Package 'binsreg'

July 23, 2021

Type Package

Title Binscatter Estimation and Inference

Date 2021-07-23

Version 0.4.3

Author Matias D. Cattaneo, Richard K. Crump, Max H. Farrell, Yingjie Feng

Maintainer Yingjie Feng <fengyingjiepku@gmail.com>

Description Provides tools for statistical analysis using the binscatter methods developed by Cattaneo, Crump, Farrell and Feng (2021a) <arXiv:1902.09608> and Cattaneo, Crump, Farrell and Feng (2021b) <arXiv:1902.09615>. Binscatter provides a flexible way of describing the mean relationship between two variables based on partitioning/binning of the independent variable of interest. binsreg(), binsqreg() and binsglm() implement binscatter least squares regression, quantile regression and generalized linear regression respectively, with particular focus on constructing binned scatter plots. They also implement robust (pointwise and uniform) inference of regression functions and derivatives thereof. binstest() implements hypothesis testing procedures for parametric functional forms of and nonparametric shape restrictions on the regression function. binspwc() implements hypothesis testing procedures for pairwise group comparison of binscatter estimators. binsregselect() implements data-driven procedures for selecting the number of bins for binscatter estimation. All the commands allow for covariate adjustment, smoothness restrictions and clustering.

Depends R (>= 3.1)

License GPL-2

Encoding UTF-8

LazyData true

Imports ggplot2, sandwich, quantreg, splines, matrixStats

Roxygen list(old_usage = TRUE)

RoxygenNote 7.1.1

R topics documented:

binsreg-package 2
binsglm 2
binspwc 7
binsqreg 11
binsreg 16
binsregselect 20
binstest 23

Index **27**

`binsreg-package` *Binsreg Package Document*

Description

Binscatter provides a flexible, yet parsimonious way of visualizing and summarizing large data sets and has been a popular methodology in applied microeconomics and other social sciences. The binsreg package provides tools for statistical analysis using the binscatter methods developed in Cattaneo, Crump, Farrell and Feng (2021a). `binsreg` implements binscatter least squares regression with robust inference and plots, including curve estimation, pointwise confidence intervals and uniform confidence band. `binsqreg` implements binscatter quantile regression with robust inference and plots, including curve estimation, pointwise confidence intervals and uniform confidence band. `binsglm` implements binscatter generalized linear regression with robust inference and plots, including curve estimation, pointwise confidence intervals and uniform confidence band. `binstest` implements binscatter-based hypothesis testing procedures for parametric specifications of and shape restrictions on the unknown function of interest. `binspwc` implements hypothesis testing procedures for pairwise group comparison of binscatter estimators. `binsregselect` implements data-driven number of bins selectors for binscatter implementation using either quantile-spaced or evenly-spaced binning/partitioning. All the commands allow for covariate adjustment, smoothness restrictions, and clustering, among other features.

The companion software article, Cattaneo, Crump, Farrell and Feng (2021b), provides further implementation details and empirical illustration. For related Stata and R packages useful for nonparametric data analysis and statistical inference, visit https://nppackages.github.io/.

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

`binsglm` *Data-Driven Binscatter Generalized Linear Regression with Robust Inference Procedures and Plots*

Description

binsglm implements binscatter generalized linear regression with robust inference procedures and plots, following the results in Cattaneo, Crump, Farrell and Feng (2021a). Binscatter provides a flexible way to describe the relationship between two variables, after possibly adjusting for other covariates, based on partitioning/binning of the independent variable of interest. The main purpose of this function is to generate binned scatter plots with curve estimation with robust pointwise confidence intervals and uniform confidence band. If the binning scheme is not set by the user, the companion function binsregselect is used to implement binscatter in a data-driven way. Hypothesis testing about the function of interest can be conducted via the companion function binstest.

Usage

```
binsglm(y, x, w = NULL, data = NULL, at = NULL, family = gaussian(),
  deriv = 0, nolink = F, dots = c(0, 0), dotsgrid = 0,
  dotsgridmean = T, line = NULL, linegrid = 20, ci = NULL,
  cigrid = 0, cigridmean = T, cb = NULL, cbgrid = 20, polyreg = NULL,
  polyreggrid = 20, polyregcigrid = 0, by = NULL, bycolors = NULL,
  bysymbols = NULL, bylpatterns = NULL, legendTitle = NULL,
  legendoff = F, nbins = NULL, binspos = "qs", binsmethod = "dpi",
  nbinsrot = NULL, samebinsby = F, randcut = NULL, nsims = 500,
  simsgrid = 20, simsseed = NULL, vce = "HC1", cluster = NULL,
  asyvar = F, level = 95, noplot = F, dfcheck = c(20, 30),
  masspoints = "on", weights = NULL, subset = NULL, plotxrange = NULL,
  plotyrange = NULL)
```

