Package 'naivereg'

October 13, 2022

Type Package

Title Nonparametric Additive Instrumental Variable Estimator and Related IV Methods

Version 1.0.5

Author Qingliang Fan, KongYu He, Wei Zhong

Maintainer Qingliang Fan <michaelqfan@xmu.edu.cn>

Description In empirical studies, instrumental variable (IV) regression is the signature method to solve the endogeneity problem. If we enforce the exogeneity condition of the IV, it is likely that we end up with a large set of IVs without knowing which ones are good. Also, one could face the model uncertainty for structural equation, as large micro dataset is commonly available nowadays. This package uses adaptive group lasso and B-spline methods to select the nonparametric components of the IV function, with the linear function being a special case (naivereg). The package also incorporates two stage least squares estimator (2SLS), generalized method of moment (GMM), generalized empirical likelihood (GEL) methods post instrument selection, logistic-regression instrumental variables estimator (LIVE, for dummy endogenous variable problem), double-selection plus instrumental variable estimator (DS-IV) and double selection plus logistic regression instrumental variable estimator (DS-LIVE), where the double selection methods are useful for high-dimensional structural equation models. The naivereg is nonparametric version of 'ivregress' in 'Stata' with IV selection and high dimensional features. The package is based on the paper by Q. Fan and W. Zhong, ``Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective" (2018), Journal of Business & Economic Statistics <doi:10.1080/07350015.2016.1180991> as well as a series of working papers led by the same authors.

Imports grpreg,splines,gmm,stats,ncvreg,glmnet

License GPL (>= 2)

Depends R (>= 3.5.0)

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

NeedsCompilation no

Repository CRAN

Date/Publication 2020-03-18 15:40:14 UTC

R topics documented:

`DSIV` *Double-Selection Plus Instrumental Variable Estimator*

Description

A three-step approach to estimate the endogenous treatment effect using high-dimensional instruments and double selection. It is applicable in the following scenarios: first, there is a known endogeneity problem for the treatment variable. Second, the treatment effect model has a large number of control variables, such as the large micro survey data.

Usage

```
DSIV(
  y,
  x,
  z,
  D,
 family = c("gaussian", "binomial", "poisson", "multinomial", "cox", "mgaussian"),
  criterion = c("BIC", "EBIC"),
  alpha = 1,
  nlambda = 100,
  ...
)
```

Arguments

`y`	Response variable, an N x 1 vector.
`x`	Control variables, an N x p1 matrix.
`z`	Instrumental variables, an N x p2 matrix.
`D`	Endogenous treatment variable.

family	Quantitative for family="gaussian", or family="poisson" (non-negative counts). For family="binomial" should be either a factor with two levels, or a two-column matrix of counts or proportions (the second column is treated as the target class; for a factor, the last level in alphabetical order is the target class). For family="multinomial", can be a nc>=2 level factor, or a matrix with nc columns of counts or proportions. For either "binomial" or "multinomial", if y is presented as a vector, it will be coerced into a factor. For family="cox", y should be a two-column matrix with columns named 'time' and 'status'. The latter is a binary variable, with '1' indicating death, and '0' indicating right censored. The function Surv() in package survival produces such a matrix. For family="mgaussian", y is a matrix of quantitative responses.
criterion	The criterion by which to select the regularization parameter. One of "BIC", "EBIC", default is "BIC".
alpha	The elasticnet mixing parameter, with 0<=alpha<= 1. alpha=1 is the lasso penalty, and alpha=0 the ridge penalty.
nlambda	The number of lambda values, default is 100.
...	other arguments, see help(glmnet).

Details

The DS-IV algorithm consists of the following three steps: In the first step, regress the outcome variable y on control variables x using the regularization method, estimate the coefficients beta and select the important control variables set denoted by c1. In the second step, regress the treatment variable d on instrumental variables w and control variables x, estimate the optimal instrument d and obtain the second important control variables set denoted by cx. In the third step, obtain the DS-IV estimator of the endogenous of the endogenous treatment effect based on the estimated optimal instrument d and the union (c3) of the selected control variables.

