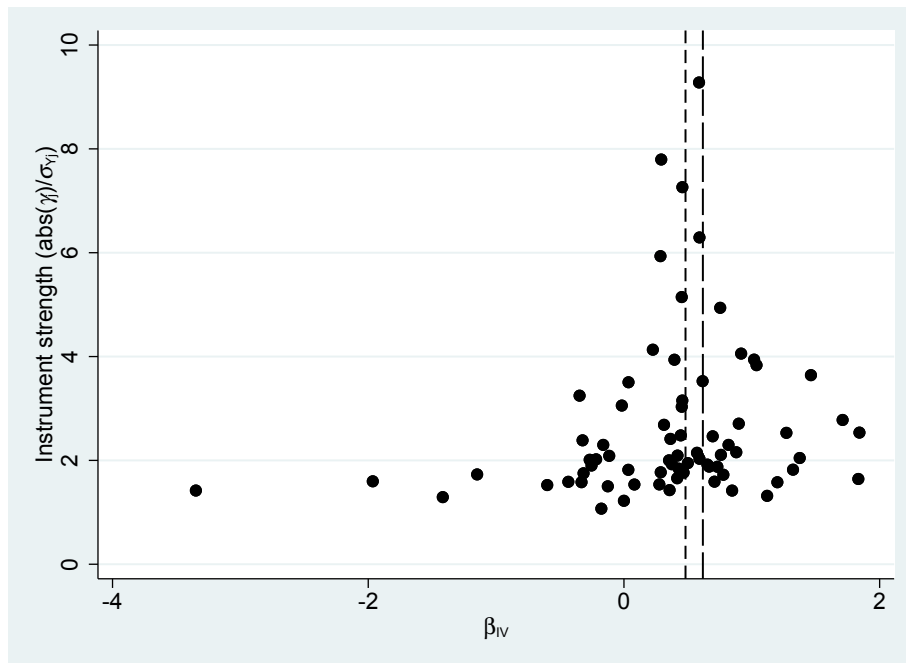

- MR-Egger estimate: long dashed line
- IVW estimate: dashed line

Inverse variance weighted (IVW) regression:

- Summary data version of TSLS with independent instruments (Angrist & Pischke)
- Notation:
  - $\widehat{\Gamma}_j$: genotype-disease associations (SEs: σ_{Yj})
  - $\widehat{\gamma}_i$: genotype-phenotype associations (SEs: σ_{Xj})
- With L instruments
- and instrument specific ratio estimates: $\widehat{\beta}_j = \widehat{\Gamma}_j / \widehat{\gamma}_j$

$$\widehat{\beta}_{\text{IVW}} = \frac{\sum_{j=1}^{L} w_j \widehat{\beta}_j}{\sum_{j=1}^{L} w_j}, \quad w_j = \frac{\widehat{\gamma}_j^2}{\sigma_{Yj}^2}$$

- Estimate biased when one or more instruments exhibit directional pleiotropy

IVW estimate

```
. mregger chdbeta ldlcbeta [aw=1/(chdse^2)] if sel1==1, ivw fe
                                                Number of genotypes = 73
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
chdbeta						
ldlcbeta	.4815055	.038221	12.60	0.000	.4065938	.5564173

```
. lincom ldlcbeta, or
 ( 1)  [chdbeta]ldlcbeta = 0
```

	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
(1)	1.618509	.061861	12.60	0.000	1.501694	1.744412

MR-Egger regression

- Proposed by 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 $\text{var}(\widehat{\beta}_j) = \frac{\sigma^2_{Yj}}{\widehat{\gamma}_j}$.

MR-Egger regression

Model:

$$\widehat{\Gamma}_j = \beta_0 + \beta_1 \widehat{\gamma}_j + \varepsilon_j, \ \varepsilon_j \sim N(0, \ \sigma^2) \text{ weighted by } \frac{1}{\sigma^2_{yj}}$$

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

MR-Egger estimate
With I^2_{GX} statistic

```
. mregger chdbeta ldlcbeta [aw=1/(chdse^2)] if sel1==1, tdist gxse(ldlcse)
                                              Number of genotypes = 73
```

	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
sign(ldlcbeta)*chdbeta						
slope	.6173131	.1034573	5.97	0.000	.4110251	.8236012
_cons	-.0087706	.0054812	-1.60	0.114	-.0196998	.0021585

