Package ‘lm.br’

January 27, 2015

Type Package
Title Linear Model with Breakpoint
Version 2.6
Date 2014-11-13
Copyright 'lm.br' uses the design and some R-code of 'lm' copyright (C) 2013 The R Foundation for Statistical Computing, and of 'lm.gls' copyright (C) 1994-2005 W. N. Venables and B. D. Ripley.

Description Exact significance tests for a changepoint in linear or multivariate linear regression. Confidence regions with exact coverage probabilities for the changepoint.

License GPL (>= 2)
Depends R(>= 3.0.1), Rcpp (>= 0.11.0)
Imports methods
LinkingTo Rcpp

Author Marc Adams [aut, cre], authors of R function 'lm' [ctb] (general interface), authors of 'lm.gls' [ctb] (interface and R code for covariate weights), U.S. NIST [ctb] (C++ code for TNT::Vector template)

Maintainer Marc Adams <lm.br.pkg@gmail.com>

Repository CRAN
Repository/R-Forge/Project blmr
Repository/R-Forge/Revision 49
Repository/R-Forge/DateTimeStamp 2014-11-13 14:58:20
Date/Publication 2014-11-14 08:16:36

NeedsCompilation yes

R topics documented:

'lm.br' .......................................................... 2
ci .............................................................. 4
Description

Exact significance tests for a changepoint in linear or multivariate linear regression. Confidence intervals and confidence regions with exact coverage probabilities for the changepoint.

Usage

```r
lm.br(formula, type = "LL", data, subset, weights, inverse = FALSE,
      var.known = FALSE, na.action, contrasts, offset, ...)
```

Arguments

- `formula`: a formula expression as for regression models, of the form response ~ predictors, see `formula`.
- `type`: "LL", "LT" or "TL" which stand for line-line, line-threshold or threshold-line, defined below.
- `data`: an optional data-frame that assigns values in `formula`.
- `subset`: expression saying which subset of the data to use.
- `weights`: vector or positive-definite matrix.
- `inverse`: if TRUE then `weights` specifies the inverse of the weights vector or matrix, as for a covariance matrix.
- `var.known`: is the variance known?
- `na.action`: a function to filter missing data.
- `contrasts`: an optional list, see 'contrasts.arg' in `model.matrix.default`.
- `offset`: a constant vector to be subtracted from the responses vector.
- `...`: other arguments to `lm.fit` or `lm.wfit`.

Details

A broken-line model consists of two straight lines joined at a changepoint. Three versions are

- **LL**  \( y = \alpha + B \times \min(x - \theta, 0) + Bp \times \max(x - \theta, 0) + e \)
- **LT**  \( y = \alpha + B \times \min(x - \theta, 0) + e \)
- **TL**  \( y = \alpha + Bp \times \max(x - \theta, 0) + e \)
where \( e \sim \text{Normal}(0, \text{var} \cdot \text{inv(weights)}) \). The LT and TL models omit \( \alpha \) if the formula is without intercept, such as \( 'y~x+0' \). The same models apply for a multivariate formula such as \( 'y~x_1 + x_2 + \ldots + x_n' \) where \( \alpha \) becomes the coefficient of the \( '1'-\text{vector} \) and \( \theta \) the changepoint for the coefficient of the first predictor term, \( 'x_1' \).

Parameters \( \theta \), \( \alpha \), \( B \), \( B_p \), \( \text{var} \) are unknown, but \( \text{weights} \) is known. Exact inferences about the changepoint \( \theta \) or \( (\theta, \alpha) \) are based on the distribution of its likelihood-ratio statistic, conditional on sufficient statistics for the other parameters.

Test for the presence of a changepoint by the significance level of a postulate value outside the range of \( x \)-values. Thus, in the LL model \( '\text{sl}(\min(x_1) - 1)' \) would give the exact significance level of the null hypothesis "single line" versus the alternate hypothesis "broken line."

**Value**

\'lm.br\' returns a list that includes a C++ object with accessor functions. Functions \( \text{sl}, \text{ci} \) and \( \text{cr} \) get significance levels, confidence intervals, and confidence regions for the changepoint’s \( x \)-coordinate or \((x,y)\)-coordinates. Other functions are \( \text{mle} \) to get maximum likelihood estimates and \( \text{sety} \) to set new \( y \)-values. The returned object also lists ‘coefficients’, ‘fitted.values’ and ‘residuals’, the same as for an ‘lm’ output list.

**Note**

If variance is known, then \( \text{var} = 1 \) and ‘weights’ is the inverse of the variances vector or variance-covariance matrix.