Value

An object of type DSIV which is a list with the following components:

yhat	The estimated value of y.
betaD	The coefficient of endogenous variable D.
betaX	The coefficient of control variables x.
c1	Variable indication of the selected in the first step (control variables x).
cx	Variable indication of selected control variables in the second step.
cz	Variable indication of selected instrumental variables in the second step.
c2	Variable indication of the selected in the second step. The number less than or equal to p1 is an indication of control variables, the number greater than p1 and less than or equal to (p1 + p2) is an indication of instrument variables.
c3	Union of c1 and cx on control variables.
family	Same as above.
criterion	Same as above.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Wei Zhong, Yang Gao, Wei Zhou and Qingliang Fan (2020), "Endogenous Treatment Effect Estimation Using High-Dimensional Instruments and Double Selection", working paper

Examples

```
library(naivereg)
data("DSIVdata")
y=DSIVdata[,1]
x=DSIVdata[,2:51]
z=DSIVdata[,52:71]
D=DSIVdata[,72]
res = DSIV(y,x,z,D,family='gaussian', criterion='EBIC')
res$c1 #Variable indication of the selected in the first step (control variables x).
res$cx #Variable indication of selected control variables in the second step.
res$cz #Variable indication of selected instrumental variables in the second step.
res$c3 #Union of c1 and cx on control variables
```

DSIVdata *The data generating for the DSIV*

Description

##The data generation process is as follows

library(MASS)

n=100

mu<-rep(0,50)

var<-matrix(,50,50)

for(i in 1:50)

for(j in 1:50)

var[i,j] <-0.5^(abs(i-j))

x<-mvrnorm(n,mu,var)#generate x

mu<-rep(0,20)

varz<-matrix(,20,20)

for(i in 1:20)

for(j in 1:20)

varz[i,j] <-0.5^(abs(i-j))

z<-mvrnorm(n,mu,varz)#generate iv

mu1<-c(0,0)

v<-c(1,0.9,0.9,1)

var1<-matrix(v,2,2)

epsilon<-mvrnorm(n,mu1,var1)#generate error term

D=1.9*x[,2]+2.5*x[,3]+1.4*x[,5]+x[,6]+x[,1]+1.6*z[,1]+1.9*z[,3]+1.7*z[,2]+epsilon[,2]

y=0.75*D+1*x[,1]+2*x[,6]+.11*x[,2]+.18*x[,3]+.12*x[,5]+epsilon[,1]

- Columns 1: Response variable y, an Nx1 vector.
- Columns 2-51: control variables x, an Nxp1 matrix.
- Columns 52-71: Instrumental variables,an Nxp2 matrix.
- Columns 72: Endogenous treatment variable.

Usage

```
data(DSIVdata)
```

DSLIVE *DS-LIVE*

Description

Double selection plus logistic regression instrumental variable estimator (DS-LIVE). A three-step approach to estimate the dummy endogenous treatment effect using high-dimensional instruments in a penalized logistic regression model and double selection. This method accommodates the binary endogenous variable as well as the high-dimensionality for both the reduced form and structural equation models.

Usage

```
DSLIVE(
  y,
  x,
  z,
  D,
  criterion = c("BIC", "CV"),
  penalty = c("SCAD", "MCP", "lasso"),
 family = c("gaussian", "binomial", "poisson", "multinomial", "cox", "mgaussian"),
  alpha = 1,
  gamma = 3.7,
  nfolds = 10,
  nlambda = 100,
  ...
)
```

