Software code in R for performing instrumental variable analyses for Mendelian randomization investigations

maintained by Stephen Burgess

September 4, 2015

This is a non-traditional publication to provide software code for the Mendelian randomization community in a single document. It will be updated when necessary as new methods are developed. Hopefully, this will become a collaborative resource than can be authored by the community rather than a single-author manuscript. However, Stephen Burgess retains the prerogative to exert editorial control.

Currently, it only contains R code. If someone wants to write Stata code or code for any other software package, this could be included in a separate document.

Contributors:

- Stephen Burgess (sb452@medschl.cam.ac.uk)

Contents

1 Introduction and notation

```
### Dimensions
N # sample size
K # number of genetic variants

### Individual-level data
g # genetic variant(s), matrix dimension N x K
x # risk factor/exposure, vector length N
y # outcome, vector length N

###### Summary data
bx   # genetic associations with exposure, vector length K
by   # genetic associations with outcome, vector length K
bxse # standard errors of genetic associations with exposure
byse # standard errors of genetic associations with outcome
```

2 Mendelian randomization analysis with individual-level data

This section on standard Mendelian randomization methods with individual-level data in a single dataset is based on Burgess et al. [2015], which in turn is based on Chapter 4 of Burgess and Thompson [2015]. We consider in turn the ratio of coefficients method, two-stage methods, likelihood-based methods, and semi-parametric methods.

2.1 Ratio of coefficients (Wald) method – single instrument

The ratio of coefficients method, or the Wald method is the simplest way of estimating the causal effect of the risk factor on the outcome (original paper). The ratio method uses a single instrumental variable (IV), which can be a single SNP or an allele score (see Burgess and Thompson [2013] for background on allele scores).

```
## A. Ratio estimate (continuous outcome)

bx   = lm(x~g)$coef[2]
bxse = summary(lm(x~g))$coef[2,2]
by   = lm(y~g)$coef[2]
byse = summary(lm(y~g))$coef[2,2]

beta_ratio = by/bx
```

See Greenland [2000] or Martens et al. [2006] for an introduction to instrumental variable methods and causal estimation, or Lawlor et al. [2008] for a specific Mendelian randomization perspective.

```
## B. Asymptotic standard error (poor with weak instruments)

# 1. Delta method approximation (summarized data)

se_ratio_approx = byse/bx
        # first order approximation
se_ratio_approx = sqrt(byse^2/bx^2 + by^2*bxse^2/bx^4 - 2*theta*by/bx^3)
        # second order approximation
        # theta is correlation between numerator
        #   and denominator in ratio estimate

# 2. Two-stage least squares method for standard error (individual-level
    data)

library(sem)
se_tsls = sqrt(tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$V[1,1])


## C. Valid confidence intervals with weak instruments
```

```
# 1. Fieller's theorem
  f0 = by^2 - qt(0.975, N)^2 * byse^2
  f1 = bx^2 - qt(0.975, N)^2 * bxse^2
  f2 = by*bx
   D = f2^2 - f0*f1

  if(D>0) {
    r1 = (f2-sqrt(D))/f1
    r2 = (f2+sqrt(D))/f1
if(f1>0) { cat("Confidence interval is a closed interval [a,b]: \n a=",
    r1, ", b=", r2, sep="")) }
if(f1<0) { cat("Confidence interval is the union of two open intervals
    (-Inf, a], [b, +Inf): \n a=", r2, ", b=", r1, sep="")) }
          }
if(D<0|D==0) { cat("No finite confidence interval exists other than the
    entire real line.") }

# 2. Anderson--Rubin

library(ivpack)
ivmodel = ivreg(y~x|g, x=TRUE)
anderson.rubin.ci(ivmodel)
         # As with Fieller's theorem, interval may be a closed interval,
             the union of two open intervals, or undefined
```

A reference for Fieller's theorem is Buonaccorsi [2005] (original reference is Fieller [1954], a web-based tool is available at `spark.rstudio.com/sb452/fieller`. A reference for the Anderson–Rubin method is Mikusheva [2010] (original reference is Anderson and Rubin [1949]).

```
## D. Binary outcome, logistic-linear model (assuming case--control data)

bx   = lm(x[y==0]~g[y==0])$coef[2]
bxse = summary(lm(x[y==0]~g[y==0]))$coef[2,2]
by   = glm(y~g, family=binomial)$coef[2]
byse = summary(lm(y~g))$coef[2,2]
```

Genetic associations with the risk factor are estimated in control participants only (see Didelez and Sheehan [2007], Bowden and Vansteelandt [2011]). This is for three main reasons: to avoid reverse causation, to avoid biases due to outcome-dependent sampling, and because the controls are a more representative sample of the population as a whole. There are some technical issues relating to the ratio estimate with a binary outcome and a logistic regression model due to the non-collapsibility of odds ratios Greenland et al. [1999], but it is a consistent estimator under the null Burgess and CHD CRP Genetics Collaboration [2013]; Vansteelandt et al. [2011].

2.2 Two-stage least squares method

```
## A. Continuous outcome
library(sem)
beta_tsls = tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$coef[1]
se_tsls  = sqrt(tsls(y, cbind(x, rep(1,N)), cbind(g, rep(1,N)),
   w=rep(1,N))$V[1,1])

library(ivpack)
ivmodel = ivreg(y~x|g, x=TRUE)
summary(ivmodel)
beta_tsls = ivreg(y~x|g, x=TRUE)$coef[2]
se_tsls  = summary(ivreg(y~x|g, x=TRUE))$coef[2,2]
```

The same point estimate (although not the standard error) can be obtained using sequential regression:

```
seqreg_tsls = lm(y~lm(x~g)$fitted)
```

.

```
## B. Binary outcome, logistic-linear model (assuming case--control data)

glm(y~lm(x~g)$fitted)
```

References

Anderson, T. and Rubin, H. 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics*, 21(1):570–582. (page 5).

Bowden, J. and Vansteelandt, S. 2011. Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine*, 30(6):678–694. (page 5).

Buonaccorsi, J. 2005. *Encyclopedia of Biostatistics*, chapter Fieller's theorem, pages 1951–1952. Wiley. (page 5).

Burgess, S. and CHD CRP Genetics Collaboration 2013. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32(27):4726–4747. (page 5).

Burgess, S., Small, D. S., and Thompson, S. G. 2015. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research.* (page 4).

Burgess, S. and Thompson, S. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, 42(4):1134–1144. (page 4).

Burgess, S. and Thompson, S. G. 2015. *Mendelian randomization: methods for using genetic variants in causal estimation.* Chapman & Hall. (page 4).

Didelez, V. and Sheehan, N. 2007. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. (page 5).

Fieller, E. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 16(2):175–185. (page 5).

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729. (page 4).

Greenland, S., Robins, J., and Pearl, J. 1999. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46. (page 5).

Lawlor, D., Harbord, R., Sterne, J., Timpson, N., and Davey Smith, G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163. (page 4).

Martens, E., Pestman, W., de Boer, A., Belitser, S., and Klungel, O. 2006. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267. (page 4).

Mikusheva, A. 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247. (page 5).

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. 2011. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422. (page 5).