Arguments

y	outcome variable. A vector.
x	independent variable of interest. A vector.
w	control variables. A matrix, a vector or a formula.
data	an optional data frame containing variables in the model.
at	value of w at which the estimated function is evaluated. The default is at="mean", which corresponds to the mean of w. Other options are: at="median" for the median of w, at="zero" for a vector of zeros. at can also be a vector of the same length as the number of columns of w (if w is a matrix) or a data frame containing the same variables as specified in w (when data is specified). Note that when at="mean" or at="median", all factor variables (if specified) are excluded from the evaluation (set as zero).
family	a description of the error distribution and link function to be used in the generalized linear model. (See family for details of family functions.)
deriv	derivative order of the regression function for estimation, testing and plotting. The default is deriv=0, which corresponds to the function itself. If nolink=TRUE, deriv cannot be greater than 1.
nolink	if true, the function within the inverse link function is reported instead of the conditional mean function for the outcome.
dots	a vector. dots=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for point estimation and plotting as "dots". The default is dots=c(0,0), which corresponds to piecewise constant (canonical binscatter)

dotsgrid	number of dots within each bin to be plotted. Given the choice, these dots are point estimates evaluated over an evenly-spaced grid within each bin. The default is dotsgrid=0, and only the point estimates at the mean of x within each bin are presented.
dotsgridmean	If true, the dots corresponding to the point estimates evaluated at the mean of x within each bin are presented. By default, they are presented, i.e., dotsgridmean=T.
line	a vector. line=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for plotting as a "line". By default, the line is not included in the plot unless explicitly specified. Recommended specification is line=c(3,3), which adds a cubic B-spline estimate of the regression function of interest to the binned scatter plot.
linegrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the line=c(p,s) option. The default is linegrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for fitting/plotting the line.
ci	a vector. ci=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints used for constructing confidence intervals. By default, the confidence intervals are not included in the plot unless explicitly specified. Recommended specification is ci=c(3,3), which adds confidence intervals based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.
cigrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the ci=c(p,s) option. The default is cigrid=1, which corresponds to 1 evenly-spaced evaluation point within each bin for confidence interval construction.
cigridmean	If true, the confidence intervals corresponding to the point estimates evaluated at the mean of x within each bin are presented. The default is cigridmean=T.
cb	a vector. cb=c(p,s) sets a the piecewise polynomial of degree p with s smoothness constraints used for constructing the confidence band. By default, the confidence band is not included in the plot unless explicitly specified. Recommended specification is cb=c(3,3), which adds a confidence band based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.
cbgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the cb=c(p,s) option. The default is cbgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.
polyreg	degree of a global polynomial regression model for plotting. By default, this fit is not included in the plot unless explicitly specified. Recommended specification is polyreg=3, which adds a cubic (global) polynomial fit of the regression function of interest to the binned scatter plot.
polyreggrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the polyreg=p option. The default is polyreggrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.
polyregcigrid	number of evaluation points of an evenly-spaced grid within each bin used for constructing confidence intervals based on polynomial regression set by the polyreg=p option. The default is polyregcigrid=0, which corresponds to not plotting confidence intervals for the global polynomial regression approximation.

by	a vector containing the group indicator for subgroup analysis; both numeric and string variables are supported. When by is specified, binsreg implements estimation and inference for each subgroup separately, but produces a common binned scatter plot. By default, the binning structure is selected for each subgroup separately, but see the option samebinsby below for imposing a common binning structure across subgroups.
bycolors	an ordered list of colors for plotting each subgroup series defined by the option by.
bysymbols	an ordered list of symbols for plotting each subgroup series defined by the option by.
bylpatterns	an ordered list of line patterns for plotting each subgroup series defined by the option by.
legendTitle	String, title of legend.
legendoff	If true, no legend is added.
nbins	number of bins for partitioning/binning of x. If not specified, the number of bins is selected via the companion function binsregselect in a data-driven, optimal way whenever possible.
binspos	position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options are "es" for evenly-spaced binning, or a vector for manual specification of the positions of inner knots (which must be within the range of x).
binsmethod	method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.
nbinsrot	initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.
samebinsby	if true, a common partitioning/binning structure across all subgroups specified by the option by is forced. The knots positions are selected according to the option binspos and using the full sample. If nbins is not specified, then the number of bins is selected via the companion command binsregselect and using the full sample.
randcut	upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.
nsims	number of random draws for constructing confidence bands. The default is nsims=500, which corresponds to 500 draws from a standard Gaussian random vector of size [(p+1)*J -(J-1)*s].
simsgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum operation needed to construct confidence bands. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum operator.
simsseed	seed for simulation.
vce	Procedure to compute the variance-covariance matrix estimator. Options are • "const" homoskedastic variance estimator. • "HC0" heteroskedasticity-robust plug-in residuals variance estimator without weights. • "HC1" heteroskedasticity-robust plug-in residuals variance estimator with hc1 weights. Default.

- "HC2" heteroskedasticity-robust plug-in residuals variance estimator with hc2 weights.
- "HC3" heteroskedasticity-robust plug-in residuals variance estimator with hc3 weights.

`cluster` cluster ID. Used for compute cluster-robust standard errors.

`asyvar` If true, the standard error of the nonparametric component is computed and the uncertainty related to control variables is omitted. Default is `asyvar=FALSE`, that is, the uncertainty related to control variables is taken into account.

`level` nominal confidence level for confidence interval and confidence band estimation. Default is `level=95`.

`noplot` If true, no plot produced.

`dfcheck` adjustments for minimum effective sample size checks, which take into account number of unique values of `x` (i.e., number of mass points), number of clusters, and degrees of freedom of the different stat models considered. The default is `dfcheck=c(20,30)`. See Cattaneo, Crump, Farrell and Feng (2021b) for more details.

`masspoints` how mass points in `x` are handled. Available options:

- "on" all mass point and degrees of freedom checks are implemented. Default.
- "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted.
- "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted.
- "off" "noadjust" and "nolocalcheck" are set simultaneously.
- "veryfew" forces the function to proceed as if `x` has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.

`weights` an optional vector of weights to be used in the fitting process. Should be `NULL` or a numeric vector. For more details, see `lm`.

`subset` Optional rule specifying a subset of observations to be used.