Arguments

y	Response variable, an N x 1 vector.
x	Control variables, an N x p1 matrix.
z	Instrumental variables, an N x p2 matrix.
D	Endogenous treatment variable, the value of endogenous variable is 0 or 1 (binary).
criterion	The criterion by which to select the regularization parameter. One of "BIC", "CV", CV means cross-validation, default is "BIC".
penalty	This parameter takes effect when the creterion is CV. Quantitative for family="gaussian", or family="poisson" (non-negative counts). For family="binomial" should be either a factor with two levels, or a two-column matrix of counts or proportions (the second column is treated as the target class; for a factor, the last level in alphabetical order is the target class). For family="multinomial", can be a nc>=2 level factor, or a matrix with nc columns of counts or proportions. For either "binomial" or "multinomial", if y is presented as a vector, it will be coerced into a factor. For family="cox", y should be a two-column matrix with columns named 'time' and 'status'. The latter is a binary variable, with '1' indicating death, and '0' indicating right censored. The function Surv() in package survival produces such a matrix. For family="mgaussian", y is a matrix of quantitative responses.
family	Only applied to the first step in the algorithm, the regression of y on x. Quantitative for family="gaussian", or family="poisson" (non-negative counts). For family="binomial" should be either a factor with two levels, or a two-column matrix of counts or proportions (the second column is treated as the target class; for a factor, the last level in alphabetical order is the target class). For family="multinomial", can be a nc>=2 level factor, or a matrix with nc columns of counts or proportions. For either "binomial" or "multinomial", if y is presented as a vector, it will be coerced into a factor. For family="cox", y should be a two-column matrix with columns named 'time' and 'status'. The latter is a binary variable, with '1' indicating death, and '0' indicating right censored. The function Surv() in package survival produces such a matrix. For family="mgaussian", y is a matrix of quantitative responses.
alpha	Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD penalty and the ridge, or L2 penalty. alpha=1 is equivalent to MCP/SCAD penalty, while alpha=0 would be equivalent to ridge regression. However, alpha=0 is not supported; alpha may be arbitrarily small, but not exactly 0.
gamma	The tuning parameter of the MCP/SCAD penalty. Default is 3.7.
nfolds	This parameter takes effect when the creterion is CV. The response number of folds - default is 10. Although nfolds can be as large as the sample size (leave-one-out CV), it is not recommended for large datasets. Smallest value allowable is nfolds=3.
nlambda	The number of lambda values, default is 100.
...	other arguments, see help(glmnet) or help(cv.ncvreg).

Details

The DS-IV algorithm consists of the following three steps: In the first step, it estimates the coefficients (betaX) and select the important control variables set (denoted by c1) which are helpful to predict the outcome variable y using regularization methods for the data (y; x). In the second step, using a penalized logistic regression model, it selects both important control variables x (the selected control variables set is denoted by cx) and instrumental variables z for the endogenous treatment D. This step is crucial in the algorithm. Because it can estimate the optimal instrument using high-dimensional IVs as well as select additional important control variables which might be missed in the first step but are nonetheless important to the treatment variable. In the third step, it computes the post-double-selection LIVE estimator for the dummy endogenous treatment effect based on the predicted treatment variable D and the union of selected control variables in the first two variable selection steps denoted by c3 = (c1 union cx).

Value

An object of type `DSLIVE` which is a list with the following components:

`betaD`	The coefficient of endogenous variable D.
`betaX`	The coefficient of control variables x.
`c1`	Variable indication of the selected in the first step (control variables x).
`cx`	Variable indication of selected control variables in the second step.
`cz`	Variable indication of selected instrumental variables in the second step.
`c2`	Variable indication of the selected in the second step. The number less than or equal to p1 is an indication of control variables, the number greater than p1 and less than or equal to (p1 + p2) is an indication of instrument variables.
`c3`	Union of c1 and cx on control variables.
`family`	Same as above.
`criterion`	Same as above.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Wei Zhong, Wei Zhou, Qingliang Fan and Yang Gao (2020), "Dummy Endogenous Treatment Effect Estimation Using High-Dimensional Instrumental Variables", working paper.

Examples

```
library(naivereg)
data("DSLIVEdata")
y=DSLIVEdata[,1]
x=DSLIVEdata[,2:201]
z=DSLIVEdata[,202:221]
D=DSLIVEdata[,222]
res = DSLIVE(y,x,z,D,family='gaussian', criterion='BIC')
```

```
res$c1 # Variable indication of the selected in the first step (control variables x).
res$cx # Variable indication of selected control variables in the second step.
res$cz # Variable indication of selected instrumental variables in the second step.
res$c3 # Union of c1 and cx on control variables.
```

DSLIVEdata *The data generating for the DSLIVE*

Description

##The data generation process is as follows mu<-rep(0,200)

var<-matrix(,200,200)

for(i in 1:200)

for(j in 1:200)

var[i,j] <-0.5^(abs(i-j))

x<-mvrnorm(100,mu,var)

mu<-rep(0,20)

varz<-matrix(,20,20)

for(i in 1:20)

for(j in 1:20)

varz[i,j] <-0.5^(abs(i-j))

z<-mvrnorm(100,mu,varz)

mu1<-c(0,0)

v<-c(1,0.9,0.9,1)

var1<-matrix(v,2,2)

epsilon<-mvrnorm(100,mu1,var1)

