Package ‘PLRModels’

February 19, 2015

Type Package
Title Statistical inference in partial linear regression models
Version 1.1
Date 2014-01-01
Author German Aneiros Perez and Ana Lopez Cheda
Maintainer German Aneiros Perez <ganeiros@udc.es>
Description This package provides statistical inference tools applied to Partial Linear Regression (PLR) models. Specifically, point estimation, confidence intervals estimation, bandwidth selection, goodness-of-fit tests and analysis of covariance are considered. Kernel-based methods, combined with ordinary least squares estimation, are used and time series errors are allowed. In addition, these techniques are also implemented for both parametric (linear) and nonparametric regression models.
License GPL
NeedsCompilation no
Repository CRAN
Date/Publication 2014-01-01 21:37:58

R topics documented:

- PLRModels-package
- barnacles1
- barnacles2
- np.ancova
- np.cv
- np.est
- np.gcv
- np.gof
- par.ancova
- par.ci
- par.est
Par.gof ................................................................. 25
plrm.ancova ......................................................... 27
plrm.beta ............................................................ 32
plrm.ci ............................................................... 34
plrm.cv ............................................................... 37
plrm.est ............................................................. 41
plrm.gcv ............................................................. 44
plrm.gof ............................................................. 47

Index 51

PLRModels-package Statistical inference in partial linear regression models

Description
This package provides statistical inference tools applied to Partial Linear Regression (PLR) models. Specifically, point estimation, confidence intervals estimation, bandwidth selection, goodness-of-fit tests and analysis of covariance are considered. Kernel-based methods, combined with ordinary least squares estimation, are used and time series errors are allowed. In addition, these techniques are also implemented for both parametric (linear) and nonparametric regression models.

Details

Package: PLRModels
Type: Package
Version: 1.1
Date: 2014-01-01
License: GPL

The most important functions are those directly related with the PLR models; that is, plrm.gcv, plrm.cv, plrm.beta, plrm.est, plrm.gof, plrm.ancova and plrm.ci. Although the other functions included in the package are auxiliary ones, they can be used independently.

Author(s)
Authors: German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
Maintainer: German Aneiros Perez <ganeiros@udc.es>

barnacles1 Sales of barnacles in Cedeira
barnacles2

Description

Information about sales and prices of barnacles in two galician towns for each month from 2004 to 2013. The data have been transformed using the logarithm function.

Usage

data(barnacles2)

Format

A matrix containing 3 columns:
barnacles2[, 1] contains the number of sales (in kg) of barnacles in Cedeira’s fish market;
barnacles2[, 2] contains the prices (in euro/kg) of the barnacles in Cedeira’s fish market;
barnacles2[, 3] contains the number of sales (in kg) of barnacles in Carino’s fish market.

Source

http://www.pescadegalicia.com/

barnacles2

Sales of barnacles in Cangas

Description

Information about sales and prices of barnacles in two galician towns for each month from 2004 to 2013. The data have been transformed using the logarithm function.

Usage

data(barnacles2)

Format

A matrix containing 3 columns:
barnacles2[, 1] contains the number of sales (in kg) of barnacles in Cango’s fish market;
barnacles2[, 2] contains the prices (in euro/kg) of the barnacles in Cango’s fish market;
barnacles2[, 3] contains the number of sales (in kg) of barnacles in Baiona’s fish market.

Source

http://www.pescadegalicia.com/
Nonparametric analysis of covariance

Description

This routine tests the equality of $L$ nonparametric regression curves ($m_1, \ldots, m_L$) from samples $(Y_{ki}, t_i): i = 1, \ldots, n, k = 1, \ldots, L$, where:

$$Y_{ki} = m_k(t_i) + \epsilon_{ki}.$$

The unknown functions $m_k$ are smooth, fixed equally spaced design is considered, and the random errors, $\epsilon_{ki}$, are allowed to be time series. The test statistic used for testing the null hypothesis, $H_0 : m_1 = \ldots = m_L$, derives from a Cramer-von-Mises-type functional based on different distances between nonparametric estimators of the regression functions.

Usage

```r
np.ancova(data = data, h.seq = NULL, w = NULL, estimator = "NW", kernel = "quadratic", time.series = FALSE, Tau.eps = NULL, h0 = NULL, lag.max = 50, p.max = 3, q.max = 3, ic = "BIC", num.lb = 10, alpha = 0.05)
```

Arguments

- `data` contains the values of the response variable, $Y_k$, for each model $k (k = 1, \ldots, L)$.
- `data[, , L+1]` contains the values of the explanatory (common) variable, $t$, for each model $k (k = 1, \ldots, L)$.
- `h.seq` the statistic test is performed using each bandwidth in the vector `h.seq` (the same bandwidth is used to estimate all the regression functions). If `NULL` (the default), 10 equidistant values between 0 and the first half of the range of $t_i$ are considered.
- `w` support interval of the weight function in the test statistic. If `NULL` (the default), $(q_{0.1}, q_{0.9})$ is considered, where $q_p$ denotes the quantile of order $p$ of $t_i$.
- `estimator` allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.
- `kernel` allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.
- `time.series` it denotes whether the data are independent (FALSE) or if data is a time series (TRUE). The default is FALSE.
- `Tau.eps` `Tau.eps[, , k]` contains the sum of autocovariances associated to the random errors of the regression model $k (k = 1, \ldots, L)$. If `NULL` (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted nonparametric regression model and, then, it obtains the sum of the autocovariances of such ARMA model.
\textbf{np.ancova}

- \textbf{h0} if \texttt{Tau. eps=NULL}, \texttt{h0} contains the pilot bandwidth used for obtaining the residuals to construct the default for \texttt{Tau. eps}. If \texttt{NULL} (the default), a quarter of the range of \( t_i \) is considered.

- \textbf{lag.max} if \texttt{Tau. eps=NULL}, \texttt{lag. max} contains the maximum delay used to construct the default for \texttt{Tau. eps}. The default is 50.

- \textbf{p.max} if \texttt{Tau. eps=NULL}, the ARMA models are selected between the models ARMA(\( p,q \)) with 0\( \leq p \leq p.\max \) and 0\( \leq q \leq q.\max \). The default is 3.

- \textbf{q.max} if \texttt{Tau. eps=NULL}, the ARMA models are selected between the models ARMA(\( p,q \)) with 0\( \leq p \leq p.\max \) and 0\( \leq q \leq q.\max \). The default is 3.

- \textbf{ic} if \texttt{Tau. eps=NULL}, \texttt{ic} contains the information criterion used to suggest the ARMA models. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).

- \textbf{num.lb} if \texttt{Tau. eps=NULL}, it checks the suitability of the selected ARMA models according to the Ljung-Box test and the t-test. It uses up to \texttt{num. lb} delays in the Ljung-Box test. The default is 10.

- \textbf{alpha} if \texttt{Tau. eps=NULL}, \texttt{alpha} contains the significance level which the ARMA models are checked. The default is 0.05.

\textbf{Details}

A weight function (specifically, the indicator function \( 1_{[w[1],w[2]]} \)) is introduced in the test statistic to allow elimination (or at least significant reduction) of boundary effects from the estimate of \( m(t_i) \).

If \texttt{Tau. eps=NULL} and the routine is not able to suggest an approximation for \texttt{Tau. eps}, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct \texttt{Tau. eps}, the procedures suggested in Muller and Stadmuller (1988) and Herrmann et al. (1992) can be followed.

For more details, see Vilar-Fernandez and Gonzalez-Manteiga (2004).

\textbf{Value}

A list with a dataframe containing:

- \texttt{h. seq} sequence of bandwidths used in the test statistic.
- \texttt{Q.m} values of the test statistic (one for each bandwidth in \texttt{h. seq}).
- \texttt{Q.m.normalised} normalised value of \texttt{Q.m}.
- \texttt{p.value} p-values of the corresponding statistic tests (one for each bandwidth in \texttt{h. seq}).

Moreover, if \texttt{data} is a time series and \texttt{Tau. eps} is not specified:

- \texttt{pv.Box.test} p-values of the Ljung-Box test for the model fitted to the residuals.
- \texttt{pv.t.test} p-values of the t-test for the model fitted to the residuals.
- \texttt{ar.ma} ARMA orders for the model fitted to the residuals.

\textbf{Author(s)}

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

Other related functions are `np.est`, `par.ancova` and `plrm.ancova`.

Examples

```r
# EXAMPLE 1: REAL DATA
data <- matrix(10,120,2)
data(barnacles1)
barnacles1 <- as.matrix(barnacles1)
data[,1] <- barnacles1[,1]
data <- diff(data, 12)
data[,2] <- 1:nrow(data)
data2 <- matrix(10,120,2)
data(barnacles2)
barnacles2 <- as.matrix(barnacles2)
data2[,1] <- barnacles2[,1]
data2 <- diff(data2, 12)
data2[,2] <- 1:nrow(data2)
data3 <- matrix(0, nrow(data), ncol(data)+1)
data3[,1] <- data[,1]
data3[,2:3] <- data2
np.ancova(data=data3)

# EXAMPLE 2: SIMULATED DATA
## Example 2.1: dependent data: true null hypothesis

set.seed(1234)
# We generate the data
n <- 100
t <- (1:n)-0.5/n
ml <- function(t) {0.25*t*(1-t)}
f <- ml(t)
espsilon1 <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
```
\[
\begin{align*}
y_1 & \leftarrow f + \epsilon_1 \\
\epsilon_2 & \leftarrow \text{arima.sim(} \text{list(order = c(0,0,1), ma=0.5), sd = 0.02, n = n)} \\
y_2 & \leftarrow f + \epsilon_2 \\
data_{\text{eq}} & \leftarrow \text{cbind(}y_1, y_2, t) \\
\end{align*}
\]

# We apply the test
np.ancova(data_{eq}, \text{time.series=TRUE})

### Example 2.2: dependent data: false null hypothesis

# We generate the data
n <- 100 \\
t <- ((1:n)-0.5)/n \\
\(m_3\) <- function(t) \{0.25*t*(1-t)\} \\
\(m_4\) <- function(t) \{0.25*t*(1-t)*0.75\} \\
\(f_3\) <- \(m_3(t)\) \\
\(\epsilon_3\) <- \text{arima.sim(} \text{list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)} \\
y_3 & \leftarrow f_3 + \epsilon_3 \\
\(f_4\) <- \(m_4(t)\) \\
\(\epsilon_4\) <- \text{arima.sim(} \text{list(order = c(0,0,1), ma=0.5), sd = 0.02, n = n)} \\
y_4 & \leftarrow f_4 + \epsilon_4 \\
data_{\text{neq}} & \leftarrow \text{cbind(}y_3, y_4, t) \\
# We apply the test
np.ancova(data_{neq}, \text{time.series=TRUE})

---

**np.cv**

*Cross-validation bandwidth selection in nonparametric regression models*

**Description**

From a sample \((Y_i, t_i) : i = 1, \ldots, n\), this routine computes, for each \(l_n\) considered, an optimal bandwidth for estimating \(m\) in the regression model

\[
Y_i = m(t_i) + \epsilon_i.
\]

The regression function, \(m\), is a smooth but unknown function, and the random errors, \(\epsilon_i\), are allowed to be time series. The optimal bandwidth is selected by means of the leave-(\(2l_n + 1\))-out cross-validation procedure. Kernel smoothing is used.

