Package 'RDHonest'

December 16, 2024

Title Honest Inference in Regression Discontinuity Designs

Version 1.0.1

Description Honest and nearly-optimal confidence intervals in fuzzy and sharp regression discontinuity designs and for inference at a point based on local linear regression. The implementation is based on Armstrong and Kolesár (2018) <doi:10.3982/ECTA14434>, and Kolesár and Rothe (2018) <doi:10.1257/aer.20160945>. Supports covariates, clustering, and weighting.

Depends R (>= 4.3.0)

License GPL-3

Encoding UTF-8

LazyData true

Imports stats, Formula, withr

Suggests spelling, ggplot2, testthat, knitr, rmarkdown, formatR

Config/testthat/edition 3

Language en-US

URL https://github.com/kolesarm/RDHonest

BugReports https://github.com/kolesarm/RDHonest/issues

RoxygenNote 7.3.2

VignetteBuilder knitr

NeedsCompilation no

Author Michal Kolesár [aut, cre, cph] (<https://orcid.org/0000-0002-2482-7796>), Tim Armstrong [ctb]

Maintainer Michal Kolesár <kolesarmi@googlemail.com>

Repository CRAN

Date/Publication 2024-12-16 15:40:02 UTC

Contents

---

cghs *Oreopoulos (2006) UK general household survey dataset*

Description

Oreopoulos (2006) UK general household survey dataset

Usage

```
cghs
```

Format

A data frame with 73,954 rows and 2 variables:

earnings Annual earnings in 1998 (UK pounds)

yearat14 Year individual turned 14

Source

American Economic Review data archive, doi:10.1257/000282806776157641

References

Philip Oreopoulos. Estimating average and local average treatment effects when compulsory education schooling laws really matter. American Economic Review, 96(1):152–175, 2006. doi:10.1257/000282806776157641

CVb *Critical values for CIs based on a biased Gaussian estimator.*

Description

Computes the critical value $cv_{1-\alpha}(B)$ such that the confidence interval $X \pm cv_{1-\alpha}(B)$ has coverage $1-\alpha$, where the estimator X is normally distributed with variance equal to 1 and maximum bias at most B.

Usage

```
CVb(B, alpha = 0.05)
```

Arguments

B	Maximum bias, vector of non-negative numbers.
alpha	Determines CI level, $1-\alpha$. Scalar between 0 and 1.

Value

Vector of critical values, one for each value of maximum bias supplied by B.

Examples

```
## 90% critical value:
CVb(B = 1, alpha = 0.1)
## Usual 95% critical value
CVb(0)
## Returns vector with 3 critical values
CVb(B = c(0, 0.5, 1), alpha = 0.05)
```

headst *Head Start data from Ludwig and Miller (2007)*

Description

Subset of Ludwig-Miller (2007) data. Counties with missing poverty rate, or with both outcomes missing (hs and mortality) were removed. In the original dataset, Yellowstone County, MT (oldcode = 27056) was entered twice, here the duplicate is removed. Yellowstone National Park, MT (oldcode = 27057) is also removed due to it being an outlier for both outcomes. Counties with oldcode equal to (3014, 32032, 47010, 47040, 47074, 47074, 47078, 47079, 47096) matched more than one FIPS entry, so the county labels may not be correct. Mortality data is missing for Alaska.

Usage

```
headst
```

Format

A data frame with 3,127 rows and 18 variables:

statefp State FIPS code

countyfp County FIPS code

oldcode ID in Ludwig-Miller dataset

povrate Poverty rate in 1960 relative to 300th poorest county (which had poverty rate 59.1984)

mortHS Average Mortality rate per 100,000 for children aged 5-9 over 1973–83 due to causes addressed as part of Head Start's health services

mortInj Average Mortality rate per 100,000 for children aged 5-9 over 1973–83 due to injury

hs90 High school completion rate in 1990 census, ages 18-24

pop County population (1960 census)

sch1417 Percent attending school, ages 14-17 (1960 census)

sch534 Percent attending school, ages 5-34 (1960 census)

hs60 High school completion rate in 1960 census, ages 25+

pop1417 Population aged 14-17 (1960 census)

pop534 Population aged 5-34 (1960 census)

pop25 Population aged 25+ (1960 census)

urban Percent urban (1960 census)

black Percent black (1960 census)

statepc State postal code

county County name

Source

Douglas Miller's former website, http://web.archive.org/web/20190619165949/http://faculty.econ.ucdavis.edu:80/faculty/dlmiller/statafiles/