D=rep(0,nrow(x))

p=.9*x[,4]+.8*x[,1]+1.96*x[,2]+1.85*x[,3]+.7*x[,5]+1.16*z[,1]+.95*z[,3]+1.7*z[,2]+epsilon[,2]

for(kk in 1:length(p))

D[kk]<-rbinom(n=1,size = 1,prob = exp(p[kk])/(1+exp(p[kk])))

y=0.75*D+3*x[,1]+2*x[,5]+1.5*x[,4]+0*3*x[,7]+0*1.5*x[,8]+.15*x[,2]+.18*x[,3]+epsilon[,1]

- Columns 1: Response variable y, an Nx1 vector.
- Columns 2-201: control variables x, an Nxp1 matrix.
- Columns 202-221: Instrumental variables,an Nxp2 matrix.
- Columns 222: Endogenous treatment variable,the value of endogenous variable is 0 or 1 (binary).

Usage

```
data(DSLIVEdata)
```

IVselect *Selecting instrument variables using group lasso and B-splines in naivereg*

Description

This shows which IVs are selected in the naivereg function.

Usage

```
IVselect(
  z,
  x,
  max.degree = 10,
  criterion = c("BIC", "AIC", "GCV", "AICc", "EBIC"),
  df.method = c("default", "active"),
  penalty = c("grLasso", "grMCP", "grSCAD", "gel", "cMCP"),
  endogenous.index = c(),
  IV.intercept = FALSE,
  family = c("gaussian", "binomial", "poisson")
)
```

Arguments

z	The instrument variables matrix.
x	The design matrix.
max.degree	The upper limit value of degree of B-splines when using BIC/AIC to choose the tuning parameters, default is BIC.
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC", default is "BIC".
df.method	How should effective model parameters be calculated? One of: "active", which counts the number of nonzero coefficients; or "default", which uses the calculated df returned by grpreg, default is "default".
penalty	The penalty to be applied to the model. For group selection, one of grLasso, grMCP, or grSCAD. For bi-level selection, one of gel or cMCP. Default is " grLasso".
endogenous.index	Specify which variables in design matrix are endogenous variables, the variable corresponds to the value 1 is endogenous variables, the variable corresponds to the value 0 is exogenous variable, default is all endogenous variables.
IV.intercept	Intercept of instrument variables, default is “FALSE”.
family	Either "gaussian" or "binomial", depending on the response, default is "gaussian".

Details

See naivereg.

Value

An object of type `IVselect` which is a list with the following components:

`degree`	Degree of B-splines.
`criterion`	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc", "EBIC", default is "BIC".
`ind`	The index of selected instrument variables.
`ind.b`	The index of selected instrument variables after B-splines.
`IVselect`	The instrument variables after B-splines.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Q. Fan and W. Zhong (2018), "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective," Journal of Business & Economic Statistics, doi: 10.1080/07350015.2016.1180991.

Caner, M. and Q. Fan (2015), Hybrid GEL Estimators: Instrument Selection with Adaptive Lasso, Journal of Econometrics, 187, 256–274.

Examples

```
#IV selecting with group Lasso an B-splines
library(naivereg)
data("naivedata")
x=naivedata[,1]
y=naivedata[,2]
z=naivedata[,3:22]
IV = IVselect(z,x)
IV$IVselect #show the IV selected after B-splines
```

`LIVE` *Logistic-regression Instrumental Variables Estimator*

Description

Binary endogenous variables are commonly encountered in program evaluations using observational data. This is a two-stage approach to estimate the dummy endogenous treatment effect using high-dimensional instrumental variables (IV). In the first stage, we use a penalized logistic reduced form model to accommodate both the binary nature of the endogenous treatment and the high-dimensionality of instrumental variables. In the second stage, we replace the original treatment variable by its estimated propensity score and run a least squares regression to obtain a penalized Logistic-regression Instrumental Variables Estimator (LIVE). If the structural equation model is also high-dimensional, one could use DS-LIVE in this package for selecting both the control variables and IVs.