`plotxrange` a vector. `plotxrange=c(min,max)` specifies a range of the x-axis for binscatter plot. Observations outside the range are dropped in the plot.

`plotyrange` a vector. `plotyrange=c(min,max)` specifies a range of the y-axis for binscatter plot. Observations outside the range are dropped in the plot.

Value

`bins_plot` A `ggplot` object for binscatter plot.

`data.plot` A list containing data for plotting. Each item is a sublist of data frames for each group. Each sublist may contain the following data frames:

- `data.dots` Data for dots. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.line` Data for line. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.ci` Data for CI. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `ci.l` and `ci.r`, left and right boundaries of each confidence intervals.

- `data.cb` Data for CB. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `cb.l` and `cb.r`, left and right boundaries of the confidence band.
- `data.poly` Data for polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.polyci` Data for confidence intervals based on polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `polyci.l` and `polyci.r`, left and right boundaries of each confidence intervals.

`cval.by` A vector of critical values for constructing confidence band for each group.

`opt` A list containing options passed to the function, as well as `N.by` (total sample size for each group), `Ndist.by` (number of distinct values in `x` for each group), `Nclust.by` (number of clusters for each group), and `nbins.by` (number of bins for each group), and `byvals` (number of distinct values in `by`).

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

`binsregselect`, `binstest`.

Examples

```
x <- runif(500); d <- 1*(runif(500)<=x)
## Binned scatterplot
binsglm(d, x, family=binomial())
```

`binspwc` *Data-Driven Pairwise Group Comparison using Binscatter Methods*

Description

`binspwc` implements hypothesis testing procedures for pairwise group comparison of binscatter estimators, following the results in Cattaneo, Crump, Farrell and Feng (2021a). If the binning scheme is not set by the user, the companion function `binsregselect` is used to implement binscatter in a data-driven way. Binned scatter plots based on different methods can be constructed using the companion functions `binsreg`, `binsqreg` or `binsglm`. Hypothesis testing about the function of interest can also be conducted via the companion function `binstest`.

Usage

```
binspwc(y, x, w = NULL, data = NULL, estmethod = "reg",
  family = gaussian(), quantile = NULL, deriv = 0, at = NULL,
  nolink = F, by = NULL, pwc = c(3, 3), testtype = "two-sided",
  lp = Inf, bins = c(0, 0), bynbins = NULL, binspos = "qs",
  binsmethod = "dpi", nbinsrot = NULL, samebinsby = FALSE,
  randcut = NULL, nsims = 500, simsgrid = 20, simsseed = NULL,
  vce = NULL, cluster = NULL, asyvar = F, dfcheck = c(20, 30),
  masspoints = "on", weights = NULL, subset = NULL, numdist = NULL,
  numclust = NULL, ...)
```

Arguments

y	outcome variable. A vector.
x	independent variable of interest. A vector.
w	control variables. A matrix, a vector or a formula.
data	an optional data frame containing variables used in the model.
estmethod	estimation method. The default is estmethod="reg" for tests based on binscatter least squares regression. Other options are "qreg" for quantile regression and "glm" for generalized linear regression. If estmethod="glm", the option family must be specified.
family	a description of the error distribution and link function to be used in the generalized linear model when estmethod="glm". (See family for details of family functions.)
quantile	the quantile to be estimated. A number strictly between 0 and 1.
deriv	derivative order of the regression function for estimation, testing and plotting. The default is deriv=0, which corresponds to the function itself.
at	value of w at which the estimated function is evaluated. The default is at="mean", which corresponds to the mean of w. Other options are: at="median" for the median of w, at="zero" for a vector of zeros. at can also be a vector of the same length as the number of columns of w (if w is a matrix) or a data frame containing the same variables as specified in w (when data is specified). Note that when at="mean" or at="median", all factor variables (if specified) are excluded from the evaluation (set as zero).
nolink	if true, the function within the inverse link function is reported instead of the conditional mean function for the outcome.
by	a vector containing the group indicator for subgroup analysis; both numeric and string variables are supported. When by is specified, binsreg implements estimation and inference for each subgroup separately, but produces a common binned scatter plot. By default, the binning structure is selected for each subgroup separately, but see the option samebinsby below for imposing a common binning structure across subgroups.
pwc	a vector. pwc=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for testing the difference between groups. The default is pwc=c(3,3), which corresponds to a cubic B-spline estimate of the function of interest for each group.
testtype	type of pairwise comparison test. The default is testtype="two-sided", which corresponds to a two-sided test of the form H0: mu_1(x)=mu_2(x). Other options are: testtype="left" for the one-sided test form H0: mu_1(x)<=mu_2(x) and testtype="right" for the one-sided test of the form H0: mu_1(x)>=mu_2(x).

