Package: RobustIV (via r-universe)

September 9, 2024

Type Package

Title Robust Instrumental Variable Methods in Linear Models

Version 0.2.4

Description Inference for the treatment effect with possibly invalid instrumental variables via TSHT('Guo et al.' (2016) <arXiv:1603.05224>) and SearchingSampling('Guo' (2021) <arXiv:2104.06911>), which are effective for both low- and high-dimensional covariates and instrumental variables; test of endogeneity in high dimensions ('Guo et al.' (2016) <arXiv:1609.06713>).

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

URL https://github.com/zijguo/RobustIV

Imports glmnet, MASS, Matrix, igraph, intervals, CVXR

NeedsCompilation no

Depends R (>= 2.10)

Repository https://zijguo.r-universe.dev

RemoteUrl https://github.com/zijguo/robustiv

RemoteRef HEAD

RemoteSha 66a2212486405fc4a1ce5188f8982338a8f4da62

Contents

`cf` *Control-Function*

Description

Implement the control function method for estimation and inference of nonlinear treatment effects.

Usage

```
cf(formula, d1 = NULL, d2 = NULL)
```

Arguments

`formula`	A formula describing the model to be fitted.
`d1`	The baseline treatment value.
`d2`	The target treatment value.

Details

For example, the formula `Y ~ D + I(D^2)+X|Z+I(Z^2)+X` describes the models $Y = \alpha_0 + D\beta_1 + D^2\beta_2 + X\phi + u$ and $D = \gamma_0 + Z\gamma_1 + Z^2\gamma_2 + X\psi + v$. Here, the outcome is `Y`, the endogenous variables are `D` and `I(D^2)`, the baseline covariates are `X`, and the instrument variables are `Z`. The formula environment follows the formula environment in the ivreg function in the AER package. The linear term of the endogenous variable, for example, `D`, must be included in the formula for the outcome model. If either one of `d1` or `d2` is missing or `NULL`, `CausalEffect` is calculated assuming that the baseline value `d1` is the median of the treatment and the target value `d2` is `d1+1`.

References

Guo, Z. and D. S. Small (2016), Control function instrumental variable estimation of nonlinear causal effect models, *The Journal of Machine Learning Research* 17(1), 3448–3482.

Examples

```
Y <- mroz[,"lwage"]
D <- mroz[,"educ"]
Z <- as.matrix(mroz[,c("motheduc","fatheduc","huseduc")])
X <- as.matrix(mroz[,c("exper","expersq","age")])
cf.model <- cf(Y~D+I(D^2)+X|Z+I(Z^2)+X)
summary(cf.model)
```

endo.test *Endogeneity test in high dimensions*

Description

Conduct the endogeneity test with high dimensional and possibly invalid instrumental variables.

Usage

```
endo.test(
  Y,
  D,
  Z,
  X,
  intercept = TRUE,
  invalid = FALSE,
  method = c("Fast.DeLasso", "DeLasso", "OLS"),
  voting = c("MP", "MaxClique"),
  alpha = 0.05,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
invalid	If TRUE, the method is robust to the presence of possibly invalid IVs; If FALSE, the method assumes all IVs to be valid. (default = FALSE)
method	The method used to estimate the reduced form parameters. "OLS" stands for ordinary least squares, "DeLasso" stands for the debiased Lasso estimator, and "Fast.DeLasso" stands for the debiased Lasso estimator with fast algorithm. (default = "Fast.DeLasso")
voting	The voting option used to estimate valid IVs. 'MP' stnads for majority and plurality voting, 'MaxClique' stands for maximum clique in the IV voting matrix. (default = 'MP')
alpha	The significance level for the confidence interval. (default = 0.05)
tuning.1st	The tuning parameter used in the 1st stage to select relevant instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)
tuning.2nd	The tuning parameter used in the 2nd stage to select valid instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)

Details

When `voting = MaxClique` and there are multiple maximum cliques, the null hypothesis is rejected if one of maximum clique rejects the null. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01\log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

`endo.test` returns an object of class "endotest", which is a list containing the following components:

`Q`	The test statistic.
`Sigma12`	The estimated covaraince of the regression errors.
`VHat`	The set of selected vaild IVs.
`p.value`	The p-value of the endogeneity test.
`check`	The indicator that $H_0 : \Sigma_{12} = 0$ is rejected.

References

Guo, Z., Kang, H., Tony Cai, T. and Small, D.S. (2018), Testing endogeneity with high dimensional covariates, *Journal of Econometrics*, Elsevier, vol. 207(1), pages 175-187.

Examples