Usage

```
LIVE(
  y,
  x,
  z,
  penalty = c("SCAD", "MCP", "lasso"),
  nfolds = 5,
  endogenous.index = c(),
  gamma = 3.7,
  alpha = 1,
  lambda.min = 0.05,
  nlambda = 100,
  ...
)
```

Arguments

y	Response variable, an N x 1 vector.
x	The design matrix, including endogenous variable, the value of endogenous variable is 0 or 1 (binary).
z	The instrumental variables matrix.
penalty	The penalty to be applied to the model. Either "SCAD" (the default), "MCP", or "lasso".
nfolds	The response number of folds - default is 5. Although nfolds can be as large as the sample size (leave-one-out CV), it is not recommended for large datasets. Smallest value allowable is nfolds=3.
endogenous.index	Specify which variables in design matrix are endogenous variables, the variable corresponds to the value 1 is endogenous variables, the variable corresponds to the value 0 is exogenous variable, the default is all endogenous variables.
gamma	The tuning parameter of the MCP/SCAD penalty. Default is 3.7.
alpha	Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD penalty and the ridge, or L2 penalty. alpha=1 is equivalent to MCP/SCAD penalty, while alpha=0 would be equivalent to ridge

	regression. However, alpha=0 is not supported; alpha may be arbitrarily small, but not exactly 0.
`lambda.min`	The smallest value for lambda, as a fraction of lambda.max, default is 0.05.
`nlambda`	The number of lambda values, default is 100.
`...`	other arguments.

Details

This is a two stage estimation. In the first stage, a high-dimensional logistic reduced form model with penalty (such as SCAD, lasso, etc.) is used to approximate the optimal instrument. In the second stage, we replace the original treatment variable by its estimated propensity score and run a least squares regression to obtain the penalized Logistic-regression Instrumental Variables Estimator (LIVE). The large dimensional IV could be the original variables or the functional transformations such as series, B-spline functions, etc.

Value

An object of type `LIVE` which is a list with the following components:

`coefficients`	The coefficients of x.
`lambda.min`	The value of lambda that gives minimum cvm.
`ind`	The selected variables of z.
`Xhat`	The xhat estimated by z.
`IVnum`	The number of instrumented variables after filtering.
`penalty`	Same as above.
`alpha`	Same as above.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Wei Zhong, Wei Zhou, Qingliang Fan and Yang Gao (2020), "Dummy Endogenous Treatment Effect Estimation Using High-Dimensional Instrumental Variables", working paper.

Examples

```
#Logistic-regression Instrumental Variables Estimator
data("LIVEdata")
y=LIVEdata[,1]
x=LIVEdata[,2]
z=LIVEdata[,3:52]
res = LIVE(y,x,z,penalty='SCAD',gamma = 3.7,alpha = 1,lambda.min = 0.05)
```

LIVEdata *The data generating for the LIVE*

Description

The data generating for the LIVE.

- Columns 1: The response variable y.
- Columns 2: The design matrix x.
- Columns 3-52: TheThe instrumental variables matrix z.

Usage

```
data(LIVEdata)
```

naive.gel *Estimete the parameters with gel after IV selecting*

Description

Hybrid gel estimator after selecting IVs in the reduced form equation.

Usage

```
naive.gel(
  g,
  x,
  z,
  max.degree = 10,
  criterion = c("BIC", "AIC", "GCV", "AICc", "EBIC"),
  df.method = c("default", "active"),
  penalty = c("grLasso", "grMCP", "grSCAD", "gel", "cMCP"),
  endogenous.index = c(),
  IV.intercept = FALSE,
  family = c("gaussian", "binomial", "poisson"),
  ...
)
```

Arguments

g	A function of the form $g(\theta, x)$ and which returns a $n \times q$ matrix with typical element $g_i(\theta, x_t)$ for $i = 1, ...q$ and $t = 1, ..., n$. This matrix is then used to build the q sample moment conditions. It can also be a formula if the model is linear (see details gel).
x	The design matrix, without an intercept.

z	The instrument variables matrix.
max.degree	The upper limit value of degree of B-splines when using BIC/AIC to choose the tuning parameters, default is BIC.
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC","EBIC", "GCV", "AICc"; default is "BIC".
df.method	How should effective model parameters be calculated? One of: "active", which counts the number of nonzero coefficients; or "default", which uses the calculated df returned by grpreg. default is "default".
penalty	The penalty to be applied to the model. For group selection, one of grLasso, grMCP, or grSCAD. For bi-level selection, one of gel or cMCP. Default is " grLasso".
endogenous.index	Specify which variables in design matrix are endogenous variables, the variable corresponds to the value 1 is endogenous variables, the variable corresponds to the value 0 is exogenous variable, the default is all endogenous variables.
IV.intercept	Intercept of instrument variables, default is "FALSE".
family	Either "gaussian" or "binomial", depending on the response, default is "gaussian".
...	Arguments passed to gel (such as type,kernel...,detail see gel).