lp	an Lp metric used for (two-sided) parametric model specification testing and/or shape restriction testing. The default is lp=Inf, which corresponds to the sup-norm of the t-statistic. Other options are lp=q for a positive integer q.
bins	A vector. Degree and smoothness for bin selection.
bynbins	a vector of the number of bins for partitioning/binning of x, which is applied to the binscatter estimation for each group. If not specified, the number of bins is selected via the companion function binsregselect in a data-driven way whenever possible.
binspos	position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options are "es" for evenly-spaced binning, or a vector for manual specification of the positions of inner knots (which must be within the range of x).
binsmethod	method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.
nbinsrot	initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.
samebinsby	if true, a common partitioning/binning structure across all subgroups specified by the option by is forced. The knots positions are selected according to the option binspos and using the full sample. If nbins is not specified, then the number of bins is selected via the companion command binsregselect and using the full sample.
randcut	upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.
nsims	number of random draws for hypothesis testing. The default is nsims=500, which corresponds to 500 draws from a standard Gaussian random vector of size [(p+1)*J-(J-1)*s].
simsgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum (infimum or Lp metric) operation needed to construct hypothesis testing procedures. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum (infimum or Lp metric) operator.
simsseed	seed for simulation.
vce	procedure to compute the variance-covariance matrix estimator. For least squares regression and generalized linear regression, the allowed options are the same as that for binsreg or binsqreg. For quantile regression, the allowed options are the same as that for binsqreg.
cluster	cluster ID. Used for compute cluster-robust standard errors.
asyvar	If true, the standard error of the nonparametric component is computed and the uncertainty related to control variables is omitted. Default is asyvar=FALSE, that is, the uncertainty related to control variables is taken into account.
dfcheck	adjustments for minimum effective sample size checks, which take into account number of unique values of x (i.e., number of mass points), number of clusters, and degrees of freedom of the different stat models considered. The default is dfcheck=c(20,30). See Cattaneo, Crump, Farrell and Feng (2021b) for more details.
masspoints	how mass points in x are handled. Available options:

- "on" all mass point and degrees of freedom checks are implemented. Default.
- "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted.
- "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted.
- "off" "noadjust" and "nolocalcheck" are set simultaneously.
- "veryfew" forces the function to proceed as if x has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.

weights — an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. For more details, see lm.

subset — optional rule specifying a subset of observations to be used.

numdist — Number of distinct for selection. Used to speed up computation.

numclust — Number of clusters for selection. Used to speed up computation.

... — optional arguments to control bootstrapping if estmethod="qreg" and vce="boot". See boot.rq.

Value

stat — A matrix. Each row corresponds to the comparison between two groups. The first column is the test statistic. The second and third columns give the corresponding group numbers. The null hypothesis is mu_i(x)<=mu_j(x), mu_i(x)=mu_j(x), or mu_i(x)>=mu_j(x) for group i (given in the second column) and group j (given in the third column). The group number corresponds to the list of group names given by opt$byvals.

pval — A vector of p-values for all pairwise group comparisons.

opt — A list containing options passed to the function, as well as N.by (total sample size for each group), Ndist.by (number of distinct values in x for each group), Nclust.by (number of clusters for each group), and nbins.by (number of bins for each group), and byvals (number of distinct values in by).

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

binsregselect, binstest.

Examples

```
x <- runif(500); y <- sin(x)+rnorm(500); t <- 1*(runif(500)>0.5)
## Binned scatterplot
binspwc(y,x, by=t)
```

`binsqreg` *Data-Driven Binscatter Quantile Regression with Robust Inference Procedures and Plots*

Description

`binsqreg` implements binscatter quantile regression with robust inference procedures and plots, following the results in Cattaneo, Crump, Farrell and Feng (2021a). Binscatter provides a flexible way to describe the quantile relationship between two variables, after possibly adjusting for other covariates, based on partitioning/binning of the independent variable of interest. The main purpose of this function is to generate binned scatter plots with curve estimation with robust pointwise confidence intervals and uniform confidence band. If the binning scheme is not set by the user, the companion function `binsregselect` is used to implement binscatter in a data-driven way. Hypothesis testing about the function of interest can be conducted via the companion function `binstest`.

Usage

```
binsqreg(y, x, w = NULL, data = NULL, at = NULL, quantile = 0.5,
  deriv = 0, dots = c(0, 0), dotsgrid = 0, dotsgridmean = T,
  line = NULL, linegrid = 20, ci = NULL, cigrid = 0, cigridmean = T,
  cb = NULL, cbgrid = 20, polyreg = NULL, polyreggrid = 20,
  polyregcigrid = 0, by = NULL, bycolors = NULL, bysymbols = NULL,
  bylpatterns = NULL, legendTitle = NULL, legendoff = F, nbins = NULL,
  binspos = "qs", binsmethod = "dpi", nbinsrot = NULL, samebinsby = F,
  randcut = NULL, nsims = 500, simsgrid = 20, simsseed = NULL,
  vce = "nid", cluster = NULL, asyvar = F, level = 95, noplot = F,
  dfcheck = c(20, 30), masspoints = "on", weights = NULL,
  subset = NULL, plotxrange = NULL, plotyrange = NULL, ...)
```

Arguments

`y`	outcome variable. A vector.
`x`	independent variable of interest. A vector.
`w`	control variables. A matrix, a vector or a `formula`.
`data`	an optional data frame containing variables in the model.
`at`	value of `w` at which the estimated function is evaluated. The default is `at="mean"`, which corresponds to the mean of `w`. Other options are: `at="median"` for the median of `w`, `at="zero"` for a vector of zeros. `at` can also be a vector of the same length as the number of columns of `w` (if `w` is a matrix) or a data frame containing the same variables as specified in `w` (when `data` is specified). Note that when `at="mean"` or `at="median"`, all factor variables (if specified) are excluded from the evaluation (set as zero).
`quantile`	the quantile to be estimated. A number strictly between 0 and 1.
`deriv`	derivative order of the regression function for estimation, testing and plotting. The default is `deriv=0`, which corresponds to the function itself.

dots — a vector. dots=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for point estimation and plotting as "dots". The default is dots=c(0,0), which corresponds to piecewise constant (canonical binscatter)

dotsgrid — number of dots within each bin to be plotted. Given the choice, these dots are point estimates evaluated over an evenly-spaced grid within each bin. The default is dotsgrid=0, and only the point estimates at the mean of x within each bin are presented.