```
n = 500; L = 11; s = 3; k = 10; px = 10;
alpha = c(rep(3,s),rep(0,L-s)); beta = 1; gamma = c(rep(1,k),rep(0,L-k))
phi<-(1/px)*seq(1,px)+0.5; psi<-(1/px)*seq(1,px)+1
epsilonSigma = matrix(c(1,0.8,0.8,1),2,2)
Z = matrix(rnorm(n*L),n,L)
X = matrix(rnorm(n*px),n,px)
epsilon = MASS::mvrnorm(n,rep(0,2),epsilonSigma)
D =  0.5 + Z %*% gamma + X %*% psi + epsilon[,1]
Y = -0.5 + Z %*% alpha + D * beta + X %*% phi + epsilon[,2]
endo.test.model <- endo.test(Y,D,Z,X,invalid = TRUE)
summary(endo.test.model)
```

`lineardata` *lineardata*

Description

Psuedo data provided by Youjin Lee, which is generated mimicing the structure of Framingham Heart Study data.

Usage

```
data(lineardata)
```

Format

A data.frame with 1445 observations on 12 variables:

- **Y:** The globulin level.
- **D:** The LDL-C level.
- **Z.1:** SNP genotypes.
- **Z.2:** SNP genotypes.
- **Z.3:** SNP genotypes.
- **Z.4:** SNP genotypes.
- **Z.5:** SNP genotypes.
- **Z.6:** SNP genotypes.
- **Z.7:** SNP genotypes.
- **Z.8:** SNP genotypes.
- **age:** the age of the subject.
- **sex:** the sex of the subject.

Source

The Framingham Heart Study data supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University.

Examples

```
data(lineardata)
```

`mroz` *mroz*

Description

Data provided by Ernst R. Berndt, which was used by Mroz (1987) and Wooldridge (2010)

Usage

```
data(mroz)
```

Format

A data.frame with 428 observations on 22 variables:

- **inlf:** =1 if in lab frce, 1975
- **hours:** hours worked, 1975
- **kidslt6:** # kids < 6 years
- **kidsge6:** # kids 6-18
- **age:** woman's age in yrs
- **educ:** years of schooling
- **wage:** est. wage from earn, hrs
- **repwage:** rep. wage at interview in 1976
- **hushrs:** hours worked by husband, 1975
- **husage:** husband's age
- **huseduc:** husband's years of schooling
- **huswage:** husband's hourly wage, 1975
- **faminc:** family income, 1975
- **mtr:** fed. marg. tax rte facing woman
- **motheduc:** mother's years of schooling
- **fatheduc:** father's years of schooling
- **unem:** unem. rate in county of resid.
- **city:** =1 if live in SMSA
- **exper:** actual labor mkt exper
- **nwifeinc:** (faminc - wage*hours)/1000
- **lwage:** log(wage)
- **expersq:** exper^2

Source

https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111531041

References

Mroz, Thomas, (1987), The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica*, 55, issue 4, p. 765-99.
Jeffrey M Wooldridge, 2010. "Econometric Analysis of Cross Section and Panel Data," MIT Press Books, The MIT Press, edition 2, volume 1, number 0262232588, December.

Examples

```
data(mroz)
```

nonlineardata *nonlineardata*

Description

Psuedo data provided by Youjin Lee, which is generated mimicing the structure of Framingham Heart Study data.

Usage

```
data(nonlineardata)
```

Format

A data.frame with 3733 observations on 9 variables:

- **Y:** The incidence of cardiovascular diseases.
- **bmi:** The BMI level.
- **insulin:** The insulin level.
- **Z.1:** SNP genotypes.
- **Z.2:** SNP genotypes.
- **Z.3:** SNP genotypes.
- **Z.4:** SNP genotypes.
- **age:** the age of the subject.
- **sex:** the sex of the subject.

Source

The Framingham Heart Study data supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University.

Examples

```
data(nonlineardata)
```

`pretest` *Prestest estimator*

Description

This function implements the pretest estimator by comparing the control function and the TSLS estimators.

Usage