**Usage**

np.cv(data = data, h.seq = NULL, num.h = 50, w = NULL, num.ln = 1, ln.0 = 0, step.ln = 2, estimator = "NW", kernel = "quadratic")
Arguments

data  
data[, 1] contains the values of the response variable, \( Y \);
data[, 2] contains the values of the explanatory variable, \( t \).

h.seq  
sequence of considered bandwidths in the CV function. If NULL (the default), num. h equidistant values between zero and a quarter of the range of \( t_i \) are considered.

num.h  
number of values used to build the sequence of considered bandwidths. If h. seq is not NULL, num. h=length(h. seq). Otherwise, the default is 50.

w  
support interval of the weight function in the CV function. If NULL (the default), (\( q_{0.1}, q_{0.9} \)) is considered, where \( q_p \) denotes the quantile of order \( p \) of \( t_i \).

num.ln  
number of values for \( l_n \): \( 2l_n + 1 \) observations around each point \( t_i \) are eliminated to estimate \( m(t_i) \) in the CV function. The default is 1.

ln.0  
minimum value for \( l_n \). The default is 0.

step.ln  
distance between two consecutive values of \( l_n \). The default is 2.

estimator  
allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

kernel  
allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

Details

A weight function (specifically, the indicator function \( 1_{[w[1],w[2]]} \)) is introduced in the CV function to allow elimination (or at least significant reduction) of boundary effects from the estimate of \( m(t_i) \).

For more details, see Chu and Marron (1991).

Value

h.opt  
dataframe containing, for each \( l_n \) considered, the selected value for the bandwidth.

CV.opt  
CV.opt[k] is the minimum value of the CV function when the k-th value of \( l_n \) is considered.

CV  
matrix containing the values of the CV function for each bandwidth and \( l_n \) considered.

w  
support interval of the weight function in the CV function.

h.seq  
sequence of considered bandwidths in the CV function.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

Other related functions are: `np.est`, `np.gcv`, `plrm.est`, `plrm.gcv` and `plrm.cv`.

Examples

```r
# EXAMPLE 1: REAL DATA
data <- matrix(10, 120, 2)
data(barnacles1)
barnacles1 <- as.matrix(barnacles1)
data[, 1] <- barnacles1[, 1]
data <- diff(data, 12)
data[, 2] <- 1:nrow(data)

aux <- np.cv(data, ln.0=1, step.ln=1, num.ln=2)
aux$opt
plot.ts(aux$CV)

par(mfrow=c(2,1))
plot(aux$h.seq, aux$CV[, 1], xlab="h", ylab="CV", type="l", main="ln=1")
plot(aux$h.seq, aux$CV[, 2], xlab="h", ylab="CV", type="l", main="ln=2")

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
m <- function(t) {0.25*t*(1-t)}
f <- m(t)

epsilon <- rnorm(n, 0, 0.01)
y <- f + epsilon
data_ind <- matrix(c(y, t), nrow=100)

# We apply the function
a <- np.cv(data_ind)
a$CV.opt

CV <- a$CV
h <- a$h.seq
plot(h, CV, type="l")
```
np.est

Nonparametric estimate of the regression function

Description

This routine computes estimates for \( m(newt_j) \) \((j = 1, ..., J)\) from a sample \((Y_i, t_i) : i = 1, ..., n,\) where:

\[ Y_i = m(t_i) + \epsilon_i. \]

The regression function, \( m, \) is a smooth but unknown function, and the random errors, \( \epsilon_i, \) are allowed to be time series. Kernel smoothing is used.

Usage

\[
\text{np.est(data = data, h.seq = NULL, newt = NULL, estimator = "NW", kernel = "quadratic")}
\]

Arguments

- **data**: \( \text{data[, 1]} \) contains the values of the response variable, \( Y; \) \( \text{data[, 2]} \) contains the values of the explanatory variable, \( t. \)
- **h.seq**: the considered bandwidths. If NULL (the default), only one bandwidth, selected by means of the cross-validation procedure, is used.
- **newt**: values of the explanatory variable where the estimates are obtained. If NULL (the default), the considered values will be the values of \( \text{data[, 2]} \).
- **estimator**: allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.
- **kernel**: allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

Details


Value

\( \text{YHAT: a length(newt) x length(h.seq) matrix containing the estimates for } m(newt_j) \)

\((j = 1, ..., \text{length(newt)})\) using the different bandwidths in \( \text{h.seq}. \)

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

Other related functions are: `np.gcv`, `np.cv`, `plrm.est`, `plrm.gcv` and `plrm.cv`.

Examples

```r
# EXAMPLE 1: REAL DATA
data <- matrix(10,120,2)
data(barnacles1)
barnacles1 <- as.matrix(barnacles1)
data[,1] <- barnacles1[,1]
data <- diff(data, 12)
data[,2] <- 1:nrow(data)

aux <- np.gcv(data)
h <- aux$h.opt
ajuste <- np.est(data=data, h=h)
plot(data[,2], ajuste, type="l", xlab="t", ylab="m(t)"
plot(data[,1], ajuste, xlab="y", ylab="y.hat", main="y.hat vs y")
abline(0,1)
residuos <- data[,1] - ajuste
mean(residuos^2)/var(data[,1])

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
m <- function(t) {0.25*t*(1-t)}
f <- m(t)

epsilon <- rnorm(n, 0, 0.01)
y <- f + epsilon
data_ind <- matrix(c(y,t),nrow=100)

# We estimate the nonparametric component of the PLR model
# (CV bandwidth)
est <- np.est(data_ind)
plot(t, est, type="l", lty=2, ylab=""
points(t, 0.25*t*(1-t), type="l")
legend(x="topleft", legend = c("m", "m hat"), col=c("black", "black"), lty=1)
```

np.est
np.gcv

Generalized cross-validation bandwidth selection in nonparametric regression models

Description

From a sample \((Y_i, t_i) : i = 1, ..., n\), this routine computes an optimal bandwidth for estimating \(m\) in the regression model

\[
Y_i = m(t_i) + \epsilon_i.
\]

The regression function, \(m\), is a smooth but unknown function. The optimal bandwidth is selected by means of the generalized cross-validation procedure. Kernel smoothing is used.

Usage

```
np.gcv(data = data, h.seq=NULL, num.h = 50, estimator = "NW", kernel = "quadratic")
```

Arguments

- `data` : `data[, 1]` contains the values of the response variable, \(Y\); `data[, 2]` contains the values of the explanatory variable, \(t\).
- `h.seq` : sequence of considered bandwidths in the GCV function. If `NULL` (the default), `num.h` equidistant values between zero and a quarter of the range of \(t_i\) are considered.
- `num.h` : number of values used to build the sequence of considered bandwidths. If `h.seq` is not `NULL`, `num.h=length(h.seq)`. Otherwise, the default is 50.
- `estimator` : allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.
- `kernel` : allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

Details


Value

- `h.opt` : selected value for the bandwidth.
- `GCV.opt` : minimum value of the GCV function.
- `GCV` : vector containing the values of the GCV function for each considered bandwidth.
- `h.seq` : sequence of considered bandwidths in the GCV function.
np.gcv

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are: np.est, np.cv, plrm.est, plrm.gcv and plrm.cv.

Examples

# EXAMPLE 1: REAL DATA
data <- matrix(10,120,2)
data(barnacles1)
barnacles1 <- as.matrix(barnacles1)
data[,1] <- barnacles1[,1]
data <- diff(data, 12)
data[,2] <- 1:nrow(data)
aux <- np.gcv(data)
aux$h.opt
plot(aux$h.seq, aux$GCV, xlab="h", ylab="GCV", type="l")

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
m <- function(t) {0.25*t*(1-t)}
f <- m(t)
epsilon <- rnorm(n, 0, 0.01)
y <- f + epsilon
data_ind <- matrix(c(y,t),nrow=100)

# We apply the function
a <- np.gcv(data_ind)
a$GCV.opt

GCV <- a$GCV
h <- a$h.seq
plot(h, GCV, type="l")
**np.gof**

*Goodness-of-Fit tests in nonparametric regression models*

**Description**

This routine tests the equality of a nonparametric regression curve, \( m \), and a given function, \( m_0 \), from a sample \((Y_i, t_i) : i = 1, ..., n\), where:

\[
Y_i = m(t_i) + \epsilon_i.
\]

The unknown function \( m \) is smooth, fixed equally spaced design is considered, and the random errors, \( \epsilon_i \), are allowed to be time series. The test statistic used for testing the null hypothesis, \( H_0 : m = m_0 \), derives from a Cramer-von-Mises-type functional distance between a nonparametric estimator of \( m \) and \( m_0 \).

**Usage**

```r
np.gof(data = data, m0 = NULL, h.seq = NULL, w = NULL, estimator = "NW", kernel = "quadratic", time.series = FALSE, Tau.eps = NULL, h0 = NULL, lag.max = 50, p.max = 3, q.max = 3, ic = "BIC", num.lb = 10, alpha = 0.05)
```

**Arguments**

- `data`: `data[, 1]` contains the values of the response variable, \( Y \);
  `data[, 2]` contains the values of the explanatory variable, \( t \).
- `m0`: the considered function in the null hypothesis. If `NULL` (the default), the zero function is considered.
- `h.seq`: the statistic test is performed using each bandwidth in the vector `h.seq`. If `NULL` (the default), 10 equidistant values between zero and a quarter of the range of \( t_i \) are considered.
- `w`: support interval of the weight function in the test statistic. If `NULL` (the default), \((q_{0.1}, q_{0.9})\) is considered, where \( q_p \) denotes the quantile of order \( p \) of \( t_i \).
- `estimator`: allows us the choice between "NW" (Nadaraya-Watson) or "LLP" (Local Linear Polynomial). The default is "NW".
- `kernel`: allows us the choice between "gaussian", "quadratic" (Epanechnikov kernel), "triweight" or "uniform" kernel. The default is "quadratic".
- `time.series`: it denotes whether the data are independent (`FALSE`) or if data is a time series (`TRUE`). The default is `FALSE`.
- `Tau.eps`: it contains the sum of autocovariances associated to the random errors of the regression model. If `NULL` (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted nonparametric regression model and, then, it obtains the sum of the autocovariances of such ARMA model.
**Details**

A weight function (specifically, the indicator function \(1_{[w[1],w[2]]}\)) is introduced in the test statistic to allow elimination (or at least significant reduction) of boundary effects from the estimate of \(m(t_i)\).