dotsgridmean — If true, the dots corresponding to the point estimates evaluated at the mean of x within each bin are presented. By default, they are presented, i.e., dotsgridmean=T.

line — a vector. line=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for plotting as a "line". By default, the line is not included in the plot unless explicitly specified. Recommended specification is line=c(3,3), which adds a cubic B-spline estimate of the regression function of interest to the binned scatter plot.

linegrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the line=c(p,s) option. The default is linegrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for fitting/plotting the line.

ci — a vector. ci=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints used for constructing confidence intervals. By default, the confidence intervals are not included in the plot unless explicitly specified. Recommended specification is ci=c(3,3), which adds confidence intervals based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.

cigrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the ci=c(p,s) option. The default is cigrid=1, which corresponds to 1 evenly-spaced evaluation point within each bin for confidence interval construction.

cigridmean — If true, the confidence intervals corresponding to the point estimates evaluated at the mean of x within each bin are presented. The default is cigridmean=T.

cb — a vector. cb=c(p,s) sets a the piecewise polynomial of degree p with s smoothness constraints used for constructing the confidence band. By default, the confidence band is not included in the plot unless explicitly specified. Recommended specification is cb=c(3,3), which adds a confidence band based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.

cbgrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the cb=c(p,s) option. The default is cbgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.

polyreg — degree of a global polynomial regression model for plotting. By default, this fit is not included in the plot unless explicitly specified. Recommended specification is polyreg=3, which adds a cubic (global) polynomial fit of the regression function of interest to the binned scatter plot.

polyreggrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the polyreg=p option. The default is polyreggrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.

polyregcigrid — number of evaluation points of an evenly-spaced grid within each bin used for constructing confidence intervals based on polynomial regression set by the

polyreg=p option. The default is polyregcigrid=0, which corresponds to not plotting confidence intervals for the global polynomial regression approximation.

by — a vector containing the group indicator for subgroup analysis; both numeric and string variables are supported. When by is specified, binsreg implements estimation and inference for each subgroup separately, but produces a common binned scatter plot. By default, the binning structure is selected for each subgroup separately, but see the option samebinsby below for imposing a common binning structure across subgroups.

bycolors — an ordered list of colors for plotting each subgroup series defined by the option by.

bysymbols — an ordered list of symbols for plotting each subgroup series defined by the option by.

bylpatterns — an ordered list of line patterns for plotting each subgroup series defined by the option by.

legendTitle — String, title of legend.

legendoff — If true, no legend is added.

nbins — number of bins for partitioning/binning of x. If not specified, the number of bins is selected via the companion function binsregselect in a data-driven, optimal way whenever possible.

binspos — position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options are "es" for evenly-spaced binning, or a vector for manual specification of the positions of inner knots (which must be within the range of x).

binsmethod — method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.

nbinsrot — initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.

samebinsby — if true, a common partitioning/binning structure across all subgroups specified by the option by is forced. The knots positions are selected according to the option binspos and using the full sample. If nbins is not specified, then the number of bins is selected via the companion command binsregselect and using the full sample.

randcut — upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.

nsims — number of random draws for constructing confidence bands. The default is nsims=500, which corresponds to 500 draws from a standard Gaussian random vector of size [(p+1)*J-(J-1)*s].

simsgrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum operation needed to construct confidence bands. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum operator.

simsseed — seed for simulation.

vce — Procedure to compute the variance-covariance matrix estimator (see summary.rq for more details). Options are

- "iid" which presumes that the errors are iid and computes an estimate of the asymptotic covariance matrix as in KB(1978).
- "nid" which presumes local (in quantile) linearity of the the conditional quantile functions and computes a Huber sandwich estimate using a local estimate of the sparsity.
- "ker" which uses a kernel estimate of the sandwich as proposed by Powell (1991).
- "boot" which implements one of several possible bootstrapping alternatives for estimating standard errors including a variate of the wild bootstrap for clustered response. See boot.rq for further details.

cluster — cluster ID. Used for compute cluster-robust standard errors.

asyvar — If true, the standard error of the nonparametric component is computed and the uncertainty related to control variables is omitted. Default is asyvar=FALSE, that is, the uncertainty related to control variables is taken into account.

level — nominal confidence level for confidence interval and confidence band estimation. Default is level=95.

noplot — If true, no plot produced.

dfcheck — adjustments for minimum effective sample size checks, which take into account number of unique values of x (i.e., number of mass points), number of clusters, and degrees of freedom of the different statistical models considered. The default is dfcheck=c(20,30). See Cattaneo, Crump, Farrell and Feng (2021b) for more details.

masspoints — how mass points in x are handled. Available options:

- "on" all mass point and degrees of freedom checks are implemented. Default.
- "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted.
- "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted.
- "off" "noadjust" and "nolocalcheck" are set simultaneously.
- "veryfew" forces the function to proceed as if x has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.

weights — an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. For more details, see lm.

subset — optional rule specifying a subset of observations to be used.

plotxrange — a vector. plotxrange=c(min,max) specifies a range of the x-axis for plotting. Observations outside the range are dropped in the plot.

plotyrange — a vector. plotyrange=c(min,max) specifies a range of the y-axis for plotting. Observations outside the range are dropped in the plot.

... — optional arguments to control bootstrapping. See boot.rq.