```
pretest(formula, alpha = 0.05)
```

Arguments

`formula`	A formula describing the model to be fitted.
`alpha`	The significant level. (default = `0.05`)

Details

For example, the formula `Y ~ D + I(D^2)+X|Z+I(Z^2)+X` describes the model where $Y = \alpha_0 + D\beta_1 + D^2\beta_2 + X\phi + u$ and $D = \gamma_0 + Z\gamma_1 + Z^2\gamma_2 + X\psi + v$. Here, the outcome is `Y`, the endogenous variables are `D` and `I(D^2)`, the baseline covariates are `X`, and the instrument variables are `Z`. The formula environment follows the formula environment in the ivreg function in the AER package. The linear term of the endogenous variable, for example, `D`, must be included in must be included in the formula for the outcome model.

Value

`pretest` returns an object of class "pretest", which is a list containing the following components:

`coefficients`	The estimate of the coefficients in the outcome model.
`vcov`	The estimated covariance matrix of coefficients.
`Hausman.stat`	The Hausman test statistic used to test the validity of control function.
`p.value`	The p-value of Hausman test.
`cf.check`	the indicator that the control function is valid.

References

Guo, Z. and D. S. Small (2016), Control function instrumental variable estimation of nonlinear causal effect models, *The Journal of Machine Learning Research* 17(1), 3448–3482.

Examples

```
Y <- mroz[,"lwage"]
D <- mroz[,"educ"]
Z <- as.matrix(mroz[,c("motheduc","fatheduc","huseduc")])
X <- as.matrix(mroz[,c("exper","expersq","age")])
pretest.model <- pretest(Y~D+I(D^2)+X|Z+I(Z^2)+X)
summary(pretest.model)
```

`ProbitControl` *Causal inference in probit outcome models with possibly invalid IVs*

Description

Perform causal inference in the probit outcome model with possibly invalid IVs under the majority rule.

Usage

```
ProbitControl(
  Y,
  D,
  Z,
  X = NULL,
  intercept = TRUE,
  invalid = FALSE,
  d1 = NULL,
  d2 = NULL,
  w0 = NULL,
  bs.Niter = 40
)
```

Arguments

`Y`	The outcome observation, a vector of length n.
`D`	The treatment observation, a vector of length n.
`Z`	The instrument observation of dimension $n \times p_z$.
`X`	The covariates observation of dimension $n \times p_x$.
`intercept`	Whether the intercept is included. (default = `TRUE`)
`invalid`	If `TRUE`, the method is robust to the presence of possibly invalid IVs; If `FALSE`, the method assumes all IVs to be valid. (default = `FALSE`)
`d1`	A treatment value for computing CATE(d1,d2|w0).
`d2`	A treatment value for computing CATE(d1,d2|w0).
`w0`	A vector for computing CATE(d1,d2|w0).
`bs.Niter`	The number of bootstrap resampling size for computing the confidence interval.

Value

`ProbitControl` returns an object of class "SpotIV", which is a list containing the following components:

`betaHat`	The estimate of the model parameter in front of the treatment.
`beta.sdHat`	The estimated standard error of betaHat.
`cateHat`	The estimate of CATE(d1,d2|w0).
`cate.sdHat`	The estimated standard deviation of `cateHat`.
`SHat`	The estimated set of relevant IVs.
`VHat`	The estimated set of relevant and valid IVs.
`Maj.pass`	The indicator that the majority rule is satisfied.

References

Li, S., Guo, Z. (2020), Causal Inference for Nonlinear Outcome Models with Possibly Invalid Instrumental Variables, Preprint *arXiv:2010.09922*.

Examples

```
Y <- mroz[,"lwage"]
D <- mroz[,"educ"]
Z <- as.matrix(mroz[,c("motheduc","fatheduc","huseduc","exper","expersq")])
X <- mroz[,"age"]
Y0 <- as.numeric((Y>median(Y)))
d2 = median(D); d1 = d2+1;
w0 = apply(cbind(Z,X)[which(D == d2),], 2, mean)
Probit.model <- ProbitControl(Y0,D,Z,X,d1 = d1,d2 = d2,w0 = w0)
summary(Probit.model)
```

`SearchingSampling` *Searching-Sampling*

Description

Construct Searching and Sampling confidence intervals for the causal effect, which provides the robust inference of the treatment effect in the presence of invalid instrumental variables in both low-dimensional and high-dimensional settings. It is robust to the mistakes in separating valid and invalid instruments.

Usage