References

Jens Ludwig and Douglas L. Miller. Does head start improve children's life chances? Evidence from a regression discontinuity design. Quarterly Journal of Economics, 122(1):159–208, February 2007. doi:10.1162/qjec.122.1.159

lee08 *Lee (2008) US House elections dataset*

Description

Lee (2008) US House elections dataset

Usage

```
lee08
```

Format

A data frame with 6,558 rows and 2 variables:

voteshare Vote share in next election

margin Democratic margin of victory

Source

Mostly Harmless Econometrics data archive, https://economics.mit.edu/people/faculty/josh-angrist/mhe-data-archive

References

David S. Lee. Randomized experiments from non-random selection in U.S. House elections. Journal of Econometrics, 142(2):675–697, 2008. doi:10.1016/j.jeconom.2007.05.004

rcp *Battistin, Brugiavini, Rettore, and Weber (2009) retirement consumption puzzle dataset*

Description

Battistin, Brugiavini, Rettore, and Weber (2009) retirement consumption puzzle dataset

Usage

```
rcp
```

Format

A data frame with 30,006 rows and 6 variables:

survey_year Survey year

elig_year Years to/from eligibility (males)

retired Retirement status (males)

food Total household food expenditure

c Total household consumption

cn Total household expenditure on non-durable goods

education Educational attainment (males), one of: "none", "elementary school", "lower secondary", "vocational studies", "upper secondary", "college or higher")

family_size Family size

Source

American Economic Review data archive, doi:10.1257/aer.99.5.2209

References

Erich Battistin, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. The retirement consumption puzzle: Evidence from a regression discontinuity approach. American Economic Review, 99(5):2209–2226, 2009. doi:10.1257/aer.99.5.2209

`RDHonest` *Honest inference in RD*

Description

Calculate estimators and bias-aware CIs for the sharp or fuzzy RD parameter, or for value of the conditional mean at a point.

Usage

```
RDHonest(
  formula,
  data,
  subset,
  weights,
  cutoff = 0,
  M,
  kern = "triangular",
  na.action,
  opt.criterion = "MSE",
  h,
  se.method = "nn",
```

```
  alpha = 0.05,
  beta = 0.8,
  J = 3,
  sclass = "H",
  T0 = 0,
  point.inference = FALSE,
  sigmaY2,
  sigmaD2,
  sigmaYD,
  clusterid
)
```

Arguments

formula	an object of class "formula" (or one that can be coerced to that class). The formula syntax is outcome ~ running_variable for inference at a point. For sharp RD, it is outcome ~ running_variable if there are no covariates, or outcome ~ running_variable | covariates if covariates are present. For fuzzy RD, it is outcome | treatment ~ running_variable | covariates, with covariates optional.
data	optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the outcome and running variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which the function is called.
subset	optional vector specifying a subset of observations to be used in the fitting process.
weights	Optional vector of weights to weight the observations (useful for aggregated data). The weights are interpreted as the number of observations that each aggregated data point averages over. Disregarded if optimal kernel is used.
cutoff	specifies the RD cutoff in the running variable. For inference at a point, specifies the point x_0 at which to calculate the conditional mean.
M	Bound on second derivative of the conditional mean function, a numeric vector of length one. For fuzzy RD, M needs to be a numeric vector of length two, specifying the smoothness of the conditional mean for the outcome and treatment, respectively.
kern	specifies the kernel function used in the local regression. It can either be a string equal to "triangular" ($k(u) = (1 - \|u\|)_+$), "epanechnikov" ($k(u) = (3/4)(1 - u^2)_+$), or "uniform" ($k(u) = (\|u\| < 1)/2$), or else a kernel function. If equal to "optimal", use the finite-sample optimal linear estimator under Taylor smoothness class, instead of a local linear estimator.
na.action	function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options (usually na.omit). Another possible value is na.fail
opt.criterion	Optimality criterion that the bandwidth is designed to optimize. The options are: "MSE" Finite-sample maximum MSE "FLCI" Length of (fixed-length) two-sided confidence intervals.