Details

See naivereg and gel

Value

An object of type `naive.gel` which is a list with the following components:

degree	Degree of B-splines.
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC"; default is "BIC".
ind	The index of selected instrument variables.
ind.b	The index of selected instrument variables after B-splines.
gel	Gel object, detail see gel.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Q. Fan and W. Zhong (2018), "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective," Journal of Business & Economic Statistics, doi: 10.1080/07350015.2016.1180991.

Caner, M. and Fan, Q. (2015), Hybrid GEL Estimators: Instrument Selection with Adaptive Lasso, Journal of Econometrics, Volume 187, 256–274.

Examples

```
# gel estimation after IV selection
n = 200
phi<-c(.2,.7)
thet <- 0.2
sd <- .2
set.seed(123)
x <- matrix(arima.sim(n = n, list(order = c(2,0,1), ar = phi, ma = thet, sd = sd)), ncol = 1)
y <- x[7:n]
ym1 <- x[6:(n-1)]
ym2 <- x[5:(n-2)]
H <- cbind(x[4:(n-3)], x[3:(n-4)], x[2:(n-5)], x[1:(n-6)])
g <- y ~ ym1 + ym2
x <- H
naive.gel(g, cbind(ym1,ym2),x, tet0 =c(0,.3,.6))
```

`naive.gmm` *Estimete the parameters with gmm after IV selecting*

Description

Hybrid gmm estimator after selecting IVs in the reduced form equation.

Usage

```
naive.gmm(
  g,
  x,
  z,
  max.degree = 10,
  criterion = c("BIC", "AIC", "GCV", "AICc", "EBIC"),
  df.method = c("default", "active"),
  penalty = c("grLasso", "grMCP", "grSCAD", "gel", "cMCP"),
  endogenous.index = c(),
  IV.intercept = FALSE,
  family = c("gaussian", "binomial", "poisson"),
  ...
)
```

Arguments

g	A function of the form $g(\theta, x)$ and which returns a $n \times q$ matrix with typical element $g_i(\theta, x_t)$ for $i = 1, ...q$ and $t = 1, ..., n$. This matrix is then used to build the q sample moment conditions. It can also be a formula if the model is linear (see details gmm).
x	The design matrix, without an intercept.
z	The instrument variables matrix.

max.degree	The upper limit value of degree of B-splines when using BIC/AIC to choose the tuning parameters, default is BIC.
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC", default is "BIC".
df.method	How should effective model parameters be calculated? One of: "active", which counts the number of nonzero coefficients; or "default", which uses the calculated df returned by grpreg, default is "default".
penalty	The penalty to be applied to the model. For group selection, one of grLasso, grMCP, or grSCAD. For bi-level selection, one of gel or cMCP, default is " grLasso".
endogenous.index	Specify which variables in design matrix are endogenous variables, the variable corresponds to the value 1 is endogenous variables, the variable corresponds to the value 0 is exogenous variable, the default is all endogenous variables.
IV.intercept	Intercept of instrument variables, default is “FALSE”.
family	Either "gaussian" or "binomial", depending on the response.default is " gaussian ".
...	Arguments passed to gmm (such as type,kernel...,detail see gmm).

Details

See naivereg and gmm.

Value

An object of type naive.gmm which is a list with the following components:

degree	Degree of B-splines.
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC", default is "BIC".
ind	The index of selected instrument variables.
ind.b	The index of selected instrument variables after B-splines.
gmm	Gmm object, detail see gmm.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Q. Fan and W. Zhong (2018), “Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective,” Journal of Business & Economic Statistics, doi: 10.1080/07350015.2016.1180991.

Caner, M. and Fan, Q. (2015), Hybrid GEL Estimators: Instrument Selection with Adaptive Lasso, Journal of Econometrics, Volume 187, 256–274.