Value

bins_plot — A ggplot object for binscatter plot.

data.plot — A list containing data for plotting. Each item is a sublist of data frames for each group. Each sublist may contain the following data frames:

- `data.dots` Data for dots. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.line` Data for line. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.ci` Data for CI. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `ci.l` and `ci.r`, left and right boundaries of each confidence intervals.
- `data.cb` Data for CB. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `cb.l` and `cb.r`, left and right boundaries of the confidence band.
- `data.poly` Data for polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.polyci` Data for confidence intervals based on polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `polyci.l` and `polyci.r`, left and right boundaries of each confidence intervals.

`cval.by` A vector of critical values for constructing confidence band for each group.

`opt` A list containing options passed to the function, as well as `N.by` (total sample size for each group), `Ndist.by` (number of distinct values in `x` for each group), `Nclust.by` (number of clusters for each group), and `nbins.by` (number of bins for each group), and `byvals` (number of distinct values in `by`).

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

`binsregselect`, `binstest`.

Examples

```
x <- runif(500); y <- sin(x)+rnorm(500)
## Binned scatterplot
binsqreg(y,x)
```

binsreg *Data-Driven Binscatter Least Squares Regression with Robust Inference Procedures and Plots*

Description

binsreg implements binscatter least squares regression with robust inference procedures and plots, following the results in Cattaneo, Crump, Farrell and Feng (2021a). Binscatter provides a flexible way to describe the mean relationship between two variables, after possibly adjusting for other covariates, based on partitioning/binning of the independent variable of interest. The main purpose of this function is to generate binned scatter plots with curve estimation with robust pointwise confidence intervals and uniform confidence band. If the binning scheme is not set by the user, the companion function binsregselect is used to implement binscatter in a data-driven (optimal) way. Hypothesis testing about the regression function can be conducted via the companion function binstest.

Usage

```
binsreg(y, x, w = NULL, data = NULL, at = NULL, deriv = 0,
  dots = c(0, 0), dotsgrid = 0, dotsgridmean = T, line = NULL,
  linegrid = 20, ci = NULL, cigrid = 0, cigridmean = T, cb = NULL,
  cbgrid = 20, polyreg = NULL, polyreggrid = 20, polyregcigrid = 0,
  by = NULL, bycolors = NULL, bysymbols = NULL, bylpatterns = NULL,
  legendTitle = NULL, legendoff = F, nbins = NULL, binspos = "qs",
  binsmethod = "dpi", nbinsrot = NULL, samebinsby = F, randcut = NULL,
  nsims = 500, simsgrid = 20, simsseed = NULL, vce = "HC1",
  cluster = NULL, asyvar = F, level = 95, noplot = F, dfcheck = c(20,
  30), masspoints = "on", weights = NULL, subset = NULL,
  plotxrange = NULL, plotyrange = NULL)
```

Arguments

y	outcome variable. A vector.
x	independent variable of interest. A vector.
w	control variables. A matrix, a vector or a formula.
data	an optional data frame containing variables used in the model.
at	value of w at which the estimated function is evaluated. The default is at="mean", which corresponds to the mean of w. Other options are: at="median" for the median of w, at="zero" for a vector of zeros. at can also be a vector of the same length as the number of columns of w (if w is a matrix) or a data frame containing the same variables as specified in w (when data is specified). Note that when at="mean" or at="median", all factor variables (if specified) are excluded from the evaluation (set as zero).
deriv	derivative order of the regression function for estimation, testing and plotting. The default is deriv=0, which corresponds to the function itself.
dots	a vector. dots=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for point estimation and plotting as "dots". The default is dots=c(0,0), which corresponds to piecewise constant (canonical binscatter)

dotsgrid — number of dots within each bin to be plotted. Given the choice, these dots are point estimates evaluated over an evenly-spaced grid within each bin. The default is dotsgrid=0, and only the point estimates at the mean of x within each bin are presented.

dotsgridmean — If true, the dots corresponding to the point estimates evaluated at the mean of x within each bin are presented. By default, they are presented, i.e., dotsgridmean=T.

line — a vector. line=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for plotting as a "line". By default, the line is not included in the plot unless explicitly specified. Recommended specification is line=c(3,3), which adds a cubic B-spline estimate of the regression function of interest to the binned scatter plot.

linegrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the line=c(p,s) option. The default is linegrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for fitting/plotting the line.

ci — a vector. ci=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints used for constructing confidence intervals. By default, the confidence intervals are not included in the plot unless explicitly specified. Recommended specification is ci=c(3,3), which adds confidence intervals based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.

cigrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the ci=c(p,s) option. The default is cigrid=1, which corresponds to 1 evenly-spaced evaluation point within each bin for confidence interval construction.

cigridmean — If true, the confidence intervals corresponding to the point estimates evaluated at the mean of x within each bin are presented. The default is cigridmean=T.

cb — a vector. cb=c(p,s) sets a the piecewise polynomial of degree p with s smoothness constraints used for constructing the confidence band. By default, the confidence band is not included in the plot unless explicitly specified. Recommended specification is cb=c(3,3), which adds a confidence band based on cubic B-spline estimate of the regression function of interest to the binned scatter plot.

cbgrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the cb=c(p,s) option. The default is cbgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.

polyreg — degree of a global polynomial regression model for plotting. By default, this fit is not included in the plot unless explicitly specified. Recommended specification is polyreg=3, which adds a cubic (global) polynomial fit of the regression function of interest to the binned scatter plot.

polyreggrid — number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the point estimate set by the polyreg=p option. The default is polyreggrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for confidence interval construction.

