Package ‘naivereg’
February 20, 2018

Type Package
Title Nonparametric Additive Instrumental Variable Estimator: A Group
Shrinkage Estimation Perspective
Version 1.0.1
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. 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. The package incorporates two stage least squares estimator (2SLS), general-
ized method of moment (GMM), generalized empirical likelihood (GEL) methods post instru-
ment selection. It is nonparametric version of ‘ivregress’ in ‘Stata’ with IV selection and high di-
mensional features. The package is based on the paper "Nonparametric Additive Instrumen-
tal Variable Estimator: A Group Shrinkage Estimation Perspective" (2017) published on-

Imports ggrepreg, splines, gmm, stats
License GPL (>= 2)
Encoding UTF-8
LazyData true
RoxygenNote 6.0.1
NeedsCompilation no
Repository CRAN
Date/Publication 2018-02-20 10:11:27 UTC

R topics documented:

IVselect ................................................................. 2
naive.gel .............................................................. 3
naive.gmm ............................................................ 5
naivedata ............................................................... 7
naivereg ................................................................. 7
IVselect

Selecting instrument variables using group lasso and B-splines

Description

Using group lasso and B-splines to obtain the group Lasso estimator where BIC or EBIC are applied to choose the tuning parameters and (degree of B-splines). And using the group Lasso estimator to obtain the valid instrument variables.

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, without an intercept
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", or "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"

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", or "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**


**Examples**

```r
#IV selecting with group Lasso an B-splines
library(naivereg)
data("naivedata")
x=naivedata[,1]
y=naivedata[,2]
z=naivedata[,3:102]
IV = IVselect(z,x)
IV$IVselect #show the IV selected after B-splines
```

**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", "GCV", "AICc", or "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 gel (such as type, kernel..., detail see gel).

Details

See naive.reg 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", or "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**


**Examples**

```r
# gel estimate after IV selecting
n = 200
phi <- c(0.2, 0.7)
theta <- 0.2
sd <- .2
set.seed(123)
x <- matrix(arima.sim(n = n, list(order = c(2, 0, 1), ar = phi, ma = theta, sd = sd)), ncol = 1)
y <- x[n/2:n]
ym1 <- x[6:(n-1)]
ym2 <- x[5:(n-2)]
H <- cbind(c(1, n-3), c(3, n-4), x[2:(n-5)], x[1:(n-6)])
g <- y ~ ym1 + ym2
x <- cbind(y, H)
naive.gel(g, cbind(ym1, ym2), x, theta = c(0.3, 0.6))
```

---

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

---

**Description**

Using gmm to make use of selected tools

**Usage**

```r
naive.gmm(g, x, z, max.degree = 10, criterion = c("BIC", "AIC", "GCV", "AICc", "EBIC"), df.method = c("default", "active"), penalty = c("grLasso", "gLasso", "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", or "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 naive.reg 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", or "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


**naivedata**

**Examples**

```r
# gmm after IV selecting
data("naivedata")
x = naivedata[,1]
y = naivedata[,2]
z = naivedata[,3:102]
naive.gmm(y = x + x^2, cbind(x, x^2), z)
naive.gmm(y = exp(x) + x, cbind(x, exp(x)), z)
```

---

**Description**

Data generated by Monte Carlo, the first column is the response variable, the second column is the design matrix, and the rest are the instrumental variables.

**Usage**

```r
data(naivedata)
```

---

**naivereg**

**Nonparametric additive instrumental variable estimator**

**Description**

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

**Usage**

```r
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 Nx1
- **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", or "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"

Details

Consider the following structural equation with endogenous regressors $Y_i = x_{ui}^T \beta + \epsilon_i$

To solve the endogeneity problem, instrumental variables are employed to obtain a consistent estimator of the population regression coefficient $\beta$. 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.

$$x_{ii} = \mu_i + \sum_{j=1}^{p} f_{ij} z_{ij} + \xi_{ii}$$
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 $l$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:

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

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 $\beta$ in the model with two-stage least squares (2SLS).

**Value**

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

- **coefficients** a vector of coefficients
- **ste** the standard deviation of the 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", or "EBIC"; default is "BIC".
- **ind** the index of selected instrument variables
- **ind.b** the index of selected instrument variables after B-splines
- **res** the difference between the predicted $y$ and the actual $y$

**Author(s)**

Qingliang Fan, KongYu He, Wei Zhong

**References**


Examples

# naive regression
library(naivereg)
data("naivedata")
x = naivedata[,1]
y = naivedata[,2]
z = naivedata[,3:102]
# estimate with intercept
naive_intercept = naivereg(y,x,z)
# estimate without intercept, criterion: EBIC
naive_without_intercept = naivereg(y,x,z,intercept=FALSE,criterion='EBIC')
Index

*Topic datasets
  naivedata, 7

IVselect, 2

naive.gel, 3
naive.gmm, 5
naivedata, 7
naivereg, 7