Examples

```
# gmm estimation after IV selection
data("naivedata")
x=naivedata[,1]
y=naivedata[,2]
z=naivedata[,3:22]
naive.gmm(y~x+x^2,cbind(x,x^2),z)
```

`naivedata` *The data generating for the naivereg*

Description

The z and residuals are both generated from a normal distribution.

x = sin(z1)+exp(z5)+z4^2+z15+log(z19+8)+e1.

y = 0.5*x+e2.

- Columns 1: The design matrix x.
- Columns 2: The response variable y.
- Columns 3-22: TheThe instrumental variables matrix z.

Usage

```
data(naivedata)
```

`naivereg` *Nonparametric additive instrumental variable estimator*

Description

NAIVE is the nonparametric additive instrumental variable estimator with the adaptive group Lasso. It uses group lasso and B-splines to obtain the valid instrument variables where BIC are applied to choose the tuning parameters. Then we get the two-stage least squares (2SLS) estimator with selected IV.

Usage

```
naivereg(
  y,
  x,
  z,
  max.degree = 10,
  intercept = TRUE,
  criterion = c("BIC", "AIC", "GCV", "AICc", "EBIC"),
```

```
  df.method = c("default", "active"),
  penalty = c("grLasso", "grMCP", "grSCAD", "gel", "cMCP"),
  endogenous.index = c(),
  IV.intercept = FALSE,
  family = c("gaussian", "binomial", "poisson")
)
```

Arguments

y	Response variable, a matrix N x 1.
x	The design matrix, without an intercept.
z	The instrument variables matrix.
max.degree	The upper limit value of degree of B-splines when using BIC/AIC to choose the tuning parameters, default is BIC.
intercept	Estimate with intercept or not, default is "TRUE".
criterion	The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC", default is "BIC".
df.method	How should effective model parameters be calculated? One of: "active", which counts the number of nonzero coefficients; or "default", which uses the calculated df returned by grpreg, default is "default".
penalty	The penalty to be applied to the model. For group selection, one of grLasso, grMCP, or grSCAD. For bi-level selection, one of gel or cMCP, default is " grLasso".
endogenous.index	Specify which variables in design matrix are endogenous variables, the variable corresponds to the value 1 is endogenous variables, thevariable corresponds to the value 0 is exogenous variable, the default is all endogenous variables.
IV.intercept	Intercept of instrument variables, default is “FALSE”.
family	Either "gaussian" or "binomial", depending on the response.default is " gaussian ".

Details

Consider the following structural equation with endogenous regressors$Y_i = x_u^T \beta + \epsilon_i$

To solve the endogeneity problem, instrumental variables are employed to obtain a consistent estimator of the population regression coefficient β. In practice, many potential instruments, including their series terms, may be recruited to approximate the optimal instrument and improve the precision of IV estimators. On the other hand, if many irrelevant instruments are contained in the reduced form equation, the approximation of the optimal instrument is generally unsatisfactory and the IV estimator is less efficient. In some cases where the dimensionality of $z_i$is even higher than the sample size, the linear IV method fails. To address these issues, the model sparsity is usually assumed and the penalized approaches can be applied to improve the efficiency of IV estimators. In this paper we propose the first-stage parsimonious predictive models and estimate optimal instruments in IV models with potentially more instruments than the sample size n.

The performance of the linear IV estimator in the finite sample is largely dependent on the validity of linearity assumption. This phenomenon motivated us to consider a more general nonlinear reduced

form equation to capture as much information of $x_i$as possible using instruments z_i under the high-dimensional model settings. This nonparametric idea for the reduced form model is consistent with Newey (1990). We consider the following nonparametric additive reduced form model with a large number of possible instruments.

$x_{il} = \mu_l + \sum_{j=1}^{p} f_{ij} z_{ij} + \xi_{il}$

To estimate the nonparametric components above, we use B-spline basis functions by following the idea of Huang, Horowitz, and Wei (2010). Let $S_n$be the space of polynomial splines of degrees L>1 and let $\phi_k, k = 1, 2, \ldots, m_n$be normalized B-spline basis functions for S_n, where m_n is the sum of the polynomial degree L and the number of knots. Let be the $\psi_k(z_{ij}) = \phi_k(z_{ij}) - n^{-1} \sum_{i=1}^{n} \phi_k(z_{ij})$centered B-spline basis functions for the th instrument. The model can then be rewritten using an approximate linear reduced form:

$x_{il} = \mu_l + \sum_{j=1}^{p} f_{ij} \sum_{k=1}^{m_n} (\gamma_{ij}) \psi(z_{ij}) + \xi_{il}$