polyregcigrid — number of evaluation points of an evenly-spaced grid within each bin used for constructing confidence intervals based on polynomial regression set by the polyreg=p option. The default is polyregcigrid=0, which corresponds to not plotting confidence intervals for the global polynomial regression approximation.

by	a vector containing the group indicator for subgroup analysis; both numeric and string variables are supported. When by is specified, binsreg implements estimation and inference for each subgroup separately, but produces a common binned scatter plot. By default, the binning structure is selected for each subgroup separately, but see the option samebinsby below for imposing a common binning structure across subgroups.
bycolors	an ordered list of colors for plotting each subgroup series defined by the option by.
bysymbols	an ordered list of symbols for plotting each subgroup series defined by the option by.
bylpatterns	an ordered list of line patterns for plotting each subgroup series defined by the option by.
legendTitle	String, title of legend.
legendoff	If true, no legend is added.
nbins	number of bins for partitioning/binning of x. If not specified, the number of bins is selected via the companion function binsregselect in a data-driven, optimal way whenever possible.
binspos	position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options are "es" for evenly-spaced binning, or a vector for manual specification of the positions of inner knots (which must be within the range of x).
binsmethod	method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.
nbinsrot	initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.
samebinsby	if true, a common partitioning/binning structure across all subgroups specified by the option by is forced. The knots positions are selected according to the option binspos and using the full sample. If nbins is not specified, then the number of bins is selected via the companion command binsregselect and using the full sample.
randcut	upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.
nsims	number of random draws for constructing confidence bands. The default is nsims=500, which corresponds to 500 draws from a standard Gaussian random vector of size [(p+1)*J -(J-1)*s].
simsgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum operation needed to construct confidence bands. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum operator.
simsseed	seed for simulation.
vce	Procedure to compute the variance-covariance matrix estimator. Options are • "const" homoskedastic variance estimator. • "HC0" heteroskedasticity-robust plug-in residuals variance estimator without weights. • "HC1" heteroskedasticity-robust plug-in residuals variance estimator with hc1 weights. Default.

- "HC2" heteroskedasticity-robust plug-in residuals variance estimator with hc2 weights.
- "HC3" heteroskedasticity-robust plug-in residuals variance estimator with hc3 weights.

cluster — cluster ID. Used for compute cluster-robust standard errors.

asyvar — If true, the standard error of the nonparametric component is computed and the uncertainty related to control variables is omitted. Default is asyvar=FALSE, that is, the uncertainty related to control variables is taken into account.

level — nominal confidence level for confidence interval and confidence band estimation. Default is level=95.

noplot — If true, no plot produced.

dfcheck — adjustments for minimum effective sample size checks, which take into account number of unique values of x (i.e., number of mass points), number of clusters, and degrees of freedom of the different statistical models considered. The default is dfcheck=c(20,30). See Cattaneo, Crump, Farrell and Feng (2021b) for more details.

masspoints — how mass points in x are handled. Available options:

- "on" all mass point and degrees of freedom checks are implemented. Default.
- "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted.
- "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted.
- "off" "noadjust" and "nolocalcheck" are set simultaneously.
- "veryfew" forces the function to proceed as if x has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.

weights — an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. For more details, see lm.

subset — Optional rule specifying a subset of observations to be used.

plotxrange — a vector. plotxrange=c(min,max) specifies a range of the x-axis for plotting. Observations outside the range are dropped in the plot.

plotyrange — a vector. plotyrange=c(min,max) specifies a range of the y-axis for plotting. Observations outside the range are dropped in the plot.

Value

bins_plot — A ggplot object for binscatter plot.

data.plot — A list containing data for plotting. Each item is a sublist of data frames for each group. Each sublist may contain the following data frames:

- data.dots Data for dots. It contains: x, evaluation points; bin, the indicator of bins; isknot, indicator of inner knots; mid, midpoint of each bin; and fit, fitted values.
- data.line Data for line. It contains: x, evaluation points; bin, the indicator of bins; isknot, indicator of inner knots; mid, midpoint of each bin; and fit, fitted values.
- data.ci Data for CI. It contains: x, evaluation points; bin, the indicator of bins; isknot, indicator of inner knots; mid, midpoint of each bin; ci.l and ci.r, left and right boundaries of each confidence intervals.

- `data.cb` Data for CB. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `cb.l` and `cb.r`, left and right boundaries of the confidence band.
- `data.poly` Data for polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; and `fit`, fitted values.
- `data.polyci` Data for confidence intervals based on polynomial regression. It contains: `x`, evaluation points; `bin`, the indicator of bins; `isknot`, indicator of inner knots; `mid`, midpoint of each bin; `polyci.l` and `polyci.r`, left and right boundaries of each confidence intervals.

`cval.by` A vector of critical values for constructing confidence band for each group.

`opt` A list containing options passed to the function, as well as `N.by` (total sample size for each group), `Ndist.by` (number of distinct values in `x` for each group), `Nclust.by` (number of clusters for each group), and `nbins.by` (number of bins for each group), and `byvals` (number of distinct values in `by`).

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

`binsregselect`, `binstest`.

Examples

```
x <- runif(500); y <- sin(x)+rnorm(500)
## Binned scatterplot
binsreg(y,x)
```

---

binsregselect — *Data-Driven IMSE-Optimal Partitioning/Binning Selection for Binscatter*

Description

`binsregselect` implements data-driven procedures for selecting the number of bins for binscatter estimation. The selected number is optimal in minimizing integrated mean squared error (IMSE).