**References**


**See Also**

vignette( "lm.br" )
demo( testscript )

**Examples**

```r
# Smith & Cook (1980), "Straight Lines with a Change-point: A Bayesian
# reciprocal of blood creatinine L/micromol vs day after transplant.
creatinine <- c(37.3, 47.1, 51.5, 67.6, 75.9, 73.3, 69.4, 61.5, 31.8, 19.4)
day <- 1:10
sc <- lm.br(creatinine ~ day)
s$ mle()
s$ ci()
s$ cr(.90, 'af')
s$ sl(day[1] - 1.5)  # test for the presence of a changepoint
```
# A 'TL' example, data from figure 1 in Chiu et al. (2006), "Bent-cable regression theory and applications", J Am Stat Assoc, 101*, 542-553,
# log(salmon abundance) vs year.
salmon <- c( 2.50, 2.93, 2.94, 2.83, 2.43, 2.84, 3.06, 2.97, 2.94, 2.65, 2.92, 2.71, 2.93, 2.68, 2.12, 2.08, 1.81, 2.45, 1.71, 0.55, 1.30 )
year <- 1980 : 2000
chiu <- lm.br( salmon ~ year, 'tl' )
chii$ci()

# A multivariate example, using one of R's included datasets,
# automobile miles-per-gallon versus weight and horsepower.
lm.br( mpg ~ wt + hp, data = mtcars )

# An example with variance known, in the Normal approximations of binomial
# random variables using formula 2.28 of Cox and Snell (1989).
# Ex. 3.4 of Freeman (2010) "Inference for binomial changepoint data" in
# Advances in Data Analysis, ed. C Skiadas, Boston: Birkhauser, 345-352.
trials <- c( 15, 82, 82, 77, 38, 81, 12, 97, 33, 75, 85, 37, 44, 96, 76, 26, 91, 47, 41, 35 )
successes <- c( 8, 44, 47, 39, 24, 38, 3, 51, 16, 43, 47, 27, 33, 64, 41, 18, 61, 32, 33, 24 )
log_odds <- log( (successes - 0.5)/(trials - successes - 0.5) )
variances <- (trials-1)/( successestrials-successes )
group <- 1 : 20
lm.br( log_odds ~ group, 'TL', w= variances, inv= TRUE, var.known= TRUE )

# An example that shows different confidence regions from inference by
# conditional likelihood-ratio versus approximate-F.
y <- c( 1.55, 3.2, 6.3, 4.8, 4.3, 4.0, 3.5, 1.8 )
x <- 1:8
eg <- lm.br( y ~ x )
eg$cr()
eg$cr( method = 'aF' )

---

## Confidence Interval for the Changepoint

### Description

Confidence interval for 'theta', the changepoint's x-coordinate.

### Usage

```r
## S4 method for signature 'Cpp_Clmbr'
ci( CL = 0.95, method = "CLR" )
```
Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>confidence level, between 0 and 1.</td>
</tr>
<tr>
<td>method</td>
<td>&quot;CLR&quot; or &quot;AF&quot; which stand for conditional likelihood-ratio or approximate-F, see s1 for details.</td>
</tr>
</tbody>
</table>

Details

This subroutine scans to determine the postulate values of \( \theta \) that have significance level greater than 1-CL.

Examples

```r
# Data for Patient B from Smith and Cook (1980)
y <- c(37.3, 47.1, 51.5, 67.6, 75.9, 73.3, 69.4, 61.5, 31.8, 19.4)
x <- 1:10
sc <- lm.br( y ~ x )
sc$ci()
sc $ ci( 0.90 )
sc $ ci( .99, 'af' )
```

Description

Joint confidence region for \((\theta, \alpha)\), the changepoint’s \((x,y)\)-coordinates.

Usage

```r
## S4 method for signature 'Cpp_Clmbr'
CR( CL =0.95 , method ="CLR", incr, output ="G" )
```

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>confidence level, between 0 and 1.</td>
</tr>
<tr>
<td>method</td>
<td>&quot;CLR&quot; or &quot;AF&quot; which stand for conditional likelihood-ratio or approximate-F (rapid), see s1 for details.</td>
</tr>
<tr>
<td>incr</td>
<td>increment of theta values for the confidence region’s boundary-points.</td>
</tr>
<tr>
<td>output</td>
<td>&quot;G&quot;, &quot;T&quot; or &quot;V&quot; which stand for graph, text or value.</td>
</tr>
</tbody>
</table>

Details

This subroutine scans to determine the postulate values of \((\theta, \alpha)\) that have significance level greater than 1-CL. It scans first along the \((\theta, \alpha\text{-MLE})\) ridge to determine the \(\theta\) boundary limits.
Value

If 'output' is "G" or "T" then 'cr' graphs or prints-out the confidence region but does not return a value. If 'output' is "V" then 'cr' returns an N x 3 matrix of boundary points (theta, min-alpha, max-alpha).

Examples

```r
# A quick example
y <- c(2, 0, 2.001, 4, 6)
x <- 1:5
t <- lm.br(y ~ x)
t $ cr()
t$cr(.9, 'af', incr = 0.1, out='t')
```

---

`mle`  
*Maximum Likelihood Estimates*

Description

Maximum-likelihood estimates of parameters. Estimates are without bias correction except for the variance.

Usage

```r
## S4 method for signature 'Cpp_Clmbr'
mle()
```

Examples

```r
# Data for Patient B from Smith and Cook (1980)
y <- c(37.3, 47.1, 51.5, 67.6, 75.9, 73.3, 69.4, 61.5, 31.8, 19.4)
x <- 1:10
sc <- lm.br(y-x)
sc$mle()
```

---

`sety`  
*Set y-Values*

Description

Reset the response values in the C++ object.