"OCI" Given quantile of excess length of one-sided confidence intervals

The methods use conditional variance given by sigmaY2, if supplied. For fuzzy RD, sigmaD2 and sigmaYD also need to be supplied in this case. Otherwise, the methods use preliminary variance estimates based on assuming homoskedasticity on either side of the cutoff.

h — bandwidth, a scalar parameter. If not supplied, optimal bandwidth is computed according to criterion given by opt.criterion.

se.method — method for estimating standard error of the estimate, one of:

"nn" Nearest neighbor method

"EHW" Eicker-Huber-White, with residuals from local regression (local polynomial estimators only).

"supplied.var" Use conditional variance supplied by sigmaY2 instead of computing residuals. For fuzzy RD, sigmaD2 and sigmaYD also need to be supplied in this case.

alpha — determines confidence level, 1-alpha for constructing/optimizing confidence intervals.

beta — Determines quantile of excess length to optimize, if bandwidth optimizes given quantile of excess length of one-sided confidence intervals (opt.criterion="OCI"); otherwise ignored.

J — Number of nearest neighbors, if se.method="nn" is specified. Otherwise ignored.

sclass — Smoothness class, either "T" for Taylor or "H" for Hölder class.

T0 — Initial estimate of the treatment effect for calculating the optimal bandwidth. Only relevant for fuzzy RD.

point.inference

Do inference at a point determined by cutoff instead of RD.

sigmaY2 — Supply variance of outcome. Ignored when kernel is optimal.

sigmaD2 — Supply variance of treatment (fuzzy RD only).

sigmaYD — Supply covariance of treatment and outcome (fuzzy RD only).

clusterid — Vector specifying cluster membership. If supplied, se.method="EHW" is required, and standard errors use cluster-robust variance formulas.

Details

The bandwidth is calculated to be optimal for a given performance criterion, as specified by opt.criterion. Alternatively, for local polynomial estimators, the bandwidth can be specified by h. For kern="optimal", calculate optimal estimators under second-order Taylor smoothness class (sharp RD only).

Value

Returns an object of class "RDResults". The function print can be used to obtain and print a summary of the results. An object of class "RDResults" is a list containing four components. First, a data frame "coefficients" containing the following columns:

term type of parameter being estimated

`estimate` point estimate

`std.error` standard error of `estimate`

`maximum.bias` maximum bias of `estimate`

`conf.low`, `conf.high` lower (upper) end-point of a two-sided CI based on `estimate`

`conf.low.onesided`, `conf.high.onesided` lower (upper) end-point of a one-sided CIs based on `estimate`

`bandwidth` bandwidth used. If `kern="optimal"`, the smoothing parameters `bandwidth.m` and `bandwidth.p` on either side of the cutoff are reported instead

`eff.obs` number of effective observations

`leverage` maximal leverage of `estimate`

`cv` critical value used to compute two-sided CIs

`alpha` coverage level, as specified by option `alpha`

`method` `sclass` is used

`M` curvature bound used for worst-case bias calculations. For fuzzy RD, equals `(abs(estimate)*M.fs+M.rf)/first.stage`

`M.rf`, `M.fs` curvature bound for the outcome (i.e. reduced-form) and first-stage regressions. Fuzzy RD only.

`first.stage` estimate of the first-stage coefficient. Fuzzy RD only.

`kernel` kernel used

`p.value` p-value for testing the null of no effect

Second, a list called `"data"` containing the data used for estimation. This is useful mostly for internal calculations. Third, an object of class `"lm"` containing the local linear regression estimates. Finally, a `call` object containing the matched call called `"call"`.

If `kern="optimal"`, the `"lm"` object is empty, and the numeric vectors `"delta"` and `"omega"` are returned in addition. These correspond to the parameters in the modulus problem used to compute the optimal estimation weights.

