Package 'iva'

October 13, 2022

Type Package

Title Instrumental Variable Analysis in Case-Control Association Studies

Version 0.1.0

Author Han Zhang, Kai Yu

Maintainer Han Zhang <han.zhang2@nih.gov>

Imports ucminf, Formula

Description Mendelian randomization (MR) analysis is a special case of instrumental variable analysis with genetic instruments. It is used to estimate the unconfounded causal effect of an exposure. This package implements estimating and testing methods in Zhang et al. (2019) for MR analysis in case-control studies. It (1) estimates the causal effect of a quantitative exposure by the quasi empirical likelihood approach; (2) uses Lagrange multiplier test for testing the presence of causal; (3) provides a test for the presence of confounder.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2019-01-19 23:50:03 UTC

R topics documented:

edata *Exposure dataset*

Description

An example of the exposure data.

Usage

```
data("edata")
```

Format

A data frame with 1000 observations on the following 8 variables.

`d` condition status (e.g. disease or not).

`z` quantitative exposure.

`x2` character covariate

`x3` binary covariate

`g1` instrument

`g2` instrument

`g3` instrument

`id` subject IDs

Details

The column `id` is always required in exposure data.

Examples

```
data(edata)
head(edata, 2)
```

mra *Conducting Mendelian Randomization Analysis*

Description

`mra` is used to estimate causal effect of a quantitative exposure on a binary outcome in a Mendelian randomization analysis, of which outcome data is collected from a case-control study.

Usage

```
mra(oformula, odata, eformula, edata)
```

Arguments

oformula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted on the case-control dataset. More details of model specification are illustrated in 'Details' and 'Examples'.
odata	a data frame containing variables specified in oformula, including outcome (case-control status), instruments, and covariates (if any).
eformula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted on the exposure dataset. More details of model specification are illustrated in 'Details' and 'Examples'.
edata	a data frame containing variables specified in eformula, including exposure, outcome, instruments, and covariates (if any).

Details

oformula specifies the model used to fit case-control data (outcome data), including case-control status, instruments, and covariates (if any). mra relies on a feature supported by the Formula package, so that the right hand side of oformula is separated into two parts by |. In general, format for oformula is case-control status ~ covariates | instuments. For example, y ~ x1 + x2 | g1 + g2 specifies an outcome model for binary outcome y, with two covariates x1 and x2, and two instruments g1 and g2 fitted in the model. One can use y ~ 1 | g1 + g2 if no covariate to be adjusted. An intercept is always required in oformula. mra will convert character or factor variables into dummy variable, so one does not have to create dummy variables by their own unless they want to specify the baseline.

eformula specifies the model used to fit exposure data, including exposure variable, outcome status (same variable as the case-control status), instruments, and covariates (if any). The right hand side of eformula is similar to that of oformula. The left hand side of eformula also has two parts separated by |. In general, format for eformula is exposure | outcome ~ x1 | g1 + g2. mra requires to know outcome status for every sample in exposure data.

Note 1: Different covariates could be adjusted in oformula and eformula, which is quite common in practice as case-control data and exposure data may be collected for different research purpose, therefore, different covariates are measured. The instruments specified in oformula and eformula must be the same.

Note 2: The case-control data and exposure data may share some subjects. This happens when researcher picks some subjects from the case-control study to measured their exposure based on their criteria, and has another set of exposure data from other sources. As such, a subject can appear in both datasets. Both odata and edata should always have a column named as id, so that mra can account for the variation due to data overlapping. This column is needed even if your case-control and exposure datasets do not share any subject.

Value

mra returns an object of class "mra".

The function summary is used to display a summary of the results. Many generic accessor functions are supported in mra to extract useful information of the value returned by mra. See 'Note' for more details.

An object of class "mra" is a list containing the following components:

coefficients	a named vector of coefficients. bet is the causal effect. alp. and phi. are coefficients estimated for the instruments and covariates in the exposure model. a and gam. are coefficients estimated for the intercept and covariates in the outcome model. alp0 and c0 are the intercept and estimated variance of random error in the exposure model for subjects without conditions. If there are subjects with conditions in the exposure data, alp1 and c1 are estimated as the intercept and variance of random error in the exposure model for those subjects. Refer to the paper for more details of model parameterization used in mra.
residuals	the residuals for subjects without conditions in exposure data, that is exposure minus fitted values. Covariates (if any) are also adjusted.
fitted.values	the fitted mean values for subjects without conditions in exposure data. Covariates (if any) are also adjusted.
wald	Wald test.
lm	Lagrange multiplier test. Recommanded for confidence interval and hypothesis testing.
ct	test for presence of confounders. Available when exposure is also measured for some subjects with conditions.
tsr	generalized two-stage regression method. This method is deprecated as it generates a more biased estimate for causal effect with underestimated standard error, a too-narrow confidence interval, and an underpowered test.
vcov	variance-covariance matrix of coefficients. Using the generic function vcov to access this component is recommanded.
sigma2	the estimated variance of the random errors in the exposure model. c0 for subjects without conditions, c1 for subjects with conditions (if any).
call	the matched call.

Note

The generic functions summary, coef, residuals, fitted, vcov are supported.

The generic function plot is used for model diagnosis to the Assumption 2 in paper: in controls, random error of the exposure model is independent of instruments conditioned on confounders. This is done by examining the "residuals versus fits plot" of residuals (residuals) and fitted values (fitted) on subjects without conditions in the exposure data. A presence of heteroscedastic among those subjects suggests potential violation of this assumption, and caution is needed when interpreting the result.

The generic function confint can be used to compute Wald's confidence intervals for one or more parameters in fitted models (case-control and exposure models). This works fine for all parameters except for the causal effect, as we showed in paper that Wald's confidence interval for causal effect can have coverage probability much lower than its nominal level. We recommand to use the confidence interval derived from the Lagrange multiplier test, which is also more powerful than the Wald test.

Author(s)

Han Zhang

References

Zhang, H., Qin, J., Berndt, S.I., Albanes, D., Gail, M.H., Yu, K. (2018) On Mendelian Randomization in Case-Control Studies. Under review.

Examples

```
## This example estimates parameters in the
## following underlying models:
## 1. outcome model. A logistic regression model
##    d ~ z + x, of which the coefficient of
##    exposure z is the causal effect of interest;
## 2. exposure model. A quasi-likelihood model
##    z ~ g + x, of which g are used as instruments.
## In Mendelian randomization, those parameters
## could be estimated by fitting two working models
## with special parameterization:
## a. A logistic regression model d ~ g + x
## b. A quasi-likelihood model z ~ d + g + x

data(edata)
data(odata)

fit <- mra(d ~ x1 + x2 | g1 + g2 + g3,
           odata,
           z | d ~ x2 + x3 | g1 + g2 + g3,
           edata)

## summary tables for outcome model and exposure model
## and for testing the presence of confounder (if available)
summary(fit)

## causal effect estimate and its standard error
coef(fit)['bet']
sqrt(vcov(fit)['bet', 'bet'])

## Lagrange multiplier test
fit$lm

## model diagnosis
plot(fit)
```

---

odata *Case-control dataset*

Description

An example of the case-control data (or the outcome data).

Usage

```
data("odata")
```

Format

A data frame with 3500 observations on the following 7 variables.

`d` case-control status

`x1` numeric covariate

`x2` character covariate

`g1` instrument

`g2` instrument

`g3` instrument

`id` subject IDs

Details

The column `id` is always required in exposure data.

Examples

```
data(odata)
head(odata, 2)
```

Index