Usage

```r
## S4 method for signature 'Cpp_Clmbr'
sety( rWy )
```
Arguments

\texttt{rWy}

vector of 'y' values, pre-multiplied by the square-root of 'weights'.

Details

The 'rWy' vector is simply the y-vector if the model does not specify weights. The square-root of a vector 'W' is the vector 'rW' of the square-roots of the elements of 'W'. The square-root of a matrix 'W' here is the matrix 'rW' such that \( rW \cdot rW = W \) (a stricter definition than \( rW \cdot \text{transpose}(rW) = W \)). 'rW' is the inverse square-root if 'inverse' was TRUE in the 'lm.br' call.

Note

The pre-multiplied vector is more convenient as input during simulation tests. 'sety' changes the y-values for the accessor functions 'sl', 'ci', 'cr' and 'mle'. But the output list of 'coefficients', 'fitted.values' and so on, retains its values from the original call.

Examples

\begin{verbatim}
# A simulation test
x <- c(1.0, 1.1, 1.3, 1.7, 2.4, 3.9, 5.7, 7.6, 8.4, 8.6)
y <- x
LLmodel <- lm.br( y ~ x )
countCLR <- countAF <- 0
theta <- 3
for( i in 1:10000 ) {
  y <- 0 + (-1.)*pmin(x-theta,0) + (0.5)*pmax(x-theta,0) + rnorm(10)
  LLmodel$sety( y )
  stest <- LLmodel$sl( theta, 'clr', .0001, "v" )
  if( stest > .05 ) countCLR <- countCLR + 1
  stest <- LLmodel$sl( theta, 'af', .0001, "v" )
  if( stest > .05 ) countAF <- countAF + 1
  if( floor(i/1000) - i/1000 == 0 ) cat(i, countCLR/i, countAF/i, "n")
}
\end{verbatim}

\textbf{sl} 

\textit{Significance Level for Changepoint}

Description

Significance level of a postulate value for the changepoint's x-coordinate or (x,y)-coordinates.

Usage

\begin{verbatim}
## S4 method for signature 'Cpp_Clmbr'
sl( theta0, method ="CLR", accuracy =0.001, output ="T" )
sl( theta0, alpha0, method ="CLR", accuracy =0.001, output ="T" )
\end{verbatim}
Arguments

theta0  postulate value for 'theta', the changepoint’s x-coordinate.
alpha0  postulate value for 'alpha', the changepoint’s y-coordinate.
method  "CLR", "MC" or "AF" which stand for conditional likelihood-ratio, conditional likelihood-ratio by Monte Carlo or approximate-F, details below.
accuracy maximum absolute error in numerical integration for the "CLR" method, or in Monte Carlo evaluation for the "MC" method, not referenced for the "AF" method.
output  "T", "V" or "B" which stand for text, value or both.

Details

Knowles, Siegmund and Zhang (1991) reduced the conditional likelihood-ratio significance test to a probability expression based on a generic random variable.

The default method "CLR" evaluates this probability using a geometric-expectation formula that Knowles et al. also derived. This formula slightly over-estimates, but the error is negligible for significance levels below 0.20.

Method "MC" evaluates the probability expression directly by Monte Carlo simulation, which avoids the over-estimate of the "CLR" method.

Method "AF" estimates the distribution of the likelihood-ratio statistic by the related F-distribution (or chi-squared if variance known) that would be exact for a linear model. This method is not exact, but it is common for non-linear regression.

Value

'sl' prints-out the result but does not return a value if 'output' is "T". 'sl' returns a numeric value if 'output' is "V" or "B".

Note

The 'accuracy' error-limit does not include the slight over-estimate that is inherent in the "CLR" method, nor the approximation inherent in the "AF" method.

Examples

# Data for Patient B from Smith and Cook (1980)
y <- c(37.3, 47.1, 51.5, 67.6, 75.9, 73.3, 69.4, 61.5, 31.8, 19.4)
x <- 1:10
sc <- lm.br( y ~ x )

sc $ sl( 6.1 )
sc $ sl( 6.1, 'mc' )
sc $ sl( 6.1, 'mc', 0.00001 )
sc $ sl( 6.1, 88.2, 'clr' )
sc $ sl( 6.1, 88.2, 'af' )
tmp <- sc $ sl( 6.1, 88.2, 'mc', 0.001, "B" )
tmp
Index

*Topic package
  'lm.br', 2
  'lm.br', 2

  ci, 3, 4
  ci, Cpp_Clmbr-method (ci), 4
  cr, 3, 5
  cr, Cpp_Clmbr-method (cr), 5

  formula, 2
  lm.br ('lm.br'), 2
  lm.br-package ('lm.br'), 2
  lm.fit, 2
  lm.wfit, 2

  mle, 3, 6
  mle, Cpp_Clmbr-method (mle), 6
  model.matrix.default, 2

  print.lm.br ('lm.br'), 2

  sety, 3, 6
  sety, Cpp_Clmbr-method (sety), 6
  sl, 3, 5, 7
  sl, Cpp_Clmbr-method (sl), 7