

# **mrrobust: a Stata package for MR-Egger regression type analyses**

London Stata User Group Meeting 2017

8<sup>th</sup> September 2017

Tom Palmer   Wesley Spiller   Neil Davies



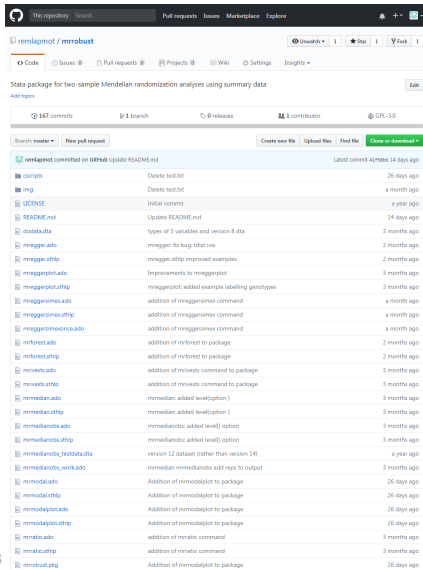
- Introduction
- GitHub and installation
- Worked example
- Stata wishes
- Discussion

- 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 (Linux 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



The screenshot displays the GitHub repository page for `remlapmot/mrrobust`. At the top, the repository name is shown along with navigation links for Code, Issues, Pull requests, Projects, Wiki, Settings, and Insights. Below this, a description states: "Stata package for two-sample Mendelian randomization analyses using summary data". The repository statistics show 167 commits, 1 branch, 0 releases, and 1 contributor. A table of files and their commit history is visible, including:

File	Commit Message	Time Ago
scripts	Delete test.txt	26 days ago
img	Delete test.txt	a month ago
LICENSE	Initial commit	a year ago
README.md	Update README.md	14 days ago
datafiles	types of 3 variables and version 8 data	3 months ago
mrregger.ado	mrregger: fix bug tdist inv	2 months ago
mrregger.athlp	mrregger.athlp: improved examples	2 months ago
mrreggerplot.ado	Improvements to mrreggerplot	3 months ago
mrreggerplot.athlp	mrreggerplot: added example labelling genotypes	3 months ago
mrreggersimex.ado	addition of mrreggersimex command	a month ago
mrreggersimex.athlp	addition of mrreggersimex command	a month ago
mrreggersimexconc.ado	addition of mrreggersimex command	a month ago
mrforest.ado	addition of mrforest to package	2 months ago
mrforest.athlp	addition of mrforest to package	2 months ago
mrivests.ado	addition of mrivests command to package	3 months ago
mrivests.athlp	addition of mrivests command to package	3 months ago
mrmedian.ado	mrmedian: added level(option)	3 months ago
mrmedian.athlp	mrmedian: added level(option)	3 months ago
mrmedianconc.ado	mrmedianconc: added level() option	3 months ago
mrmedianconc.athlp	mrmedianconc: added level() option	3 months ago
mrmedianconc_testdata.ado	version 12 dataset (rather than version 14)	a year ago
mrmedianconc_work.ado	mrmedian mrmedianconc: add raps to output	3 months ago
mrmod.ado	Addition of mrmodplot to package	26 days ago
mrmod.athlp	Addition of mrmodplot to package	26 days ago
mrmodplot.ado	Addition of mrmodplot to package	26 days ago
mrmodplot.athlp	Addition of mrmodplot to package	26 days ago
mrmodplot.athlp	Addition of mrmodplot to package	26 days ago
mrmod.ado	addition of mrmod command	3 months ago
mrmod.athlp	addition of mrmod command	3 months ago
mrmodplot.ado	Addition of mrmodplot to package	26 days ago

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

Search master • mrrobus / README.md Find file Copy path

rendapnot update README.md commit 14 days ago

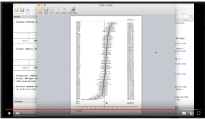
1 contributor

392 lines (88 sloc) 5.86 kB Raw Clone History

## mrrobus: 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 `mrrobus` 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:

- `mrrobus`: implements the standard instrumental variable ratio (Wald) estimate with a choice of standard errors/confidence intervals.
- `mrrobusci`: automates calling `mrrobus` on all the selected genotypes in your dataset.
- `mrrobusivw`: implements the IVW and MR-Egger regression approaches introduced in Bowden et al. 2015.
- `mrrobusstata`: implements the simulation extrapolation algorithm for the MR-Egger model.
- `mrrobusplot`: implements a scatter plot with fitted line (either from IVW, MR-Egger, or weighted median estimators) and confidence interval.
- `mrrobusci` and `mrrobusciplot`: implement the unweighted, weighted, and penalized weighted median IV estimators robust to 50% invalid instruments in Bowden et al. 2016.
- `mrrobusciplot`: implements the zero modal estimator of Hartwig et al. 2017.
- `mrrobusciplot`: plot of density used in model estimator.
- `mrrobusciplot`: 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`, `klein`, and `areadata` packages, the `heterog` command (Orsini et al.), the `metan` command (Hartis et al.), and the `gscdiag` command (Wiggins). Install those using the following commands:

- 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:

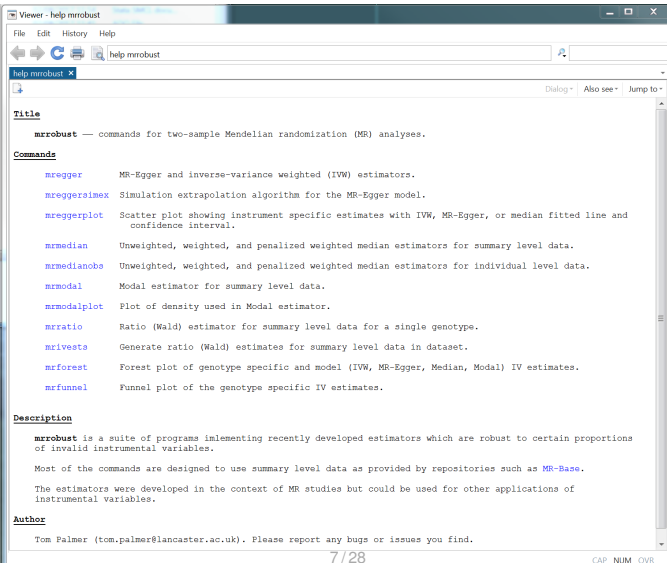
```
. net install mrrobust, from(https://raw.githubusercontent.com/remlapmot/mrrobust/master/)
```

- 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 x

Dialog - Also see - Jump to -

**Title**

**mrrobust** — commands for two-sample Mendelian randomization (MR) analyses.

**Commands**

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

**Description**

**mrrobust** is a suite of programs implementing 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

- With a single instrument IV estimator is:

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

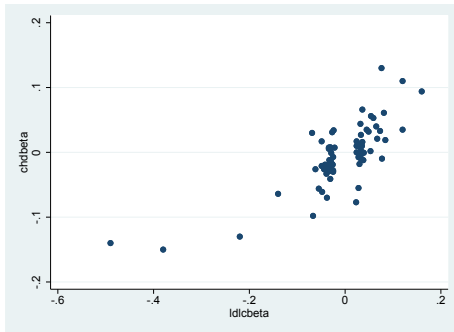
- Can obtain such associations from published GWAS
- GWAS results also now available from online databases such as MR-Base
- Two-sample Mendelian randomization
- Single genotype:

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



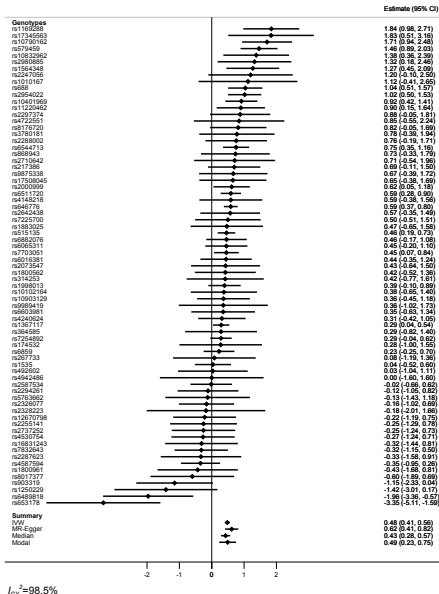
# Worked example

- 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)



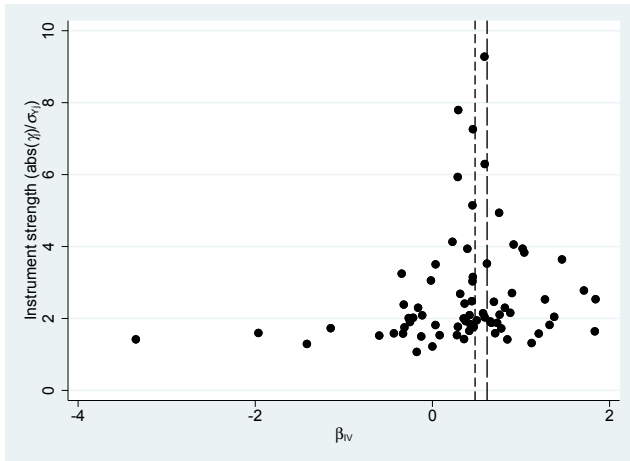
# Genotype-specific IV estimates

mrforest ..