To select the significant instruments and estimate the component functions simultaneously, we consider the following penalized objective function with an adaptive group Lasso penalty (Huang, Horowitz, and Wei 2010) for each l th endogenous variable

$L_n(\gamma_l; \lambda_n) = ||X_l - U\lambda_l||_2^2 + \lambda_n \sum_{j=1}^{p} \omega_{njl} ||\gamma_{jl}||_2$,where $\omega_{jnl} = ||\gamma_{jl}||_2^{-1}$,if $||\gamma_{jl}||_2 > 0$,$\omega_{jnl} = infty$,if $||\gamma_{jl}||_2 = 0$

By minimizing the penalized objective function with a group Lasso penalty we by minimizing the penalized objective function with a group Lasso penalty. And then we use the selected IV for β in the model with two-stage least squares (2SLS).

Value

An object of type `naivereg` which is a list with the following components:

`beta.endogenous`
: The coefficient of endogenous variable.

`beta.exogenous`
: The coefficient of exogenous variable.

`std.endogenous`
: The standard deviation of the endogenous variables' coefficients.

`std.exogenous`
: The standard deviation of the exogenous variables' coefficients.

`n`
: Number of samples.

`degree`
: Degree of B-splines.

`criterion`
: The criterion by which to select the regularization parameter. One of "AIC", "BIC", "GCV", "AICc","EBIC"; default is "BIC".

`ind`
: The index of selected instrument variables. Each row represents the instrumental variable selected for the corresponding endogenous variable. The order of the endogenous variables is from left to right in x.

`ind.b`
: The index of selected instrument variables after B-splines. Each row represents the instrumental variable selected for the corresponding endogenous variable. The order of the endogenous variables is from left to right in x.

`res`
: The difference between the predicted y and the actual y.

`t.endogenous`
: The t-value of the endogenous variables' coefficients.

`t.exogenous`
: The t-value of the exogenous variables' coefficients.

`endogenous.conf.interval.lower`
: The lower bound of 95 percent confidence interval for endogenous variables.

`endogenous.conf.interval.upper`
: The upper bound of 95 percent confidence interval for endogenous variables.

`exogenous.conf.interval.lower`
: The lower bound of 95 percent confidence interval for exogenous variables.

`exogenous.conf.interval.upper`
: The upper bound of 95 percent confidence interval for exogenous variables.

Author(s)

Qingliang Fan, KongYu He, Wei Zhong

References

Q. Fan and W. Zhong (2018), "Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective," Journal of Business & Economic Statistics, doi: 10.1080/07350015.2016.1180991.

Caner, M. and Fan, Q. (2015), Hybrid GEL Estimators: Instrument Selection with Adaptive Lasso, Journal of Econometrics, Volume 187, 256–274.

Examples

```
#naive regression
library(naivereg)
data("naivedata")
x=naivedata[,1]
y=naivedata[,2]
z=naivedata[,3:22]
#estimate with intercept
naive_intercept= naivereg(y,x,z)
#estimate without intercept,criterion:AIC
naive_without_intercept = naivereg(y,x,z,intercept=FALSE,criterion='AIC')
```

TradeAndGrowthData *Trade and growth data*

Description

This is a revisit of the seminal work of Frankel and Romer (1999, AER) "Does trade cause growth?" using new 2017 data. NAIVE method is detailed in Fan and Zhong (2018, JBES), "Nonparametric additive instrumental variable estimator: a group shrinkage estimation perspective". We found that trade still has significant effect on growth, but compared to the original studies, the effect of trade seems to be smaller in magnitude and NAIVE is able to select the important IVs which yield unbiased estimation. See details of the study including the invalid IV discussion in Fan and Wu (2020).

- N: log(economically active population in millions).

- A: log(land area).
- water: water area.
- coast: coastline.
- arable: arable percentage.
- border: land border.
- forest: forest percentage.
- lang: number of official languages.
- pm25: PM2.5.
- ww: waterways.
- rw: railway.
- in_water=T_hat*water.
- in_coast=T_hat*coast.
- in_arable=T_hat*arable.
- in_border=T_hat*border.
- in_forest=T_hat*forest.
- in_lang=T_hat*lang.
- ...

Usage

```
data(TradeAndGrowthData)
```

Index