Package ‘hcp’

February 20, 2015

Type Package

Title Change Point Estimation for Regression with Heteroscedastic Data

Description Estimation of parameters in 3-segment (i.e. 2 change-point) regression models with heteroscedastic variances is provided based on both likelihood and hybrid Bayesian approaches, with and without continuity constraints at the change points.

Version 0.1

Author Stephen J.Ganocy, Jiayang Sun, and Yulei Wang

Maintainer Yulei Wang <yxw390@case.edu>

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2014-11-05 14:14:28

R topics documented:

`FP.Sample.2` *Change Point Estimation for Regression with Heteroscedastic Data*

Description

This is the footprint data set used in Ganocy and Sun (2014).

Format

A 2-column vector containing 141 observations. The first column lists the values (fp1x) at which the footprint pressure distribution curve are measured, while the second column has the responses (fp1y), the pressures obtained at the x's.

Source

Heteroscedastic Change Point Analysis and Application to Footprint Data, 2014

References

Stephen J. Ganocy and Jiayang Sun (2014), "Heteroscedastic Change Point Analysis and Application to Footprint Data," J. Data Science.

---

hcp — *Change Point Estimation for Regression with Heteroscedastic Data*

Description

Estimation of parameters in 3-segment (i.e. 2 change-point) regression models with heteroscedastic variances is provided based on both likelihood and hybrid Bayesian approaches, with and without continuity constraints at the change points.

Usage

```
hcp(dataset, jlo, jhi, klo, khi, method = c("C-LLL", "C-LQL", "U-LLL", "MMP"),
  variance = c("121", "Common", "Differ"), plot = c("FALSE", "TRUE"),
  sigma21, r1, s1, sigma22, r2, s2, sigma23, r3, s3)
```

Arguments

`dataset`	either a data frame or matrix containing 2 columns, where the first column is the predictor or covariate and the second column is the response.
`jlo,jhi`	lower and upper bounds for the x value of the first change point.
`klo,khi`	lower and upper bounds for the x value of the second change point.
`method`	the method and model used to compute the parameters in a 3-segment (2-change-point) regression model with Gaussian errors. The default is "C-LLL", a constrained linear-linear-linear model, constrained to be continuous at the two change points. The other options are "C-LQL", "U-LLL", and "MMP". "C-LQL" is for a constrained linear-quadratic-linear model, constrained to be continuous at the two change points. "U-LLL" is for an unconstrained LLL model, which does not impose continuity constraints at the two change points. "MMP" is for a hybrid Bayesian estimation, which uses a Maximization-Maximization-Posterior procedure to estimate parameters in the LLL model without a continuity constraint.

`variance`	the default is "121", which means a heteroscedastic data with identical variance, sigma21, for the first and third segments, but (possibly) different variance, sigma22, for the second segment. The option "Common" is for a homoscedastic data with same variance, sigma21, for all three segments. The option "Differ" is for heteroscedastic data with different variances, sigma21, sigma22, and sigma23, for three segments.
`plot`	the default is "FALSE". plot = "TRUE" provides a log-likelihood surface plot. No plot for MMP method.
	* THE FOLLOWING ARE REQUIRED ONLY FOR MMP METHOD
`sigma21`	initial value of variance, sigma1 squared.
`r1,s1`	shape and scale parameters in the first Gamma prior distribution, f(x,r,s) = constant * x^(r-1) exp(-s*x).
`sigma22`	initial value of the second variance, sigma2 squared, only required for "MMP" method with "121" or "Differ" variances.
`r2,s2`	shape and scale parameters in the second Gamma prior distribution, only required for "MMP" method with "121" or "Differ" variances.
`sigma23`	initial value of the third variance, sigma3 squared, only required for "MMP" method with "Differ" variances.
`r3,s3`	shape and scale parameters in the third Gamma prior distribution, only required for "MMP" method with "Differ" variances.

Details

The default "C-LLL" method and other two methods "C-LQL" and "U-LLL" are based on the likelihood principle, which provide the estimates of two change points, coefficients and variances of each of the three regression functions.