```
Residual standard error:  1.548
I^2_GX statistic:  98.49%
```

- Additionally specifying `fe` option would calculate SEs with

  `Residual standard error: 1`

Egger regression plot

`mreggerplot ...`
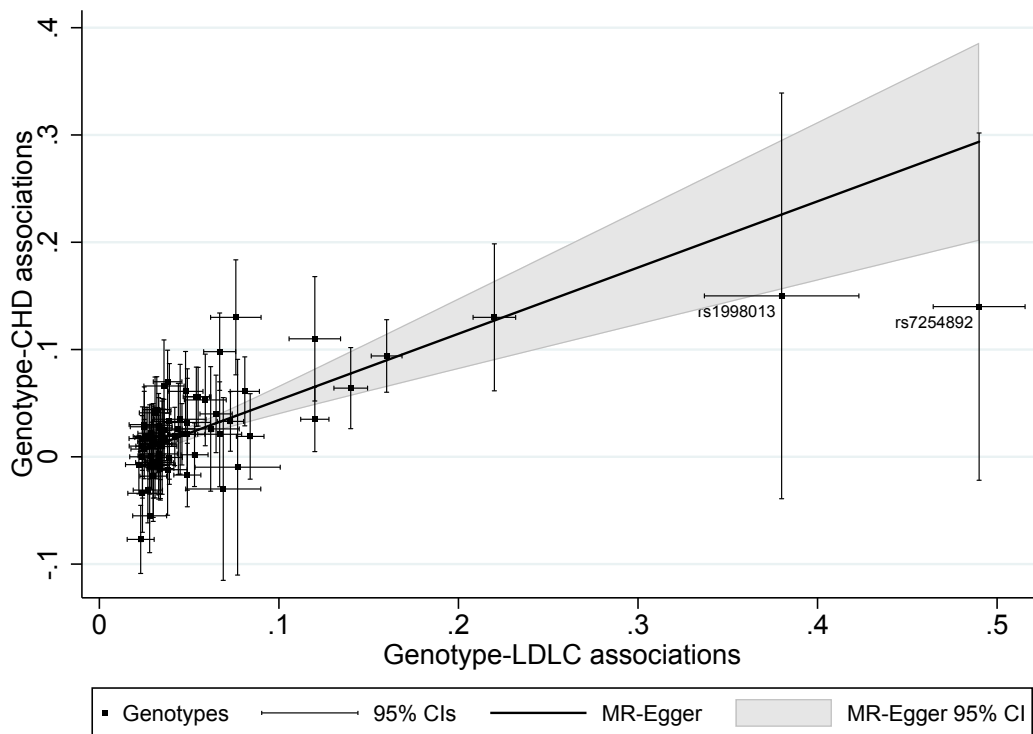
I^2_{GX} statistic

- 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.

$$Q_{GX} = \frac{\sum_{j=1}^{L}(\widehat{\gamma}_j - \overline{\overline{\gamma}})^2}{\sum_{j=1}^{L} \sigma^2_{Xj}}$$

$$I^2_{GX} = \frac{Q_{GX} - (L-1)}{Q_{GX}} = \frac{\sigma^2_{\gamma}}{\sigma^2_{\gamma} + s^2}$$

- I^2_{GX} of 0.9 represents an estimated relative bias of 10% towards the null.

Median estimator

- Essentially take the median or weighted median of the genotype-specific IV estimates

```
. mrmedian chdbeta chdse ldlcbeta ldlcse if sel1==1, weighted seed(12345)
                                                     Number of genotypes = 73
                                                     Replications = 1000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
beta	.4582573	.0624645	7.34	0.000	.3358291 .5806856

Modal estimator

- Hartwig et al., IJE, 2017
- Take the instrument specific ratio estimates
- Perform kernel density estimation - Normal density
- Find the highest point of the estimated density - mode
- Sensitive to the bandwidth parameter used in density estimation

Modal estimator

```
. mrmodalplot chdbeta chdse ldlcbeta ldlcse if sel1==1
```


- Choose value of ϕ which gives smoothest density, here $\phi = 1$.

Modal estimate

```
. mrmodal chdbeta chdse ldlcbeta ldlcse if sel1==1, weighted seed(12345) phi(.25)
                                                     Number of genotypes = 73
                                                     Replications = 1000
                                                     Phi = .25
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
beta	.5820001	.1365403	4.26	0.000	.314386	.8496142

```
. mrmodal chdbeta chdse ldlcbeta ldlcse if sel1==1, weighted seed(12345) phi(1)
                                                     Number of genotypes = 73
                                                     Replications = 1000
                                                     Phi = 1
```
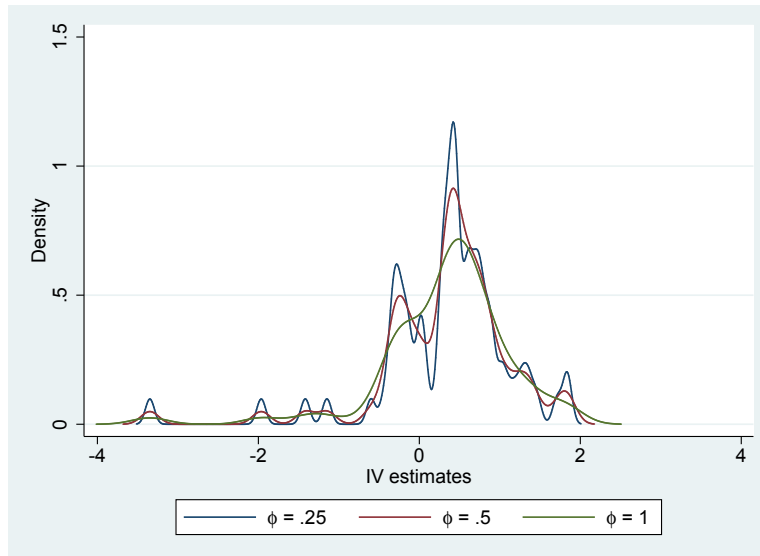
	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
beta	.4789702	.0718135	6.67	0.000	.3382183	.6197221

MR-Egger SIMEX

- Approach to assessing the NOME assumption in the weights used in IVW/MR-Egger