If Tau.eps=NULL and the routine is not able to suggest an approximation for Tau.eps, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct Tau.eps, the procedures suggested in Muller and Stadmuller (1988) and Herrmann et al. (1992) can be followed.

The implemented statistic test particularizes that one in Gonzalez Manteiga and Vilar Fernandez (1995) to the case where the considered class in the null hypothesis has only one element.

**Value**

A list with a dataframe containing:

- h.seq: sequence of bandwidths used in the test statistic.
- Q.m: values of the test statistic (one for each bandwidth in h.seq).
- Q.m.normalised: normalised value of Q.m.
- p.value: p-values of the corresponding statistic tests (one for each bandwidth in h.seq).

Moreover, if data is a time series and Tau.eps is not specified:

- pv.Box.test: p-values of the Ljung-Box test for the model fitted to the residuals.
- pv.t.test: p-values of the t.test for the model fitted to the residuals.
- ar.ma: ARMA orders for the model fitted to the residuals.

**Author(s)**

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

Other related functions are np.est.par.gof and plrm.gof.

Examples

```r
# EXAMPLE 1: REAL DATA
data <- matrix(10,120,2)
data(barnacles1)
barnacles1 <- as.matrix(barnacles1)
data[,1] <- barnacles1[,1]
data <- diff(data, 12)
data[,2] <- 1:nrow(data)

np.gof(data)
```

```r
# EXAMPLE 2: SIMULATED DATA
# Example 2a: dependent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
m <- function(t) {0.25*t*(1-t)}
f <- m(t)
f.function <- function(u) {0.25*u*(1-u)}
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y <- f + epsilon
data <- cbind(y,t)

# Example 2a.1: true null hypothesis
```

```r
```
**Description**

This routine tests the equality of $L$ vector coefficients, $(\beta_1, \ldots, \beta_L)$, from samples $(Y_{ki}, X_{k1}, \ldots, X_{kp})$: $i = 1, \ldots, n, k = 1, \ldots, L$, where: 

$$\beta_k = (\beta_{k1}, \ldots, \beta_{kp})$$

is an unknown vector parameter and

$$Y_{ki} = X_{k1} * \beta_{k1} + \ldots + X_{kp} * \beta_{kp} + \epsilon_{ki}.$$

The random errors, $\epsilon_{ki}$, are allowed to be time series. The test statistic used for testing the null hypothesis, $H_0 : \beta_1 = \ldots = \beta_L$, derives from the asymptotic normality of the ordinary least squares estimator of $\beta_k$ ($k = 1, \ldots, L$), this result giving a $\chi^2$-test.

**Usage**

```r
par.ancova(data = data, time.series = FALSE, Var.Cov.eps = NULL, p.max = 3, q.max = 3, ic = "BIC", num.lb = 10, alpha = 0.05)
```

**Arguments**

- `data` 
  - `data[, 1, k]` contains the values of the response variable, $Y_k$, for each model $k = 1, \ldots, L$;
  - `data[, 2:(p+1), k]` contains the values of the explanatory variables, $X_{k1}, \ldots, X_{kp}$, for each model $k = 1, \ldots, L$.
- `time.series` 
  - it denotes whether the data is independent (FALSE) or if data is a time series (TRUE). The default is FALSE.
- `Var.Cov.eps` 
  - `Var.Cov.eps[, , k]` contains the $n \times n$ matrix of variances-covariances associated to the random errors of the regression model $k = 1, \ldots, L$. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted linear regression model and, then, it obtains the var-cov matrix of such ARMA model.
- `p.max` 
  - if `Var.Cov.eps=NULL`, the ARMA models are selected between the models ARMA($p,q$) with $0 \leq p \leq p.max$ and $0 \leq q \leq q.max$. The default is 3.
- `q.max` 
  - if `Var.Cov.eps=NULL`, the ARMA models are selected between the models ARMA($p,q$) with $0 \leq p \leq p.max$ and $0 \leq q \leq q.max$. The default is 3.
ic if Var.Cov.eps=NULL, ic contains the information criterion used to suggest the ARMA models. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).

num.lb if Var.Cov.eps=NULL, it checks the suitability of the ARMA models according to the Ljung-Box and the t.test. It uses up to num.lb delays in the Ljung-Box test. The default is 10.

alpha if Var.Cov.eps=NULL, alpha contains the significance level (default is 0.05) which the ARMA models are checked.

Details
If Var.Cov.eps=NULL and the routine is not able to suggest an approximation for Var.Cov.eps, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct Var.Cov.eps, the procedure suggested in Domowitz (1982) can be followed.

The implemented procedure particularizes the parametric test in the routine plrm.ancova to the case where is known that the nonparametric components in the corresponding PLR models are null.

Value
A list with a dataframe containing:

Q.beta value of the test statistic.
p.value p-value of the corresponding statistic test.

Moreover, if data is a time series and Var.Cov.eps is not specified:
pv.Box.test p-values of the Ljung-Box test for the model fitted to the residuals.
pv.t.test p-values of the t.test for the model fitted to the residuals.
ar.ma ARMA orders for the model fitted to the residuals.

Author(s)
German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References

See Also
Other related functions are np.ancova and plrm.ancova.
Examples

# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data[,1],1, data[,-1])
data(barnacles2)
data2 <- as.matrix(barnacles2)
data2 <- diff(data2, 12)
data2 <- cbind(data2[,1],1, data2[,-1])
data3 <- array(0, c(nrow(data), ncol(data), 2))
data3[,1] <- data
ndata3[,2] <- data2
par.ancova(data=data3)

# EXAMPLE 2: SIMULATED DATA
## Example 2a: dependent data - true null hypothesis
set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)
x1 <- matrix(rnorm(200, 0, 1), nrow=n)
sum1 <- x1%*%beta
epsilon1 <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y1 <- sum1 + epsilon1
data1 <- cbind(y1,x1)
x2 <- matrix(rnorm(200, 1, 2), nrow=n)
sum2 <- x2%*%beta
epsilon2 <- arima.sim(list(order = c(0,0,1), ma=0.5), sd = 0.02, n = n)
y2 <- sum2 + epsilon2
data2 <- cbind(y2,x2)
data_eq <- array(cbind(data1, data2), c(100, 3, 2))
# We apply the test
par.ancova(data_eq, time.series=TRUE)

## Example 2a: dependent data - false null hypothesis
# We generate the data
n <- 100
beta3 <- c(0.05, 0.01)
beta4 <- c(0.05, 0.02)
par.ci

Confidence intervals estimation in linear regression models

Description

This routine obtains a confidence interval for the value $a^T \beta$, by asymptotic distribution and bootstrap, from a sample $(Y_i, X_{i1}, ..., X_{ip}) : i = 1, ..., n$, where:

$$a = (a_1, ..., a_p)^T$$

is an unknown vector,

$$\beta = (\beta_1, ..., \beta_p)^T$$

is an unknown vector parameter and

$$Y_i = X_{i1} \beta_1 + ... + X_{ip} \beta_p + \epsilon_i.$$

The random errors, $\epsilon_i$, are allowed to be time series.

Usage

par.ci(data=data, seed=123, CI="AD", B=1000, N=50, a=NULL, p.arima=NULL, q.arima=NULL, p.max=3, q.max=3, alpha=0.05, alpha2=0.05, num.lb=10, ic="BIC", Var.Cov.eps=NULL)

Arguments

data

data[, 1] contains the values of the response variable, $Y$;
daughter[2:(p+1)] contains the values of the explanatory variables, $X_1, ..., X_p$.

seed

the considered seed.
par.ci

method to obtain the confidence interval. It allows us to choose between: “AD” (asymptotic distribution), “B” (bootstrap) or “all” (both). The default is “AD”.

B

number of bootstrap replications. The default is 1000.

N

Truncation parameter used in the finite approximation of the MA(infinite) expression of \( \epsilon \).

a

Vector which, multiplied by \( \beta \), is used for obtaining the confidence interval of this result.

p.arima

the considered \( p \) to fit the model ARMA(\( p,q \)).

q.arima

the considered \( q \) to fit the model ARMA(\( p,q \)).

p.max

if \( \text{Var.Cov.} \text{.eps} = \text{NULL} \), the ARMA models are selected between the models ARMA(\( p, q \)) with \( 0 \leq p \leq p_{\text{max}} \) and \( 0 \leq q \leq q_{\text{max}} \). The default is 3.

q.max

if \( \text{Var.Cov.} \text{.eps} = \text{NULL} \), the ARMA models are selected between the models ARMA(\( p, q \)) with \( 0 \leq p \leq p_{\text{max}} \) and \( 0 \leq q \leq q_{\text{max}} \). The default is 3.

alpha

1 - \( \alpha \) is the confidence level of the confidence interval. The default is 0.05.

alpha2

significance level used to check (if needed) the ARMA model fitted to the residuals. The default is 0.05.

num.lb

if \( \text{Var.Cov.} \text{.eps} = \text{NULL} \), it checks the suitability of the selected ARMA model according to the Ljung-Box test and the t-test. It uses up to \( \text{num.lb} \) delays in the Ljung-Box test. The default is 10.

ic

if \( \text{Var.Cov.} \text{.eps} = \text{NULL} \), \( \text{ic} \) contains the information criterion used to suggest the ARMA model. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).

\text{Var.Cov.} \text{.eps}

\( n \times n \) matrix of variances-covariances associated to the random errors of the regression model. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted regression model and, then, it obtains the var-cov matrix of such ARMA model.

Value

A list containing:

Bootstrap

a dataframe containing \( \text{ci\.inf} \) and \( \text{ci\.sup} \), the confidence intervals using bootstrap and \( p_{\text{opt}} \) and \( q_{\text{opt}} \) (the orders for the ARMA model fitted to the residuals).