The "MMP" method is based on a hybrid Bayesian approach, in which each variance follows a Gamma distribution specified by a shape and a scale parameter: f(x,r,s) = constant * x^(r-1) exp(-s*x), and the pair of change points follows a uniform distribution on the data indices. Given the change points, the MLE of the regression coefficients are the LSE. Then the conditional posterior of the change points given the variances, and the conditional posterior of variances given the change points (with the MLE of regression coefficients plugged in as function of change points in both posteriors) can be maximized iteratively until the convergence to the final estimates of variances and change points. Hence this method is called Maximization-Maximization-Posterior (MMP) method.

Value

maxloglik: maximum log-likelihood value, provided only for the likelihood methods.

sigma2: up to three values of variance based on the variance specification, "Common", "121", "Differ".

coe: coefficients for three regression segments, beta = (a0,a1,b0,b1,b2,c0,c1). No b2 output for LLL model.

changepoints: the first and second change points in x values.

References

Stephen J. Ganocy and Jiayang Sun (2014), "Heteroscedastic Change Point Analysis and Application to Footprint Data", J of Data Science, In Press.

Examples

```
# Example 1: Test the hcp() using simulated data
# Simulate a C-LLL data set with a common variance 0.25,
# where the 2 change points are at x = 2 and x = 5.

x1<-seq(0,2,by=0.05)
x2<-seq(2.05,5,by=0.05)
x3<-seq(5.05,7,by=0.05)
y1<-2+2*x1+rnorm(length(x1),0,0.5)
y2<-5+0.5*x2+rnorm(length(x2),0,0.5)
y3<-17.5-2*x3+rnorm(length(x3),0,0.5)
z<-data.frame(c(x1,x2,x3),c(y1,y2,y3)); names(z)=c("x","y")
# So the true beta for data z is (2,2,5,0.5,17.5,-2).

# Visualizing the plot given by plot(z) shows that
# three segments are all linear, variances appear
# to be homoscedastic and the change points are
# in (1.5,2.5) and (4.5,5.5).
# Thus, we fit the following model:
hcp(z,1.5,2.5,4.5,5.5,"C-LLL","Common")

# All estimates look good in comparison to the
# real parameters.

# Can also try MMP method for the LLL model as
# below, if needed. The reasonable r1, s1 for
# MMP method are: r1 = 11, s1 = 60.
# hcp(dataset1,1.5,2.5,4.5,5.5,"MMP","Common","FALSE",0.25,11,60)


# Example 2: The footprint data from the tire industry
# in Ganocy and Sun (2014). The objective was to estimate
# the footprint length, i.e. the length between two change
# points in the data.

# Can visualize the built-in tire footprint data that
# comes with this package to set the intervals
# bracketing 2 change points.
plot(FP.Sample.2)

# Footprint data usually has three segments with
# a larger variance in the middle segment than those
# in the first and third segments, which is confirmed
# by the plot. It also appears that the middle segment
# is quadratic and the two change points are around
# 1 and 5.8, falling in the intervals (0.5,2) and (5,6.5).
# Hence, the C-LQL model with "121" variance should
```

```
# be the most adequate:
hcp(FP.Sample.2,0.5,2,5,6.5,"C-LQL","121")


# Extra: The following illustrates how to choose the
# hyperparameters, needed in an application of the
# MMP method. This MMP method is an iterative method
# so it is the slowest method.

# It also requires specifying the input for sigma21,
# r1, s1, if the variance type is "Common", where
# sigma21 is the variance for all three segments.
# r1, s1 are shape and scale parameters of the Gamma
# distribution for all three segments.

# The Bayes empirical estimates of r1, s1 are:
# r1 = mean/sigma2, s1 = (mean)^2/sigma2, where mean
# and sigma are the mean and sigma of the variance
# sigma21; they can be computed based on a sequence
# of local estimates of sigma21 using neighboring points.

# Input for sigma21, r1, s1, sigma22, r2, s2 is needed
# if the variance type is "121", where sigma21 is
# the variance for the first and third segments, and
# sigma22 is the variance for the second segment.

# Bayes empirical estimates of (r1,s1) and (r2,s2) are
# similar to that in the "Common" type. They can be
# obtained based on two sequences of local estimates
# of sigma21, sigma22, using the neighboring points
# in preliminary respective segments. They do not
# need to be too accurate.

# Here is an example of MMP method for this footprint
# data, though the C-LQL method fit above is more adequate:
# hcp(FP.Sample.2,0.5,2,5,6.5,"MMP","121","FALSE",10,1,10,35,1,35)
# Its second segment is a linear approximation to the
# quadratic curve (for an illustration of MMP, only).
```

Index