```
SearchingSampling(
  Y,
  D,
  Z,
  X = NULL,
  intercept = TRUE,
  method = c("OLS", "DeLasso", "Fast.DeLasso"),
  robust = FALSE,
  Sampling = TRUE,
  alpha = 0.05,
  CI.init = NULL,
  a = 0.6,
  rho = NULL,
  M = 1000,
  prop = 0.1,
  filtering = TRUE,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
method	The method used to estimate the reduced form parameters. "OLS" stands for ordinary least squares, "DeLasso" stands for the debiased Lasso estimator, and "Fast.DeLasso" stands for the debiased Lasso estimator with fast algorithm. (default = "OLS")
robust	If TRUE, the method is robust to heteroskedastic errors. If FALSE, the method assumes homoskedastic errors. (default = FALSE)
Sampling	If TRUE, use the proposed sampling method; else use the proposed searching method. (default=TRUE)
alpha	The significance level (default=0.05)
CI.init	An initial range for beta. If NULL, it will be generated automatically. (default=NULL)
a	The grid size for constructing beta grids. (default=0.6)
rho	The shrinkage parameter for the sampling method. (default=NULL)
M	The resampling size for the sampling method. (default = 1000)
prop	The proportion of non-empty intervals used for the sampling method. (default=0.1)

filtering	Filtering the resampled data or not. (default=TRUE)
tuning.1st	The tuning parameter used in the 1st stage to select relevant instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)
tuning.2nd	The tuning parameter used in the 2nd stage to select valid instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)

Details

When robust = TRUE, the method will be input as 'OLS'. For rho, M, prop, and filtering, they are required only for Sampling = TRUE. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01\log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

SearchingSampling returns an object of class "SS", which is a list containing the following components:

ci	1-alpha confidence interval for beta.
SHat	The set of selected relevant IVs.
VHat	The initial set of selected relevant and valid IVs.
check	The indicator that the plurality rule is satisfied.

References

Guo, Z. (2021), Causal Inference with Invalid Instruments: Post-selection Problems and A Solution Using Searching and Sampling, Preprint *arXiv:2104.06911*.

Examples

```
data("lineardata")
Y <- lineardata[,"Y"]
D <- lineardata[,"D"]
Z <- as.matrix(lineardata[,c("Z.1","Z.2","Z.3","Z.4","Z.5","Z.6","Z.7","Z.8")])
X <- as.matrix(lineardata[,c("age","sex")])
Searching.model <- SearchingSampling(Y,D,Z,X, Sampling = FALSE)
summary(Searching.model)
Sampling.model <- SearchingSampling(Y,D,Z,X)
summary(Sampling.model)
```

SpotIV *SpotIV method for causal inference in semi-parametric outcome model*

Description

Perform causal inference in the semi-parametric outcome model with possibly invalid IVs.

Usage

```
SpotIV(
  Y,
  D,
  Z,
  X = NULL,
  intercept = TRUE,
  invalid = FALSE,
  d1,
  d2,
  w0,
  M.est = TRUE,
  M = 2,
  bs.Niter = 40,
  bw = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
invalid	If TRUE, the method is robust to the presence of possibly invalid IVs; If FALSE, the method assumes all IVs to be valid. (default = FALSE)
d1	A treatment value for computing CATE(d1,d2|w0).
d2	A treatment value for computing CATE(d1,d2|w0).
w0	A value of measured covariates and instruments for computing CATE(d1,d2|w0).
M.est	If TRUE, M is estimated based on BIC, otherwise M is specified by input value of M. (default = TRUE)
M	The dimension of indices in the outcome model, from 1 to 3. (default = 2)
bs.Niter	The number of bootstrap resampling size for computing the confidence interval. (default = 40)
bw	A (M+1) by 1 vector bandwidth specification. (default = NULL)

Value

SpotIV returns an object of class "SpotIV", which "SpotIV" is a list containing the following components:

betaHat	The estimate of the model parameter in front of the treatment.
cateHat	The estimate of CATE(d1,d2|w0).
cate.sdHat	The estimated standard error of cateHat.
SHat	The set of relevant IVs.
VHat	The set of relevant and valid IVs.
Maj.pass	The indicator that the majority rule is satisfied.

References

Li, S., Guo, Z. (2020), Causal Inference for Nonlinear Outcome Models with Possibly Invalid Instrumental Variables, Preprint *arXiv:2010.09922*.

Examples