# Funnel plot

```
mrfunnel chdbeta chdse ldldbeta ldldse 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:
  - $\hat{\Gamma}_j$ : genotype-disease associations (SEs:  $\sigma_{Yj}$ )
  - $\hat{\gamma}_i$ : genotype-phenotype associations (SEs:  $\sigma_{Xj}$ )
- With  $L$  instruments
- and instrument specific ratio estimates:  $\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j$

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

- Estimate biased when one or more instruments exhibit directional pleiotropy

```
. 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  $\hat{\gamma}_j$  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  $\text{var}(\hat{\beta}_j) = \frac{\sigma_{\gamma_j}^2}{\hat{\gamma}_j}$ .

Model:

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

- 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 ldldbata [aw=1/(chdse^2)] if sel1==1, tdist gxse(ldlcse)
```

Number of genotypes = 73

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

Residual standard error: 1.548

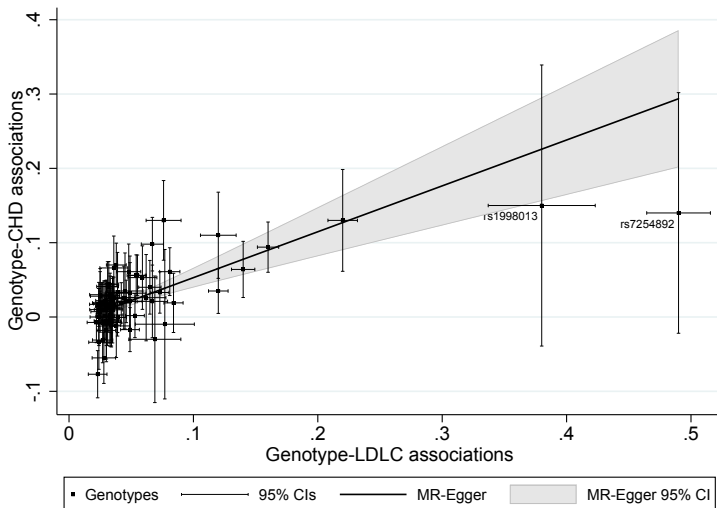
$I^2_{GX}$  statistic: 98.49%

- Additionally specifying `fe` option would calculate SEs with Residual standard error: 1



# Egger regression plot

mreggerplot ...



- 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.

$$Q_{GX} = \frac{\sum_{j=1}^L (\hat{\gamma}_j - \bar{\gamma})^2}{\sum_{j=1}^L \sigma_{Xj}^2}$$
$$I^2_{GX} = \frac{Q_{GX} - (L - 1)}{Q_{GX}} = \frac{\sigma_{\gamma}^2}{\sigma_{\gamma}^2 + s^2}$$

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

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

```
. mrmedian chdbeta chdse ldldbeta ldldse if sel1==1, weighted seed(12345)
```

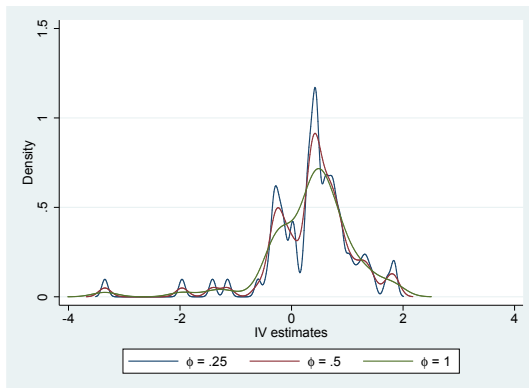
Number of genotypes = 73

Replications = 1000

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

- 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

```
. mrmodalplot chdbeta chdse ldldcbeta ldldcse if sel1==1
```



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

```
. mrmodal chdbeta chdse ldldbета 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 ldldbета 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

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

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

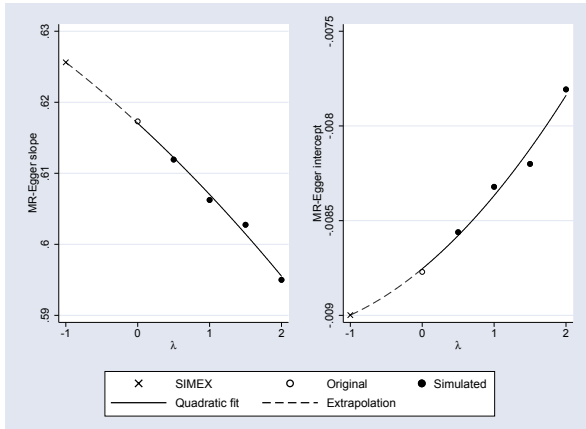
Bootstrap replications (25)

—|— 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”



- 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

- `mrrobust` package
- Install from GitHub repo
- Estimators: IVW, MR-Egger ( $I^2_{GX}$  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

- 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 genome 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?