AD

a dataframe containing \( \text{ci\.inf} \) and \( \text{ci\.sup} \), the confidence intervals using the asymptotic distribution and \( p_{\text{opt}} \) and \( q_{\text{opt}} \) (the orders for the ARMA model fitted to the residuals).

pv.Box.test

p-values of the Ljung-Box test for the model fitted to the residuals.

pv.t.test

p-values of the t.test for the model fitted to the residuals.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

A related function is `plrm.ci`.

Examples

```R
# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data[,1],1,data[,-1])

## Not run: par.ci(data, a=c(1,0,0), CI="all")
## Not run: par.ci(data, a=c(0,1,0), CI="all")
## Not run: par.ci(data, a=c(0,0,1), CI="all")

# EXAMPLE 2: SIMULATED DATA
## Example 2a: dependent data

set.seed(123)
# We generate the data
n <- 100
beta <- c(0.5, 2)

x <- matrix(rnorm(200,0,3), nrow=n)
sum <- x%*%beta
sum <- as.matrix(sum)
eps <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.1, n = n)
eps <- as.matrix(eps)
y <- sum + eps
data_parci <- cbind(y,x)

## We estimate the confidence interval of a^T * beta in the PLR model
## Not run: par.ci(data, a=c(1,0), CI="all")
## Not run: par.ci(data, a=c(0,1), CI="all")
```

`par.est`  
*Estimation in linear regression models*
Description

This routine computes the ordinary least squares estimate for $\beta$ from a sample $(Y_i, X_{i1}, ..., X_{ip})$, $i = 1, ..., n$, where:

$$\beta = (\beta_1, ..., \beta_p)$$

is an unknown vector parameter and

$$Y_i = X_{i1} \ast \beta_1 + ... + X_{ip} \ast \beta_p + \epsilon_i.$$

The random errors, $\epsilon_i$, are allowed to be time series.

Usage

par.est(data = data)

Arguments

data data[, 1] contains the values of the response variable, $Y$;
data[, 2:(p+1)] contains the values of the explanatory variables, $X_1, ..., X_p$.

Details

See Seber (1977) and Judge et al. (1980).

Value

A vector containing the corresponding estimate.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are plrm.beta and plrm.est.
Examples

# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data[, 1], 1, data[, -1])

beta <- par.est(data=data)
beta
residuos <- data[, 1] - data[, -1]*beta
mean(residuos^2)/var(data[, 1])

fitted.values <- data[, -1]*beta
plot(data[, 1], fitted.values, xlab="y", ylab="y.hat", main="y.hat vs y")
abline(0,1)

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data
set.seed(1234)
# We generate the data	n <- 100
beta <- c(0.05, 0.01)

x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x*beta
epsilon <- rnorm(n, 0, 0.01)
y <- sum + epsilon
data_ind <- matrix(c(y, x), nrow=100)

# We estimate the parametric component of the PLR model
par.est(data_ind)

## Example 2b: dependent data
set.seed(1234)
# We generate the data
x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x*beta
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y <- sum + epsilon
data_dep <- matrix(c(y, x), nrow=100)

# We estimate the parametric component of the PLR model
par.est(data_dep)
Goodness-of-Fit tests in linear regression models

Description

This routine tests the equality of the vector of coefficients, $\beta$, in a linear regression model and a given parameter vector, $\beta_0$, from a sample $(Y_i, X_{i1}, ..., X_{ip}) : i = 1, ..., n$, where:

$$\beta = (\beta_1, ..., \beta_p)$$

is an unknown vector parameter and

$$Y_i = X_{i1} \ast \beta_1 + ... + X_{ip} \ast \beta_p + \epsilon_i.$$

The random errors, $\epsilon_i$, are allowed to be time series. The test statistic used for testing the null hypothesis, $H_0 : \beta = \beta_0$, derives from the asymptotic normality of the ordinary least squares estimator of $\beta$, this result giving a $\chi^2$-test.

Usage

par.gof(data = data, beta0 = NULL, time.series = FALSE,
Var.Cov.eps = NULL, p.max = 3, q.max = 3, ic = "BIC",
num.lb = 10, alpha = 0.05)

Arguments

data          data[, 1] contains the values of the response variable, $Y$;
data[, 2:(p+1)] contains the values of the explanatory variables, $X_1, ..., X_p$.
beta0         the considered parameter vector in the null hypothesis. If NULL (the default), the zero vector is considered.
time.series   it denotes whether the data are independent (FALSE) or if data is a time series (TRUE). The default is FALSE.
Var.Cov.eps   $n \times n$ matrix of variances-covariances associated to the random errors of the regression model. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted linear regression model and, then, it obtains the var-cov matrix of such ARMA model.
p.max         if Var.Cov.eps=NULL, the ARMA model is selected between the models ARMA(p,q) with 0<=p<=p.max and 0<=q<=q.max. The default is 3.
q.max         if Var.Cov.eps=NULL, the ARMA model is selected between the models ARMA(p,q) with 0<=p<=p.max and 0<=q<=q.max. The default is 3.
ic            if Var.Cov.eps=NULL, ic contains the information criterion used to suggest the ARMA model. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).
num.lb  if \texttt{Var.Cov.eps=NULL}, it checks the suitability of the selected ARMA model according to the Ljung-Box test and the t-test. It uses up to \texttt{num.lb} delays in the Ljung-Box test. The default is 10.

alpha  if \texttt{Var.Cov.eps=NULL}, \texttt{alpha} contains the significance level which the ARMA model is checked. The default is 0.05.

Details

If \texttt{Var.Cov.eps=NULL} and the routine is not able to suggest an approximation for \texttt{Var.Cov.eps}, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct \texttt{Var.Cov.eps}, the procedure suggested in Domowitz (1982) can be followed.

The implemented procedure particularizes the parametric test in the routine \texttt{plrm.gof} to the case where is known that the nonparametric component in the corresponding PLR model is null.

Value

A list with a dataframe containing:

- \texttt{Q.beta}  value of the test statistic.
- \texttt{p.value}  p-value of the corresponding statistic test.

Moreover, if \texttt{data} is a time series and \texttt{Var.Cov.eps} is not specified:

- \texttt{pv.Box.test}  p-values of the Ljung-Box test for the model fitted to the residuals.
- \texttt{pv.t.test}  p-values of the t.test for the model fitted to the residuals.
- \texttt{ar.ma}  ARMA orders for the model fitted to the residuals.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are \texttt{np.gof} and \texttt{plrm.gof}. 
Examples

```r
# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data[,1], 1, data[,1])

## Example 1.1: false null hypothesis
par.gof(data)
## Example 1.2: true null hypothesis
par.gof(data, beta0=c(0.05, 0.01))
```

# EXAMPLE 2: SIMULATED DATA
# Example 2a: dependent data

```r
set.seed(1234)
# We generate the data
n <- 100
beta <- c(0.05, 0.01)
x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x%*%beta
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y <- sum + epsilon
data <- cbind(y, x)

## Example 2a.1: true null hypothesis
par.gof(data, beta0=c(0.05, 0.01))

## Example 2a.2: false null hypothesis
par.gof(data)
```

### plrm.ancova

**Semiparametric analysis of covariance (based on PLR models)**

**Description**

From samples \((Y_{ki}, X_{k1i}, ..., X_{kip}, t_i) : i = 1, ..., n, k = 1, ..., L\), this routine tests the null hypotheses \(H_0: \beta_1 = ... = \beta_L\) and \(H_0: m_1 = ... = m_L\), where:

\[
\beta_k = (\beta_{k1}, ..., \beta_{kp})
\]

is an unknown vector parameter;

\[
m_k(\cdot)
\]

is a smooth but unknown function and

\[
Y_{ki} = X_{k1i} \ast \beta_{k1} + ... + X_{kip} \ast \beta_{kp} + m(t_i) + \epsilon_{ki}.
\]
Fixed equally spaced design is considered for the "nonparametric" explanatory variable, \( t \), and the random errors, \( \epsilon_{ki} \), are allowed to be time series. The test statistic used for testing \( H_0 : \beta_1 = \ldots = \beta_L \) derives from the asymptotic normality of an estimator of \( \beta_k \) \((k = 1, \ldots, L)\) based on both ordinary least squares and kernel smoothing (this result giving a \( \chi^2 \)-test). The test statistic used for testing \( H_0 : m_1 = \ldots = m_L \) derives from a Cramer-von-Mises-type functional based on different distances between nonparametric estimators of \( m_k \) \((k = 1, \ldots, L)\).

**Usage**

```r
plrm.ancova(data = data, t = t, b.seq = NULL, h.seq = NULL, 
  w = NULL, estimator = "NW", kernel = "quadratic", 
  time.series = FALSE, Var.Cov.ups = NULL, Tau.ups = NULL, 
  b0 = NULL, h0 = NULL, lag.max = 50, p.max = 3, q.max = 3, 
  ic = "BIC", num.lb = 10, alpha = 0.05)
```

**Arguments**

- **data** contains the values of the response variable, \( Y_k \), for each model \( k = 1, \ldots, L \).
- **data**, \([2:(p+1), k]\) contains the values of the "linear" explanatory variables, \( X_{k1}, \ldots, X_{kp} \), for each model \( k = 1, \ldots, L \).
- **t** contains the values of the "nonparametric" explanatory (common) variable, \( t \), for each model \( k = 1, \ldots, L \).
- **b.seq** is the statistic test for \( H_0 : \beta_1 = \ldots = \beta_L \) is performed using each bandwidth in the vector \( b.seq \). If NULL (the default) but \( h.seq \) is not NULL, it takes \( \text{b.seq} = \text{h.seq} \). If both \( b.seq \) and \( h.seq \) are NULL, 10 equidistant values between zero and a quarter of the range of \( t_i \) are considered.
- **h.seq** is the statistic test for \( H_0 : m_1 = \ldots = m_L \) is performed using each pair of bandwidths \( \{b.seq[j], h.seq[j]\} \). If NULL (the default) but \( b.seq \) is not NULL, it takes \( \text{h.seq} = \text{b.seq} \). If both \( b.seq \) and \( h.seq \) are NULL, 10 equidistant values between zero and a quarter of the range of \( t_i \) are considered for both \( b.seq \) and \( h.seq \).
- **w** is the support interval of the weight function in the test statistic for \( H_0 : m_1 = \ldots = m_L \). If NULL (the default), \((q_{0.1}, q_{0.9})\) is considered, where \( q_p \) denotes the quantile of order \( p \) of \( t_i \).
- **estimator** allows us the choice between "NW" (Nadaraya-Watson) or "LLP" (Local Linear Polynomial). The default is "NW".
- **kernel** allows us the choice between "gaussian", "quadratic" (Epanechnikov kernel), "triweight" or "uniform" kernel. The default is "quadratic".
- **time.series** it denotes whether the data are independent (FALSE) or if data is a time series (TRUE). The default is FALSE.
- **Var.Cov.ups** contains the \( n \times n \) matrix of variances-covariances associated to the random errors of the regression model \( k \) \((k = 1, \ldots, L)\). If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterion) to the residuals from the fitted regression model and, then, it obtains the var-cov matrix of such ARMA model.
plrm.ancova 29