Note

`subset` is evaluated in the same way as variables in `formula`, that is first in `data` and then in the environment of `formula`.

References

Timothy B. Armstrong and Michal Kolesár. Optimal inference in a class of regression models. Econometrica, 86(2):655–683, March 2018. doi:10.3982/ECTA14434

Timothy B. Armstrong and Michal Kolesár. Simple and honest confidence intervals in nonparametric regression. Quantitative Economics, 11(1):1–39, January 2020. doi:10.3982/QE1199

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8):2277—-2304, August 2018. doi:10.1257/aer.20160945

Examples

```
RDHonest(voteshare ~ margin, data = lee08, kern = "uniform", M = 0.1, h = 10)
RDHonest(cn | retired ~ elig_year, data=rcp, cutoff=0, M=c(4, 0.4),
          kern="triangular", opt.criterion="MSE", T0=0, h=3)
RDHonest(voteshare ~ margin, data = lee08, subset = margin>0,
          kern = "uniform", M = 0.1, h = 10, point.inference=TRUE)
```

RDHonestBME *Honest CIs in sharp RD with discrete regressors under BME function class*

Description

Computes honest CIs for local polynomial regression with uniform kernel in sharp RD under the assumption that the conditional mean lies in the bounded misspecification error (BME) class of functions, as considered in Kolesár and Rothe (2018). This class formalizes the notion that the fit of the chosen model is no worse at the cutoff than elsewhere in the estimation window.

Usage

```
RDHonestBME(
  formula,
  data,
  subset,
  cutoff = 0,
  na.action,
  h = Inf,
  alpha = 0.05,
  order = 0,
  regformula
)
```

Arguments

formula	object of class "formula" (or one that can be coerced to that class) of the form outcome ~ running_variable
data	optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the outcome and running variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which the function is called.
subset	optional vector specifying a subset of observations to be used in the fitting process.
cutoff	specifies the RD cutoff in the running variable.
na.action	function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options (usually na.omit). Another possible value is na.fail

h	bandwidth, a scalar parameter.
alpha	determines confidence level, $1 - \alpha$
order	Order of local regression 1 for linear, 2 for quadratic, etc.
regformula	Explicitly specify regression formula to use instead of running a local polynomial regression, with y and x denoting the outcome and the running variable, and cutoff is normalized to 0. Local linear regression (order = 1) is equivalent to regformula = "y~x*I(x>0)". Inference is done on the order+2th element of the design matrix

Value

An object of class "RDResults". This is a list with at least the following elements:

"coefficients" Data frame containing estimation results, including point estimate, one- and two-sided confidence intervals, a bound on worst-case bias, bandwidth used, and the number of effective observations.

"call" The matched call.

"lm" An "lm" object containing the fitted regression.

"na.action" (If relevant) information on the special handling of NAs.

Note

subset is evaluated in the same way as variables in formula, that is first in data and then in the environment of formula.

References

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8):2277—-2304, August 2018. doi:10.1257/aer.20160945

Examples

```
RDHonestBME(log(earnings)~yearat14, data=cghs, h=3,
            order=1, cutoff=1947)
## Equivalent to
RDHonestBME(log(earnings)~yearat14, data=cghs, h=3,
            cutoff=1947, order=1, regformula="y~x*I(x>=0)")
```

RDScatter *Scatterplot of binned raw observations*

Description

Scatterplot of raw observations in which each point corresponds to an binned average.

Usage

```
RDScatter(
  formula,
  data,
  subset,
  cutoff = 0,
  na.action,
  avg = 10,
  xlab = NULL,
  ylab = NULL,
  vert = TRUE,
  propdotsize = FALSE
)
```

Arguments

formula	object of class "formula" (or one that can be coerced to that class) of the form outcome ~ running_variable
data	optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the outcome and running variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which the function is called.
subset	optional vector specifying a subset of observations to be used in the fitting process.
cutoff	specifies the RD cutoff for the running variable.
na.action	function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options (usually na.omit). Another possible value is na.fail
avg	Number of observations to average over. If set to Inf, then take averages for each possible value of the running variable (convenient when the running variable is discrete).
xlab, ylab	x- and y-axis labels
vert	Draw a vertical line at cutoff?
propdotsize	If TRUE, then size of points is proportional to number of observations that the point averages over (useful when avg=Inf). Otherwise the size of points is constant.

