Package 'controlfunctionIV'

May 8, 2026

Type Package

Title Control Function Methods with Possibly Invalid Instrumental Variables

Version 0.1.1

Description Inference with control function methods for nonlinear outcome models when the model is known ('Guo and Small' (2016) <doi:10.48550/arXiv.1602.01051>) and when unknown but semiparametric ('Li and Guo' (2021) <doi:10.48550/arXiv.2010.09922>).

License GPL-3

Encoding UTF-8

LazyData true

Imports dr, orthoDr, stats, AER, Formula

RoxygenNote 7.2.2

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

Depends R (>= 2.10)

NeedsCompilation no

Author Taehyeon Koo [aut],
Sai Li [aut],
Dylan Small [ctb],
Zijian Guo [aut, cre, cph]

Maintainer Zijian Guo <zijguo@stat.rutgers.edu>

Repository CRAN

Date/Publication 2022-12-20 03:20:02 UTC

Contents

cf *Control-Function*

Description

Implement the control function method for the 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 is `D`, the baseline covariates are `X`, and the instrument variables are `Z`. The formula environment follows that in the ivreg function in the AER package. The endogenous variable `D` must be in the first term of 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`.

Value

`cf` returns an object of class "cf", 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.
CausalEffect	The causal effect when the treatment changes from `d1` to `d2`.
CausalEffect.sd	The standard error of the causal effect estimator.
CausalEffect.ci	The 95% confidence interval of the causal effect.

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

```
data("nonlineardata")
Y <- log(nonlineardata[,"insulin"])
D <- nonlineardata[,"bmi"]
Z <- as.matrix(nonlineardata[,c("Z.1","Z.2","Z.3","Z.4")])
X <- as.matrix(nonlineardata[,c("age","sex")])
cf.model <- cf(Y~D+I(D^2)+X|Z+I(Z^2)+X)
summary(cf.model)
```

`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 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 is `D`, the baseline covariates are `X`, and the instrument variables are `Z`. The formula environment follows that in the ivreg function in the AER package. The endogenous variable `D` must be in the first term of 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 the extra IV generated by the control function.
p.value	The p-value of the Hausman test.
cf.check	The indicator that the extra IV generated by 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

```
data("nonlineardata")
Y <- log(nonlineardata[,"insulin"])
D <- nonlineardata[,"bmi"]
Z <- as.matrix(nonlineardata[,c("Z.1","Z.2","Z.3","Z.4")])
X <- as.matrix(nonlineardata[,c("age","sex")])
pretest.model <- pretest(Y~D+I(D^2)+X|Z+I(Z^2)+X)
summary(pretest.model)
```

`Probit.cf` *Causal inference in probit outcome models with possibly invalid IVs*

Description

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

Usage

```
Probit.cf(
  Y,
  D,
  Z,
  X = NULL,
  intercept = TRUE,
  invalid = TRUE,
  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 = `TRUE`)
`d1`	A treatment value for computing CATE(d1,d2|w0).
`d2`	A treatment value for computing CATE(d1,d2|w0).
`w0`	A vector of the instruments and baseline covariates for computing CATE(d1,d2|w0).
`bs.Niter`	The bootstrap resampling size for constructing the confidence interval.

Value

`Probit.cf` 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

```
data("nonlineardata")
Y <- nonlineardata[,"CVD"]
D <- nonlineardata[,"bmi"]
Z <- as.matrix(nonlineardata[,c("Z.1","Z.2","Z.3","Z.4")])
X <- as.matrix(nonlineardata[,c("age","sex")])
d1 <- median(D)+1
d2 <- median(D)
w0 <- c(rep(0,4), 30, 1)
Probit.model <- Probit.cf(Y,D,Z,X,invalid = TRUE,d1 =d1, d2 = d2,w0 = w0)
summary(Probit.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 = TRUE,
  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 = TRUE)
d1	A treatment value for computing CATE(d1,d2|w0).
d2	A treatment value for computing CATE(d1,d2|w0).
w0	A vector of the instruments and baseline covariates 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 bootstrap resampling size for constructing the confidence interval.
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

```
data("nonlineardata")
Y <- nonlineardata[,"CVD"]
D <- nonlineardata[,"bmi"]
Z <- as.matrix(nonlineardata[,c("Z.1","Z.2","Z.3","Z.4")])
X <- as.matrix(nonlineardata[,c("age","sex")])
d1 <- median(D)+1
d2 <- median(D)
w0 <- c(rep(0,4), 30, 1)
SpotIV.model <- SpotIV(Y,D,Z,X,invalid = TRUE,d1 =d1, d2 = d2,w0 = w0)
summary(SpotIV.model)
```

Index