Tau. eps \[\text{Tau. eps}[k]\] contains the sum of autocovariances associated to the random errors of the regression model \( k = 1, \ldots, L \). If\( \text{NULL}\) (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterion) to the residuals from the fitted regression model and, then, it obtains the sum of the autocovariances of such ARMA model.

\[ b_0 \]

if \( \text{Var. Cov. eps=NULL and/or Tau. eps=\text{NULL}} \), \( b_0 \) contains the pilot bandwidth for the estimator of \( \beta_k \) \( k = 1, \ldots, L \) used for obtaining the residuals to construct the default for \( \text{Var. Cov. eps and/or Tau. eps} \). If \( \text{NULL}\) (the default) but \( h_0 \) is not \( \text{NULL}\), it takes \( b_0 = h_0 \). If both \( b_0 \) and \( h_0 \) are \( \text{NULL}\), a quarter of the range of \( t_i \) is considered.

\[ h_0 \]

if \( \text{Var. Cov. eps=NULL and/or Tau. eps=\text{NULL}} \), \( h_0 \) contains the pair of pilot bandwidths for the estimator of \( m_k \) \( k = 1, \ldots, L \) used for obtaining the residuals to construct the default for \( \text{Var. Cov. eps and/or Tau. eps} \). If \( \text{NULL}\) (the default) but \( b_0 \) is not \( \text{NULL}\), it takes \( h_0 = b_0 \). If both \( b_0 \) and \( h_0 \) are \( \text{NULL}\), a quarter of the range of \( t_i \) is considered for both \( b_0 \) and \( h_0 \).

\[ \text{lag.max} \]

if \( \text{Tau. eps=\text{NULL}} \), \( \text{lag.max} \) contains the maximum delay used to construct the default for \( \text{Tau. eps} \). The default is 50.

\[ p.\text{max} \]

if \( \text{Var. Cov. eps=\text{NULL and/or Tau. eps=\text{NULL}}} \), the ARMA models are selected between the models ARMA\( (p,q) \) with \( 0 \leq p \leq p.\text{max} \) and \( 0 \leq q \leq q.\text{max} \). The default is 3.

\[ q.\text{max} \]

if \( \text{Var. Cov. eps=\text{NULL and/or Tau. eps=\text{NULL}}} \), the ARMA models are selected between the models ARMA\( (p,q) \) with \( 0 \leq p \leq p.\text{max} \) and \( 0 \leq q \leq q.\text{max} \). The default is 3.

\[ \text{ic} \]

if \( \text{Var. Cov. eps=\text{NULL and/or Tau. eps=\text{NULL}}} \), \( \text{ic} \) contains the information criterion used to suggest the ARMA models. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).

\[ \text{num.lb} \]

if \( \text{Var. Cov. eps=\text{NULL and/or Tau. eps=\text{NULL}}} \), it checks the suitability of the selected ARMA models according to the Ljung-Box test and the t-test. It uses up to \( \text{num.lb} \) delays in the Ljung-Box test. The default is 10.

\[ \alpha \]

if \( \text{Var. Cov. eps=\text{NULL and/or Tau. eps=\text{NULL}}} \), \( \alpha \) contains the significance level which the ARMA models are checked. The default is 0.05.

Details

A weight function (specifically, the indicator function \( 1_{[w[1],w[2]]} \)) is introduced in the test statistic for testing \( H_0 : m_1 = \ldots = m_L \) to allow elimination (or at least significant reduction) of boundary effects from the estimate of \( m_k(t_i) \).

If \( \text{Var. Cov. eps=} \text{NULL} \) and the routine is not able to suggest an approximation for \( \text{Var. Cov. eps} \), it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct \( \text{Var. Cov. eps} \), the procedure suggested in Aneiros-Perez and Vieu (2013) can be followed.

If \( \text{Tau. eps=} \text{NULL} \) and the routine is not able to suggest an approximation for \( \text{Tau. eps} \), it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct \( \text{Tau. eps} \), the procedures suggested in Aneiros-Perez (2008) can be followed.

Expressions for the implemented statistic tests can be seen in (15) and (16) in Aneiros-Perez (2008).
plrm.ancova

Value

A list with two dataframes:

parametric.test
a dataframe containing the bandwidths, the statistics and the p-values when one tests $H_0: \beta_1 = \ldots = \beta_L$.

nonparametric.test
a dataframe containing the bandwidths $b$ and $h$, the statistics, the normalised statistics and the p-values when one tests $H_0: m_1 = \ldots = m_L$.

Moreover, if data is a time series and Tau. eps or Var. Cov. eps are not specified:

pv.Box.test p-values of the Ljung-Box test for the model fitted to the residuals.

pv.t.test p-values of the t.test for the model fitted to the residuals.

ar.ma ARMA orders for the model fitted to the residuals.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are plrm.est, par.ancova and np.ancova.

Examples

# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

data(barnacles2)
data2 <- as.matrix(barnacles2)
data2 <- diff(data2, 12)
data2 <- cbind(data2, 1:nrow(data2))

data3 <- array(0, c(nrow(data), ncol(data)-1, 2))
data3[, , 1] <- data[, -4]
data3[, , 2] <- data2[, -4]
t <- data[, 4]
plrm.ancova(data=data, t=t)

# EXAMPLE 2: SIMULATED DATA
## Example 2a: dependent data - true null hypotheses

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)

m1 <- function(t) (0.25*t*(1-t))
f <- m1(t)
x1 <- matrix(rnorm(200,0,1), nrow=n)
sum1 <- x1*beta
epsilon1 <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y1 <- sum1 + f + epsilon1
data1 <- cbind(y1,x1)

x2 <- matrix(rnorm(200,1,2), nrow=n)
sum2 <- x2*beta
epsilon2 <- arima.sim(list(order = c(0,0,1), ma=0.5), sd = 0.02, n = n)
y2 <- sum2 + f + epsilon2
data2 <- cbind(y2,x2)

data_eq <- array(c(data1,data2), c(n,3,2))

# We apply the tests
plrm.ancova(data=data_eq, t=t, time.series=TRUE)

## Example 2b: dependent data - false null hypotheses

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
m3 <- function(t) (0.25*t*(1-t))
m4 <- function(t) (0.25*t*(1-t)*0.75)

beta3 <- c(0.05, 0.01)
beta4 <- c(0.05, 0.02)

x3 <- matrix(rnorm(200,0,1), nrow=n)
sum3 <- x3*beta3
f3 <- m3(t)
epsilon3 <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y3 <- sum3 + f3 + epsilon3
data3 <- cbind(y3,x3)

x4 <- matrix(rnorm(200,1,2), nrow=n)
sum4 <- x4*beta4
f4 <- m4(t)
epsilon4 <- arima.sim(list(order = c(0,0,1), ma=0.5), sd = 0.02, n = n)
y4 <- sum4 + f4 + epsilon4
data4 <- cbind(y4,x4)

data_neq <- array(c(data3,data4), c(n,3,2))

# We apply the tests
plrm.ancova(data=data_neq, t=t, time.series=TRUE)

plrm.beta

Semiparametric estimate for the parametric component of the regression function in PLR models

Description

This routine computes estimates for $\beta$ from a sample $(Y_i, X_{i1}, ..., X_{ip}, t_i) : i = 1, ..., n$, where:

$$\beta = (\beta_1, ..., \beta_p)$$

is an unknown vector parameter and

$$Y_i = X_{i1} \ast \beta_1 + ... + X_{ip} \ast \beta_p + m(t_i) + \epsilon_i.$$  

The nonparametric component, $m$, is a smooth but unknown function, and the random errors, $\epsilon_i$, are allowed to be time series. Ordinary least squares estimation, combined with kernel smoothing, is used.

Usage

plrm.beta(data = data, b.seq = NULL, estimator = "NW", kernel = "quadratic")

Arguments

data [ , 1] contains the values of the response variable, $Y$;
data [ , 2:(p+1)] contains the values of the "linear" explanatory variables, $X_1, ..., X_p$;
data [ , p+2] contains the values of the "nonparametric" explanatory variable, $t$.

b.seq vector of bandwidths for estimating $\beta$. If NULL (the default), only one estimate of $\beta$ is computed, the corresponding bandwidth being selected by means of the cross-validation procedure.

estimator allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

kernel allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.
Details

The expression for the estimator of $\beta$ can be seen in page 52 in Aneiros-Perez et al. (2004).

Value

A list containing:

- **BETA** $p \times \text{length}(\text{b.seq})$ matrix containing the estimate of $\beta$ for each bandwidth in $\text{h.seq}$.
- **G** $n \times p \times \text{length}(\text{b.seq})$ array containing the nonparametric estimate of $E(X_{(ij)} | t_i)$ ($i = 1, ..., n; j = 1, ..., p$) for each bandwidth in $\text{b.seq}$.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are: `plrm.est`, `plrm.gcv`, `plrm.cv`.

Examples

```r
# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

b.h <- plrm.gcv(data)$bh.opt
ajuste <- plrm.beta(data=data, b=b.h[1])
ajuste$BETA

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)
```
m <- function(t) (0.25*t*(1-t))
f <- m(t)

x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x%*%beta
epsilon <- rnorm(n, 0, 0.01)
y <- sum + f + epsilon
data_ind <- matrix(c(y,x), nrow=n)

# We estimate the parametric component of the PLR model
# (GCV bandwidth)
a <- plrm.beta(data_ind)
a$BETA

## Example 2b: dependent data

set.seed(1234)
# We generate the data
x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x%*%beta
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7, sd = 0.01, n = n)
y <- sum + f + epsilon
data_dep <- matrix(c(y,x), nrow=n)

# We estimate the parametric component of the PLR model
# (CV bandwidth)
b <- plrm.cv(data_dep, ln=0.2)$bh.opt[2,1]
a <- plrm.beta(data_dep, b=b)
a$BETA

---

**plrm.ci**

*Confidence intervals estimation in partial linear regression models*

**Description**

This routine obtains a confidence interval for the value $a^T \beta$, by asymptotic distribution and bootstrap, $(Y_i, X_{i1}, ..., X_{ip}, t_i) : i = 1, ..., n$, where:

$$a = (a_1, ..., a_p)^T$$

is an unknown vector,

$$\beta = (\beta_1, ..., \beta_p)^T$$

is an unknown vector parameter and

$$Y_i = X_{i1} \beta_1 + ... + X_{ip} \beta_p + m(t_i) + \epsilon_i.$$  

The nonparametric component, $m$, is a smooth but unknown function, and the random errors, $\epsilon_i$, are allowed to be time series.
Usage

plrm.ci(data=data, seed=123, CI="AD", B=1000, N=50, a=NULL, b1=NULL, b2=NULL, estimator="NW", kernel="quadratic", p.arima=NULL, q.arima=NULL, p.max=3, q.max=3, alpha=0.05, alpha2=0.05, num.lb=10, ic="BIC", Var.Cov.eps=NULL)

Arguments

data       data[, 1] contains the values of the response variable, Y;
data[, 2:(p+1)] contains the values of the "linear" explanatory variables, X_1, ..., X_p;
data[, p+2] contains the values of the "nonparametric" explanatory variable, t.

seed       the considered seed.