```
. mreggersimex chdbeta ldlcbeta [aw=1/chdse^2] if sel1==1, ///
>         gxse(ldlcse) seed(12345)
(running mreggersimexonce on estimation sample)

Bootstrap replications (25)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
.........................
                                          Number of genotypes = 73
                                          Bootstrap replications = 25
                                          Simulation replications = 50
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
slope	.6256194	.1166245	5.36	0.000	.3970396	.8541991
_cons	−.0089987	.0062257	−1.45	0.148	−.0212009	.0032035

MR-Egger SIMEX
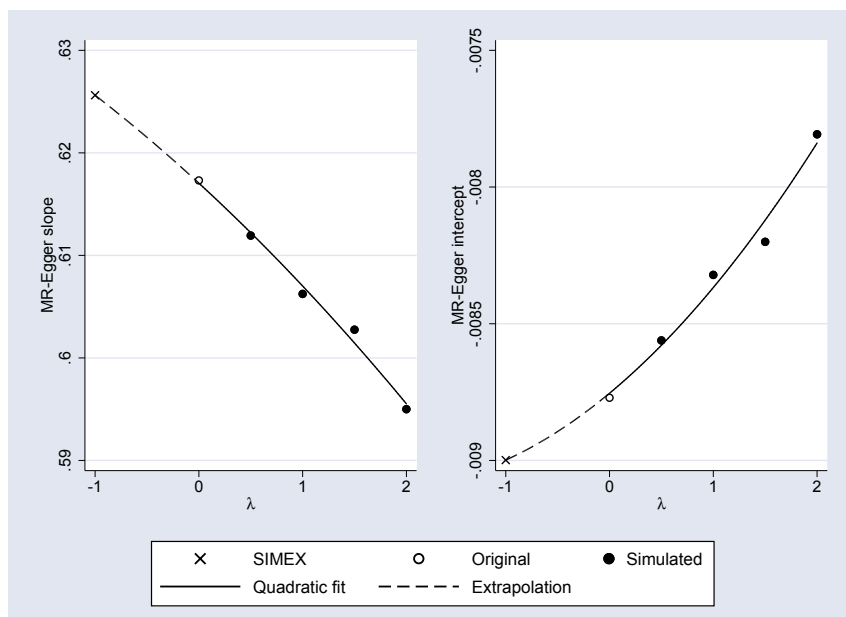
- $\lambda = 0$: original data estimate
- $\lambda = -1$: estimate from data with "no measurement error"

Stata wishes

- I often push more than 1 update to GitHub per day - would help me if I could additionally specify time in distribution date in .pkg file, current format is only:

  `d Distribution-Date: yyyymmdd`

- MR-Base uses Google authentication so Stata commands for Google, Facebook, Microsoft authentication – like R package `googleAuthR` – would be very helpful

Summary

- `mrrobust` package
- Install from GitHub repo
- Esimators: IVW, MR-Egger (I_{GX}^2 statistic), Median, Modal
- Plots: IV forest plot, Egger regression plot, modal density plot
- Testing/validation: I have cscripts for each command – on GitHub – graph commands much harder and more inconvenient to test
- To do: many methods - field developing rapidly

Bibliography

- Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015, 44, 2, 512–525.

- Bowden J, Davey Smith G, Haycock PC, Burgess S. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology, published online 7 April.

- Bowden J, Del Greco F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. 2016. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I-squared statistic. International Journal of Epidemiology.

- Davey Smith G, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease. International Journal of Epidemiology. 2003; 32, 1, 1–22

- Do R et al., 2013. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature Genetics. 45, 13451352. DOI: http://dx.doi.org/10.1038/ng.2795

- Hemani G, Zheng J, Wade KH, et al., Davey Smith G, Gaunt TR, Haycock PC. The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv, 2016, doi:10.1101/078972; http://www.mrbase.org/ .

- Yavorska OO & Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. International Journal of Epidemiology. 2017

- Yavorska O, Burgess S. MendelianRandomization: Mendelian Randomization Package. 2016, version 0.2.0. https://CRAN.R-project.org/package=MendelianRandomization

Thank you for your attention.

Any questions?