Value

An object of class "ggplot", a scatterplot the binned raw observations.

Note

subset is evaluated in the same way as variables in formula, that is first in data and then in the environment of formula.

Examples

```
RDScatter(log(earnings)~yearat14, data=cghs, cutoff=1947,
               avg=Inf, propdotsize=TRUE)
```

RDSmoothnessBound *Lower bound on smoothness constant M in sharp RD designs*

Description

Estimate a lower bound on the smoothness constant M and provide a lower confidence interval for it, using method described in supplement to Kolesár and Rothe (2018).

Usage

```
RDSmoothnessBound(
  object,
  s,
  separate = FALSE,
  multiple = TRUE,
  alpha = 0.05,
  sclass = "H"
)
```

Arguments

object	An object of class "RDResults", typically a result of a call to RDHonest.
s	Number of support points that curvature estimates should average over.
separate	If TRUE, report estimates separately for data above and below cutoff. If FALSE, report pooled estimates.
multiple	If TRUE, use multiple curvature estimates. If FALSE, only use a single curvature estimate using observations closest to the cutoff.
alpha	determines confidence level 1-alpha.
sclass	Smoothness class, either "T" for Taylor or "H" for Hölder class.

Value

Returns a data frame wit the following columns:

`estimate` Point estimate for lower bounds for M.

`conf.low` Lower endpoint for a one-sided confidence interval for M

The data frame has a single row if `separate==FALSE`; otherwise it has two rows, corresponding to smoothness bound estimates and confidence intervals below and above the cutoff, respectively.

References

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. American Economic Review, 108(8):2277—-2304, August 2018. doi:10.1257/aer.20160945

Examples

```
## Subset data to increase speed
r <- RDHonest(log(earnings)~yearat14, data=cghs,
              subset=abs(yearat14-1947)<10,
              cutoff=1947, M=0.04, h=3)
RDSmoothnessBound(r, s=2)
```

RDTEfficiencyBound *Finite-sample efficiency bounds for minimax CIs*

Description

Compute efficiency of minimax one-sided CIs at constant functions, or efficiency of two-sided fixed-length CIs at constant functions under second-order Taylor smoothness class.

Usage

```
RDTEfficiencyBound(object, opt.criterion = "FLCI", beta = 0.5)
```

Arguments

`object`	An object of class `"RDResults"`, typically a result of a call to `RDHonest`.
`opt.criterion`	Either `"FLCI"` for computing efficiency of two-sided CIs, or else `"OCI"` for minimax one-sided CIs.
`beta`	Determines quantile of excess length for evaluating minimax efficiency of one-sided CIs. Ignored if `opt.criterion=="FLCI"`.

Value

Efficiency bound, a numeric vector of length one.

References

Timothy B. Armstrong and Michal Kolesár. Optimal inference in a class of regression models. Econometrica, 86(2):655–683, March 2018. doi:10.3982/ECTA14434

Examples

```
r <- RDHonest(voteshare ~ margin, data=lee08,
              subset=abs(margin)<10, M=0.1, h=2)
RDTEfficiencyBound(r, opt.criterion="OCI")
```

`rebp` *Austrian unemployment duration data from Lalive (2008)*

Description

Subset of Lalive (2008) data for individuals in the regions affected by the REBP program

Usage

```
rebp
```

Format

A data frame with 29,371 rows and 4 variables:

age Age in years, at monthly accuracy

period Indicator for whether REBP is in place

female Indicator for female

duration unemployment duration in weeks

Source

Rafael Lalive's website, https://sites.google.com/site/rafaellalive/

References

Rafael Lalive. How do extended benefits affect unemployment duration? A regression discontinuity approach. Journal of Econometrics, 142(2):785–806, February 2008. doi:10.1016/j.jeconom.2007.05.013

Index