CI         method to obtain the confidence interval. It allows us to choose between: “AD” (asymptotic distribution), “B” (bootstrap) or “all” (both). The default is “AD”.

B          number of bootstrap replications. The default is 1000.

N          Truncation parameter used in the finite approximation of the MA(infinite) expression of ε.

a          Vector which, multiplied by beta, is used for obtaining the confidence interval of this result.

b1         the considered bandwidth to estimate the confidence interval by asymptotic distribution. If NULL (the default), it is obtained using cross-validation.

b2         the considered bandwidth to estimate the confidence interval by bootstrap. If NULL (the default), it is obtained using cross-validation.

estimator  allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

kernel     allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

p.arima    the considered p to fit the model ARMA(p,q).

q.arima    the considered q to fit the model ARMA(p,q).

p.max      if Var.Cov.eps=NULL, the ARMA models are selected between the models ARMA(p,q) with 0<=p<=p.max and 0<=q<=q.max. The default is 3.

q.max      if Var.Cov.eps=NULL, the ARMA models are selected between the models ARMA(p,q) with 0<=p<=p.max and 0<=q<=q.max. The default is 3.

alpha      1 - alpha is the confidence level of the confidence interval. The default is 0.05.

alpha2     significance level used to check (if needed) the ARMA model fitted to the residuals. The default is 0.05.

num.lb     if Var.Cov.eps=NULL, it checks the suitability of the selected ARMA model according to the Ljung-Box test and the t-test. It uses up to num.lb delays in the Ljung-Box test. The default is 10.
ic

if \texttt{Var.Cov.eps=NULL}, \texttt{ic} contains the information criterion used to suggest the ARMA model. It allows us to choose between: ”AIC”, ”AICC” or ”BIC” (the default).

\texttt{Var.Cov.eps}

$\times \times$ matrix of variances-covariances associated to the random errors of the regression model. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted regression model and, then, it obtains the var-cov matrix of such ARMA model.

\textbf{Value}

A list containing:

\begin{description}
\item[Bootstrap] a dataframe containing \texttt{ci\_inf} and \texttt{ci\_sup}, the confidence intervals using bootstrap; \texttt{p\_opt} and \texttt{q\_opt} (the orders for the ARMA model fitted to the residuals) and \texttt{b1} and \texttt{b2}, the considered bandwidths.
\item[AD] a dataframe containing \texttt{ci\_inf} and \texttt{ci\_sup}, the confidence intervals using the asymptotic distribution; \texttt{p\_opt} and \texttt{q\_opt} (the orders for the ARMA model fitted to the residuals) and \texttt{b1}, the considered bandwidth.
\item[pv.Box.test] \texttt{p-values} of the Ljung-Box test for the model fitted to the residuals.
\item[pv.t.test] \texttt{p-values} of the \texttt{t.test} for the model fitted to the residuals.
\end{description}

\textbf{Author(s)}

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

\textbf{References}


\textbf{See Also}

A related functions is \texttt{par.ci}.

\textbf{Examples}

\begin{verbatim}
# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

b.h <- plrm.gcv(data)$bh.opt
b1 <- b.h[1]
\end{verbatim}
## Description

From a sample \((Y_i, X_{i1}, \ldots, X_{ip}, t_i) : i = 1, \ldots, n\), this routine computes, for each \(l_n\) considered, an optimal pair of bandwidths for estimating the regression function of the model

\[
Y_i = X_{i1} \beta_1 + \ldots + X_{ip} \beta_p + m(t_i) + \epsilon_i,
\]

where

\[
\beta = (\beta_1, \ldots, \beta_p)
\]

is an unknown vector parameter and

\[
m(.)
\]

is a smooth but unknown function. The random errors, \(\epsilon_i\), are allowed to be time series. The optimal pair of bandwidths, \((b_{\text{opt}}, h_{\text{opt}})\), is selected by means of the leave-(\(2l_n + 1\))-out cross-validation procedure. The bandwidth \(b_{\text{opt}}\) is used in the estimate of \(\beta\), while the pair of bandwidths \((b_{\text{opt}}, h_{\text{opt}})\) is considered in the estimate of \(m\). Kernel smoothing, combined with ordinary least squares estimation, is used.
Usage

plrm.cv(data = data, b.equal.h = TRUE, b.seq=NULL, h.seq=NULL, 
num.b = NULL, num.h = NULL, w = NULL, num.ln = 1, ln.0 = 0, 
step.ln = 2, estimator = "NW", kernel = "quadratic")

Arguments

data data[,1] contains the values of the response variable, Y;
data[, 2:(p+1)] contains the values of the "linear" explanatory variables, 
\[X_1, ..., X_p;\]
data[, p+2] contains the values of the "nonparametric" explanatory variable, 
\[t.\]

b.equal.h if TRUE (the default), the same bandwidth is used for estimating both \(\beta\) and \(m\).
b.seq sequence of considered bandwidths, \(b\), in the CV function for estimating \(\beta\). If 
NULL (the default), num.b equidistant values between zero and a quarter of the 
range of \(t_i\) are considered.

h.seq sequence of considered bandwidths, \(h\), in the pair of bandwidths \((b, h)\) used 
in the CV function for estimating \(m\). If NULL (the default), num.h equidistant 
values between zero and a quarter of the range of \(t_i\) are considered.

num.b number of values used to build the sequence of considered bandwidths for esti-
mating \(\beta\). If b.seq is not NULL, num.b=length(b.seq). Otherwise, if both 
umb and num.h are NULL (the default), num.b=50 is considered; if num.b is 
NULL (the default) but num.h is not NULL, then num.b=num.h is considered; if 
b.equal.h=TRUE (the default) and both num.b and num.h are not NULL and dif-
ferent, the maximum value of num.b and num.h is considered for both.

num.h pairs of bandwidths \((b, h)\) are used for estimating \(m\), num.h being the number 
of values considered for \(h\). If h.seq is not NULL, num.h=length(h.seq). Otherwise, if both 
umb and num.h are NULL (the default), num.h=50 is considered; if num.h is NULL (the default) but num.b is not NULL, num.h=num.b is considered; if 
b.equal.h=TRUE (the default) and both num.b and num.h are not NULL and dif-
ferent, the maximum value of num.b and num.h is considered for both.

w support interval of the weight function in the CV function. If NULL (the default), 
\((q_{0.1}, q_{0.9})\) is considered, where \(q_p\) denotes the quantile of order \(p\) of \(t_i\).

num.ln number of values for \(l_n\): after estimating \(\beta\), \(2l_n + 1\) observations around each 
point \(t_i\) are eliminated to estimate \(m(t_i)\) in the CV function. The default is 1.

ln.0 minimum value for \(l_n\). The default is 0.

step.ln distance between two consecutives values of \(l_n\). The default is 2.

estimator allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear 
Polynomial). The default is “NW”.

kernel allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), 
“triweight” or “uniform” kernel. The default is “quadratic”.
Details

A weight function (specifically, the indicator function $1_{[w1, w2]}$) is introduced in the CV function to allow elimination (or at least significant reduction) of boundary effects from the estimate of $m(t_i)$.

As noted in the definition of num.ln, the estimate of $\beta$ in the CV function is obtained from all data while, once $\beta$ is estimated, $2l_n + 1$ observations around each $t_i$ are eliminated to estimate $m(t_i)$ in the CV function. Actually, the estimate of $\beta$ to be used in time $i$ in the CV function could be done eliminating such $2l_n + 1$ observations too; that possibility was not implemented because both their computational cost and the known fact that the estimate of $\beta$ is quite insensitive to the bandwidth selection.

The implemented procedure generalizes that one in expression (8) in Aneiros-Perez and Quintela-del-Rio (2001) by including a weight function (see above) and allowing two smoothing parameters instead of only one (see Aneiros-Perez et al., 2004).

Value

- **bh.opt**: dataframe containing, for each ln considered, the selected value for (b, h).
- **CV.opt**: CV.opt[k] is the minimum value of the CV function when de k-th value of ln is considered.
- **CV**: an array containing the values of the CV function for each pair of bandwidths and ln considered.
- **b.seq**: sequence of considered bandwidths, b, in the CV function for estimating $\beta$.
- **h.seq**: sequence of considered bandwidths, h, in the pair of bandwidths (b, h) used in the CV function for estimating $m$.
- **w**: support interval of the weight function in the CV function.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are: *plrm.beta, plrm.est, plrm.gcv, np.est, np.gcv* and *np.cv*. 
Examples

# EXAMPLE 1: REAL DATA
data(barnacles)
data <- as.matrix(barnacles)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

aux <- plrm.cv(data, step.ln=1, num.ln=2)
aux$h.opt
plot.ts(aux$CV[, -2])

par(mfrow=c(2, 1))
plot(aux$b.seq, aux$CV[, -2, 1], xlab="h", ylab="CV", type="l", main="ln=0")
plot(aux$b.seq, aux$CV[, -2, 2], xlab="h", ylab="CV", type="l", main="ln=1")

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n) - 0.5)/n
beta <- c(0.05, 0.01)
m <- function(t) (0.25*t*(1-t))
f <- m(t)
x <- matrix(rnorm(200, 0, 1), nrow=n)
sum <- x*beta
epsilon <- rnorm(n, 0, 0.01)
y <- sum + f + epsilon
data_ind <- matrix(c(y, x, t), nrow=100)

# We apply the function
a <- plrm.cv(data_ind)
a$CV.opt

CV <- a$CV
h <- a$h.seq
plot(h, CV, type="l")

## Example 2b: dependent data and ln.0 > 0

set.seed(1234)
# We generate the data
x <- matrix(rnorm(200, 0, 1), nrow=n)
sum <- x*beta
epsilon <- arima.sim(list(order = c(1, 0, 0), ar=0.7), sd = 0.01, n = n)
y <- sum + f + epsilon
data_dep <- matrix(c(y, x, t), nrow=100)
plrm.est

Semiparametric estimates for the unknown components of the regression function in PLR models

Description

This routine computes estimates for $\beta$ and $m(newt_j)$ ($j = 1, ..., J$) from a sample $(Y_i, X_{i1}, ..., X_{ip}, t_i)$: $i = 1, ..., n$, where:

$$\beta = (\beta_1, ..., \beta_p)$$

is an unknown vector parameter,

$$m(.)$$

is a smooth but unknown function and

$$Y_i = X_{i1} \ast \beta_1 + ... + X_{ip} \ast \beta_p \ast m(t_i) + \epsilon_i.$$ 

The random errors, $\epsilon_i$, are allowed to be time series. Kernel smoothing, combined with ordinary least squares estimation, is used.

