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

- 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)

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

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 mrobust, from(https://raw.github.com/remlapmot/mrobust/master/)

• Obtain updates with:
  . adoupdate mrobust, 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.
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

\[ \text{mrforest} \ldots \]
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:
  - $\hat{\Gamma}_j$: genotype-disease associations (SEs: $\sigma_{Yj}$)
  - $\hat{\gamma}_j$: 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
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   | 0.4815055| 0.038221  | 12.60| 0.000| 0.4065938           |
|            |          |           |     |      | 0.5564173           |
```

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

|                    | Odds Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|--------------------|------------|-----------|-----|------|---------------------|
| (1)                | 1.618509   | 0.061861  | 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 $\text{var}(\hat{\beta}_j) = \frac{\sigma^2_{\gamma_j}}{\gamma_j}$. 

13/28

14/28
MR-Egger regression

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 \( I_{GX}^2 \) statistic

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

|                | 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  |

Number of genotypes = 73

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

Genotype-CHD associations

Genotype-LDLC associations

Genotypes 95% CIs

MR-Egger 95% CI

I² 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.

$$Q_{GX} = \frac{\sum_{j=1}^{L} (\hat{\gamma}_j - \bar{\gamma})^2}{\sum_{j=1}^{L} \sigma^2_{X_j}}$$

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

- $I^2_{GX}$ 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 $\phi$ which gives smoothest density, here $\phi = 1$.

Modal estimate

. `mrmodal chdbeta chdse ldlcbeta ldlcse if sel1==1, weighted seed(12345) phi(.25)`

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>beta</td>
<td>.5820001</td>
<td>.1365403</td>
<td>4.26</td>
<td>0.000</td>
</tr>
</tbody>
</table>

. `mrmodal chdbeta chdse ldlcbeta ldlcse if sel1==1, weighted seed(12345) phi(1)`

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>beta</td>
<td>.4789702</td>
<td>.0718135</td>
<td>6.67</td>
<td>0.000</td>
</tr>
</tbody>
</table>
• Approach to assessing the NOME assumption in the weights used in IVW/MR-Egger

```
mreggersimex chdbeta ldcbeta [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  | 0.6256194 | 0.1166245 | 5.36 | 0.000 | 0.3970396, 0.8541991 |
| _cons  | -0.0089987 | 0.0062257 | -1.45 | 0.148 | -0.0212009, 0.0032035 |
```

• $\lambda = 0$: 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 ($I^2_{GX}$ 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, 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.


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