```
## Not run:
Y <- mroz[,"lwage"]
D <- mroz[,"educ"]
Z <- as.matrix(mroz[,c("motheduc","fatheduc","huseduc","exper","expersq")])
X <- mroz[,"age"]
Y0 <- as.numeric((Y>median(Y)))
d2 = median(D); d1 = d2+1;
w0 = apply(cbind(Z,X)[which(D == d2),], 2, mean)
SpotIV.model <- SpotIV(Y0,D,Z[,-5],X,d1 = d1,d2 = d2,w0 = w0[-5])
summary(SpotIV.model)

## End(Not run)
```

TSHT *Two-Stage Hard Thresholding*

Description

Perform Two-Stage Hard Thresholding method, which provides the robust inference of the treatment effect in the presence of invalid instrumental variables.

Usage

```
TSHT(
  Y,
  D,
  Z,
  X,
  intercept = TRUE,
  method = c("OLS", "DeLasso", "Fast.DeLasso"),
  voting = c("MaxClique", "MP", "Conservative"),
  robust = FALSE,
  alpha = 0.05,
  tuning.1st = NULL,
  tuning.2nd = NULL
)
```

Arguments

Y	The outcome observation, a vector of length n.
D	The treatment observation, a vector of length n.
Z	The instrument observation of dimension $n \times p_z$.
X	The covariates observation of dimension $n \times p_x$.
intercept	Whether the intercept is included. (default = TRUE)
method	The method used to estimate the reduced form parameters. "OLS" stands for ordinary least squares, "DeLasso" stands for the debiased Lasso estimator, and "Fast.DeLasso" stands for the debiased Lasso estimator with fast algorithm. (default = "OLS")
voting	The voting option used to estimate valid IVs. 'MP' stands for majority and plurality voting, 'MaxClique' stands for finding maximal clique in the IV voting matrix, and 'Conservative' stands for conservative voting procedure. Conservative voting is used to get an initial estimator of valid IVs in the Searching-Sampling method. (default= 'MaxClique').
robust	If TRUE, the method is robust to heteroskedastic errors. If FALSE, the method assumes homoskedastic errors. (default = FALSE)
alpha	The significance level for the confidence interval. (default = 0.05)
tuning.1st	The tuning parameter used in the 1st stage to select relevant instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)
tuning.2nd	The tuning parameter used in the 2nd stage to select valid instruments. If NULL, it will be generated data-dependently, see Details. (default=NULL)

Details

When robust = TRUE, the method will be input as 'OLS'. When voting = MaxClique and there are multiple maximum cliques, betaHat,beta.sdHat,ci, and VHat will be list objects where each element of list corresponds to each maximum clique. As for tuning parameter in the 1st stage and 2nd stage, if do not specify, for method "OLS" we adopt $\sqrt{\log n}$ for both tuning parameters, and for other methods we adopt $\max\left(\sqrt{2.01 \log p_z}, \sqrt{\log n}\right)$ for both tuning parameters.

Value

TSHT returns an object of class "TSHT", which is a list containing the following components:

betaHat	The estimate of treatment effect.
beta.sdHat	The estimated standard error of betaHat.
ci	The 1-alpha confidence interval for beta.
SHat	The set of selected relevant IVs.
VHat	The set of selected relevant and valid IVs.
voting.mat	The voting matrix.
check	The indicator that the majority rule is satisfied.

References

Guo, Z., Kang, H., Tony Cai, T. and Small, D.S. (2018), Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting, *J. R. Stat. Soc. B*, 80: 793-815.

Examples

```
data("lineardata")
Y <- lineardata[,"Y"]
D <- lineardata[,"D"]
Z <- as.matrix(lineardata[,c("Z.1","Z.2","Z.3","Z.4","Z.5","Z.6","Z.7","Z.8")])
X <- as.matrix(lineardata[,c("age","sex")])
TSHT.model <- TSHT(Y=Y,D=D,Z=Z,X=X)
summary(TSHT.model)
```

Index