Usage

plrm.est(data = data, b = NULL, h = NULL, newt = NULL, estimator = "NW", kernel = "quadratic")

Arguments

data

- data[, 1] contains the values of the response variable, $Y$;
- data[, 2:(p+1)] contains the values of the "linear" explanatory variables, $X_1, ..., X_p$;
- data[, p+2] contains the values of the "nonparametric" explanatory variable, $t$.

b

- bandwidth for estimating the parametric part of the model. If both $b$ and $h$ are NULL (the default), it is selected by means of the cross-validation procedure (fixing $b$=h); if $b$ is NULL (the default) but $h$ is not NULL, $b$=h is considered.

h

- (b,h) is the pair of bandwidths for estimating the nonparametric part of the model. If both $b$ and $h$ are NULL (the default), it is selected by means of the cross-validation procedure (fixing $b$=h); if $b$ is NULL (the default) but $h$ is not NULL, $b$=h is considered; if $h$ is NULL (the default) but $b$ is not NULL, $h$=b is considered.
values of the "nonparametric" explanatory variable where the estimator of \(m\) is evaluated. If NULL (the default), the considered values will be the values of \(\text{data}[,\, p+2]\).

estimator allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

kernel allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

Details

Expressions for the estimators of \(\beta\) and \(m\) can be seen in page 52 in Aneiros-Perez et al. (2004).

Value

A list containing:

- beta: a vector containing the estimate of \(\beta\).
- m.t: a vector containing the estimator of the non-parametric part, \(m\), evaluated in the design points.
- m.newt: a vector containing the estimator of the non-parametric part, \(m\), evaluated in newt.
- residuals: a vector containing the residuals: \(Y - X \ast \text{beta} - m.t\).
- fitted.values: the values obtained from the expression: \(X \ast \text{beta} + m.t\)
- b: the considered bandwidth for estimating \(\beta\).
- h: \((b, h)\) is the pair of bandwidths considered for estimating \(m\).

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are: \texttt{plrm.beta}, \texttt{plrm.gcv}, \texttt{plrm.cv}, \texttt{np.est}, \texttt{np.gcv} and \texttt{np.cv}.
Examples

# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

b.h <- plrm.gcv(data)$bh.opt
ajuste <- plrm.est(data=data, b=b.h[1], h=b.h[2])
ajuste$beta
plot(data[,4], ajuste$m, type="l", xlab="t", ylab="m(t)"

plot(data[,1], ajuste$fitted.values, xlab="y", ylab="y.hat", main="y.hat vs y")
abline(0,1)
mean(ajuste$residuals^2)/var(data[,1])

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)
m <- function(t) (0.25*t*(1-t))
f <- m(t)

x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x%*%beta
epsilon <- rnorm(n, 0, 0.01)
y <- sum + f + epsilon
data_ind <- matrix(c(y,x,t),nrow=100)

# We estimate the components of the plr model
# (CV bandwidth)
a <- plrm.est(data_ind)
a$beta

est <- a$m.t
plot(t, est, type="l", lty=2, ylab="")
points(t, 0.25*t*(1-t), type="l")
legend(x="topleft", legend = c("m", "m hat"), col=c("black", "black"), lty=c(1,2))

## Example 2b: dependent data
# We generate the data
x <- matrix(rnorm(200,0,1), nrow=n)
sum <- x%*%beta
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y <- sum + f + epsilon
data_dep <- matrix(c(y,x,t),nrow=100)

# We estimate the components of the PLR model
# (CV bandwidth)
h <- plrm.cv(data_dep, ln.0=2)$bh.opt[3,1]
a <- plrm.est(data_dep, h=h)

a$beta

est <- a$m.t
plot(t, est, type="l", lty=2, ylab="")
points(t, 0.25*t*(1-t), type="l")
legend(x="topleft", legend = c("m", "m hat"), col=c("black", "black"), lty=c(1,2))

---

plrm.gcv

Generalized cross-validation bandwidth selection in PLR models

Description

From a sample \( (Y_i, X_{i1}, ..., X_{ip}, t_i) : i = 1, ..., n \), this routine computes an optimal pair of bandwidths for estimating the regression function of the model

\[
Y_i = X_{i1} \beta_1 + ... + X_{ip} \beta_p + m(t_i) + \epsilon_i,
\]

where

\[
\beta = (\beta_1, ..., \beta_p)
\]

is an unknown vector parameter and

\[
m(\cdot)
\]

is a smooth but unknown function. The optimal pair of bandwidths, \((b_{opt}, h_{opt})\), is selected by means of the generalized cross-validation procedure. The bandwidth \(b_{opt}\) is used in the estimate of \(\beta\), while the pair of bandwidths \((b_{opt}, h_{opt})\) is considered in the estimate of \(m\). Kernel smoothing, combined with ordinary least squares estimation, is used.

Usage

\[
plrm.gcv(data = data, b.equal.h = TRUE, b.seq=NULL, h.seq=NULL, num.b = NULL, num.h = NULL, estimator = "NW", kernel = "quadratic")
\]

Arguments

data

- \(data[,1]\) contains the values of the response variable, \(Y\);
- \(data[, 2:(p+1)]\) contains the values of the "linear" explanatory variables, \(X_1, ..., X_p\);
- \(data[, p+2]\) contains the values of the "nonparametric" explanatory variable, \(t\).
b. equal.h if TRUE (the default), the same bandwidth is used for estimating both \( \beta \) and \( m \).

b. seq sequence of considered bandwidths, \( b \), in the GCV function for estimating \( \beta \).

h. seq sequence of considered bandwidths, \( h \), in the pair of bandwidths \((b, h)\) used in the GCV function for estimating \( m \).

num.b number of values used to build the sequence of considered bandwidths for estimating \( \beta \). If \( b \).seq is not NULL, \( num.b=\text{length}(b \text{.seq}) \). Otherwise, if both \( num.b \) and \( num.h \) are NULL (the default), \( num.b=50 \) is considered; if \( num.b \) is NULL (the default) but \( num.h \) is not NULL, then \( num.b=num.h \) is considered; if \( b \).equal.h=TRUE (the default) and both \( num.b \) and \( num.h \) are not NULL and different, the maximum value of \( num.b \) and \( num.h \) is considered for both.

num.h pairs of bandwidths \((b, h)\) are used for estimating \( m \), \( num.h \) being the number of values considered for \( h \). If \( h \).seq is not NULL, \( num.h=\text{length}(h \text{.seq}) \). Otherwise, if both \( num.b \) and \( num.h \) are NULL (the default), \( num.h=50 \) is considered; if \( num.h \) is NULL (the default) but \( num.b \) is not NULL, \( num.h=num.b \) is considered; if \( b \).equal.h=TRUE (the default) and both \( num.b \) and \( num.h \) are not NULL and different, the maximum value of \( num.b \) and \( num.h \) is considered for both.

estimator allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

kernel allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

Details

The implemented procedure generalizes that one in page 423 in Speckman (1988) by allowing two smoothing parameters instead of only one (see Aneiros-Perez et al., 2004).

Value

bh. opt selected value for \((b, h)\).

GCV. opt minimum value of the GCV function.

GCV matrix containing the values of the GCV function for each pair of bandwidths considered.

b. seq sequence of considered bandwidths, \( b \), in the GCV function for estimating \( \beta \). If \( b \).seq was not input by the user, it is composed by \( num.b \) equidistant values between zero and a quarter of the range of \( t_i \).

h. seq sequence of considered bandwidths, \( h \), in the pair of bandwidths \((b, h)\) used in the GCV function for estimating \( m \). If \( h \).seq was not input by the user, it is composed by \( num.h \) equidistant values between zero and a quarter of the range of \( t_i \).

Author(s)

German Aneiros Perez <ganeiros@udc.es>

Ana Lopez Cheda <ana.lopez.cheda@udc.es>
References


See Also

Other related functions are: `plrm.beta`, `plrm.est`, `plrm.cv`, `np.est`, `np.gcv` and `np.cv`.

Examples

```r
# EXAMPLE 1: REAL DATA

data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

aux <- plrm.gcv(data)
aux$h.opt
plot(aux$h.seq, aux$GCV, xlab="h", ylab="GCV", type="l")

# EXAMPLE 2: SIMULATED DATA
## Example 2a: independent data

set.seed(1234)

# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)
m <- function(t) (0.25*t*(1-t))
f <- m(t)

x <- matrix(rnorm(200, 0, 1), nrow=n)
sum <- x%*%beta
epsilon <- rnorm(n, 0, 0.01)
y <- sum + f + epsilon
data_ind <- matrix(c(y, x, t), nrow=100)

# We obtain the optimal bandwidths
a <- plrm.gcv(data_ind)
a$GCV.opt

GCV <- a$GCV
h <- a$h.seq
plot(h, GCV, type="l")
```
**Description**

From a sample \((Y_i, X_{i1}, ..., X_{ip}, t_i) : i = 1, ..., n\), this routine tests the null hypotheses \(H_0 : \beta = \beta_0\) and \(H_0 : m = m_0\), where:

\[
\beta = (\beta_1, ..., \beta_p)
\]

is an unknown vector parameter,

\[
m(.)
\]

is a smooth but unknown function and

\[
Y_i = X_{i1} \* \beta_1 + ... + X_{ip} \* \beta_p + m(t_i) + \epsilon_i.
\]

Fixed equally spaced design is considered for the "nonparametric" explanatory variable, \(t\), and the random errors, \(\epsilon_i\), are allowed to be time series. The test statistic used for testing \(H_0 : \beta = \beta_0\) derives from the asymptotic normality of an estimator of \(\beta\) based on both ordinary least squares and kernel smoothing (this result giving a \(\chi^2\)-test). The test statistic used for testing \(H_0 : m = m_0\) derives from a Cramer-von-Mises-type functional distance between a nonparametric estimator of \(m\) and \(m_0\).

**Usage**

```r
plrm.gof(data = data, beta0 = NULL, m0 = NULL, b.seq = NULL, h.seq = NULL, w = NULL, estimator = "NW", kernel = "quadratic", time.series = FALSE, Var.Cov.eps = NULL, Tau.eps = NULL, b0 = NULL, h0 = NULL, lag.max = 50, p.max = 3, q.max = 3, ic = "BIC", num.lb = 10, alpha = 0.05)
```

**Arguments**

- `data`: \([, 1]\) contains the values of the response variable, \(Y\);
- `data[, 2:(p+1)]` contains the values of the "linear" explanatory variables \(X_{1}, ..., X_{p}\);
- `data[, p+2]` contains the values of the "nonparametric" explanatory variable, \(t\).
- `beta0`: the considered parameter vector in the parametric null hypothesis. If NULL (the default), the zero vector is considered.
- `m0`: the considered function in the nonparametric null hypothesis. If NULL (the default), the zero function is considered.
- `b.seq`: the statistic test for \(H_0 : \beta = \beta_0\) is performed using each bandwidth in the vector `b.seq`. If NULL (the default) but `h.seq` is not NULL, it takes `b.seq=h.seq`. If both `b.seq` and `h.seq` are NULL, 10 equidistant values between zero and a quarter of the range of \(t_i\) are considered.
the statistic test for $H_0 : m = m_0$ is performed using each pair of bandwidths $(b_{seq}[j], h_{seq}[j])$. If NULL (the default) but $b_{seq}$ is not NULL, it takes $h_{seq} = b_{seq}$. If both $b_{seq}$ and $h_{seq}$ are NULL, 10 equidistant values between zero and a quarter of the range of $t_i$ are considered for both $b_{seq}$ and $h_{seq}$.

If NULL (the default), $(q_{0.1}, q_{0.9})$ is considered, where $q_p$ denotes the quantile of order $p$ of $t_i$.

allows us the choice between “NW” (Nadaraya-Watson) or “LLP” (Local Linear Polynomial). The default is “NW”.

allows us the choice between “gaussian”, “quadratic” (Epanechnikov kernel), “triweight” or “uniform” kernel. The default is “quadratic”.

it denotes whether the data are independent (FALSE) or if data is a time series (TRUE). The default is FALSE.

$n \times n$ matrix of variances-covariances associated to the random errors of the regression model. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted regression model and, then, it obtains the var-cov matrix of such ARMA model.

it contains the sum of autocovariances associated to the random errors of the regression model. If NULL (the default), the function tries to estimate it: it fits an ARMA model (selected according to an information criterium) to the residuals from the fitted regression model and, then, it obtains the sum of the autocovariances of such ARMA model.

if Var.Cov.eps=NULL and/or Tau.eps=NULL, b0 contains the pilot bandwidth for the estimator of $\beta$ used for obtaining the residuals to construct the default for Var.Cov.eps and/or Tau.eps. If NULL (the default) but b0 is not NULL, it takes $b_0 = h_0$. If both b0 and h0 are NULL, a quarter of the range of $t_i$ is considered.

if Var.Cov.eps=NULL and/or Tau.eps=NULL, (b0, h0) contains the pair of pilot bandwidths for the estimator of $m$ used for obtaining the residuals to construct the default for Var.Cov.eps and/or Tau.eps. If NULL (the default) but b0 is not NULL, it takes $h_0 = b_0$. If both b0 and h0 are NULL, a quarter of the range of $t_i$ is considered for both b0 and h0.

if Tau.eps=NULL, lag.max contains the maximum delay used to construct the default for Tau.eps. The default is 50.

if Var.Cov.eps=NULL and/or Tau.eps=NULL, the ARMA model is selected between the models ARMA(p,q) with $0 \leq p \leq p_{max}$ and $0 \leq q \leq q_{max}$. The default is 3.

if Var.Cov.eps=NULL and/or Tau.eps=NULL, the ARMA model is selected between the models ARMA(p,q) with $0 \leq p \leq p_{max}$ and $0 \leq q \leq q_{max}$. The default is 3.

if Var.Cov.eps=NULL and/or Tau.eps=NULL, ic contains the information criterion used to suggest the ARMA model. It allows us to choose between: "AIC", "AICC" or "BIC" (the default).

if Var.Cov.eps=NULL and/or Tau.eps=NULL, it checks the suitability of the selected ARMA model according to the Ljung-Box test and the t-test. It uses up to num.lb delays in the Ljung-Box test. The default is 10.
alpha if Var.Cov.eps=NULL and/or Tau.eps=NULL, alpha contains the significance level which the ARMA model is checked. The default is 0.05.

Details

A weight function (specifically, the indicator function \(1_{[v_1,v_2]}\)) is introduced in the test statistic for testing \(H_0 : m = m_0\) to allow elimination (or at least significant reduction) of boundary effects from the estimate of \(m(t_i)\).

If Var.Cov.eps=NULL and the routine is not able to suggest an approximation for Var.Cov.eps, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct Var.Cov.eps, the procedure suggested in Aneiros-Perez and Vieu (2013) can be followed.

If Tau.eps=NULL and the routine is not able to suggest an approximation for Tau.eps, it warns the user with a message saying that the model could be not appropriate and then it shows the results. In order to construct Tau.eps, the procedures suggested in Aneiros-Perez (2008) can be followed.

The implemented procedures generalize those ones in expressions (9) and (10) in Gonzalez-Manteiga and Aneiros-Perez (2003) by allowing some dependence condition in \((X_{i1},...,X_{ip}) : i = 1,...,n\) and including a weight function (see above), respectively.

Value

A list with two dataframes:

- **parametric.test**: a dataframe containing the bandwidths, the statistics and the p-values when one tests \(H_0 : \beta = \beta_0\)

- **nonparametric.test**: a dataframe containing the bandwidths b and h, the statistics, the normalised statistics and the p-values when one tests \(H_0 : m = m_0\)

Moreover, if data is a time series and Tau.eps or Var.Cov.eps are not specified:

- **pv.Box.test**: p-values of the Ljung-Box test for the model fitted to the residuals.
- **pv.t.test**: p-values of the t.test for the model fitted to the residuals.
- **ar.ma**: ARMA orders for the model fitted to the residuals.

Author(s)

German Aneiros Perez <ganeiros@udc.es>
Ana Lopez Cheda <ana.lopez.cheda@udc.es>

References


See Also

Other related functions are `plrm.est.par.gof` and `np.gof`.

Examples

```r
# EXAMPLE 1: REAL DATA
data(barnacles1)
data <- as.matrix(barnacles1)
data <- diff(data, 12)
data <- cbind(data, 1:nrow(data))

plrm.gof(data)
plrm.gof(data, beta0=c(-0.1, 0.35))
```

```r
# EXAMPLE 2: SIMULATED DATA
## Example 2a: dependent data

set.seed(1234)
# We generate the data
n <- 100
t <- ((1:n)-0.5)/n
beta <- c(0.05, 0.01)
m <- function(t) {0.25*t*(1-t)}
f <- m(t)
f.function <- function(u) {0.25*u*(1-u)}

x <- matrix(rnorm(200, 0, 1), nrow=n)
sum <- x%*%beta
epsilon <- arima.sim(list(order = c(1,0,0), ar=0.7), sd = 0.01, n = n)
y <- sum + f + epsilon
data <- cbind(y, x, t)

## Example 2a.1: true null hypotheses
plrm.gof(data, beta0=c(0.05, 0.01), m0=f.function, time.series=TRUE)

## Example 2a.2: false null hypotheses
plrm.gof(data, time.series=TRUE)
```
Index

*Topic **Nonparametric Statistics**
  np.ancova, 4
  np.cv, 7
  np.est, 10
  np.gcv, 12
  np.gof, 14
  plrm.ancova, 27
  plrm.beta, 32
  plrm.cv, 37
  plrm.est, 41
  plrm.gcv, 44
  plrm.gof, 47

*Topic **PLRModels**
  PLRModels-package, 2

*Topic **Regression**
  np.ancova, 4
  np.cv, 7
  np.est, 10
  np.gcv, 12
  np.gof, 14
  par.ancova, 17
  par.ci, 20
  par.est, 22
  par.gof, 25
  plrm.ancova, 27
  plrm.beta, 32
  plrm.ci, 34
  plrm.cv, 37
  plrm.est, 41
  plrm.gcv, 44
  plrm.gof, 47

*Topic **Statistical Inference**
  np.ancova, 4
  np.cv, 7
  np.est, 10
  np.gcv, 12
  np.gof, 14
  par.ancova, 17
  par.ci, 20
  par.est, 22
  par.gof, 25
  plrm.ancova, 27
  plrm.beta, 32
  plrm.ci, 34
  plrm.cv, 37
  plrm.est, 41
  plrm.gcv, 44
  plrm.gof, 47

*Topic **Time Series**
  np.ancova, 4
  np.cv, 7
  np.est, 10
  np.gcv, 12
  np.gof, 14
  par.ancova, 17
  par.ci, 20
  par.est, 22
  par.gof, 25
  plrm.ancova, 27
  plrm.beta, 32
  plrm.ci, 34
  plrm.cv, 37
  plrm.est, 41
  plrm.gcv, 44
  plrm.gof, 47

*Topic **datasets**
  barnacles1, 2
  barnacles2, 3

barnacles1, 2
barnacles2, 3

np.ancova, 4, 18, 30
np.cv, 7, 11, 13, 39, 42, 46
np.est, 6, 9, 10, 13, 16, 39, 42, 46
np.gcv, 9, 11, 12, 39, 42, 46
np.gof, 14, 26, 50

par.ancova, 6, 17, 30
par.ci, 20, 36
par.est, 22
par.gof, 16, 25, 50
plrm.ancova, 6, 18, 27
plrm.beta, 23, 32, 39, 42, 46
plrm.ci, 22, 34
plrm.cv, 9, 11, 13, 33, 37, 42, 46
plrm.est, 9, 11, 13, 23, 30, 33, 39, 41, 46, 50
plrm.gcv, 9, 11, 13, 33, 39, 42, 44
plrm.gof, 16, 26, 47
PLRModels (PLRModels-package), 2
PLRModels-package, 2