Usage

```
binsregselect(y, x, w = NULL, data = NULL, deriv = 0, bins = c(0, 0),
  binspos = "qs", binsmethod = "dpi", nbinsrot = NULL, simsgrid = 20,
  savegrid = F, vce = "HC1", useeffn = NULL, randcut = NULL,
  cluster = NULL, dfcheck = c(20, 30), masspoints = "on",
  weights = NULL, subset = NULL, norotnorm = F, numdist = NULL,
  numclust = NULL)
```

Arguments

y	outcome variable. A vector.
x	independent variable of interest. A vector.
w	control variables. A matrix, a vector or a formula.
data	an optional data frame containing variables used in the model.
deriv	derivative order of the regression function for estimation, testing and plotting. The default is deriv=0, which corresponds to the function itself.
bins	a vector. bins=c(p,s) set a piecewise polynomial of degree p with s smoothness constraints for data-driven (IMSE-optimal) selection of the partitioning/binning scheme. The default is bins=c(0,0), which corresponds to piecewise constant (canonical binscatter).
binspos	position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options is "es" for evenly-spaced binning.
binsmethod	method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.
nbinsrot	initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.
simsgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum (infimum or Lp metric) operation needed to construct confidence bands and hypothesis testing procedures. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum (infimum or Lp metric) operator.
savegrid	If true, a data frame produced containing grid.
vce	procedure to compute the variance-covariance matrix estimator. Options are • "const" homoskedastic variance estimator. • "HC0" heteroskedasticity-robust plug-in residuals variance estimator without weights. • "HC1" heteroskedasticity-robust plug-in residuals variance estimator with hc1 weights. Default. • "HC2" heteroskedasticity-robust plug-in residuals variance estimator with hc2 weights. • "HC3" heteroskedasticity-robust plug-in residuals variance estimator with hc3 weights.
useeffn	effective sample size to be used when computing the (IMSE-optimal) number of bins. This option is useful for extrapolating the optimal number of bins to larger (or smaller) datasets than the one used to compute it.

randcut	upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.
cluster	cluster ID. Used for compute cluster-robust standard errors.
dfcheck	adjustments for minimum effective sample size checks, which take into account number of unique values of x (i.e., number of mass points), number of clusters, and degrees of freedom of the different statistical models considered. The default is dfcheck=c(20,30). See Cattaneo, Crump, Farrell and Feng (2021b) for more details.
masspoints	how mass points in x are handled. Available options: • "on" all mass point and degrees of freedom checks are implemented. Default. • "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted. • "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted. • "off" "noadjust" and "nolocalcheck" are set simultaneously. • "veryfew" forces the function to proceed as if x has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.
weights	an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. For more details, see lm.
subset	optional rule specifying a subset of observations to be used.
norotnorm	if true, a uniform density rather than normal density used for ROT selection.
numdist	number of distinct for selection. Used to speed up computation.
numclust	number of clusters for selection. Used to speed up computation.

Value

nbinsrot.poly	ROT number of bins, unregularized.
nbinsrot.regul	ROT number of bins, regularized.
nbinsrot.uknot	ROT number of bins, unique knots.
nbinsdpi	DPI number of bins.
nbinsdpi.uknot	DPI number of bins, unique knots.
imse.v.rot	variance constant in IMSE expansion, ROT selection.
imse.b.rot	bias constant in IMSE expansion, ROT selection.
imse.v.dpi	variance constant in IMSE expansion, DPI selection.
imse.b.dpi	bias constant in IMSE expansion, DPI selection.
opt	A list containing options passed to the function, as well as total sample size n, number of distinct values Ndist in x, and number of clusters Nclust.
data.grid	A data frame containing grid.

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

binsreg, binstest.

Examples

```
x <- runif(500); y <- sin(x)+rnorm(500)
est <- binsregselect(y,x)
summary(est)
```

---

binstest — *Data-Driven Nonparametric Shape Restriction and Parametric Model Specification Testing using Binscatter*

Description

binstest implements binscatter-based hypothesis testing procedures for parametric functional forms of and nonparametric shape restrictions on the regression function estimators, following the results in Cattaneo, Crump, Farrell and Feng (2021a). If the binning scheme is not set by the user, the companion function binsregselect is used to implement binscatter in a data-driven way and inference procedures are based on robust bias correction. Binned scatter plots based on different methods can be constructed using the companion functions binsreg, binsqreg or binsglm.

Usage

```
binstest(y, x, w = NULL, data = NULL, estmethod = "reg",
  family = gaussian(), quantile = NULL, deriv = 0, at = NULL,
  nolink = F, testmodel = c(3, 3), testmodelparfit = NULL,
  testmodelpoly = NULL, testshape = c(3, 3), testshapel = NULL,
  testshaper = NULL, testshape2 = NULL, lp = Inf, bins = c(0, 0),
  nbins = NULL, binspos = "qs", binsmethod = "dpi", nbinsrot = NULL,
  randcut = NULL, nsims = 500, simsgrid = 20, simsseed = NULL,
  vce = NULL, cluster = NULL, asyvar = F, dfcheck = c(20, 30),
  masspoints = "on", weights = NULL, subset = NULL, numdist = NULL,
  numclust = NULL, ...)
```

Arguments

y	outcome variable. A vector.
x	independent variable of interest. A vector.
w	control variables. A matrix, a vector or a formula.
data	an optional data frame containing variables used in the model.
estmethod	estimation method. The default is estmethod="reg" for tests based on binscatter least squares regression. Other options are "qreg" for quantile regression and "glm" for generalized linear regression. If estmethod="glm", the option family must be specified.

family	a description of the error distribution and link function to be used in the generalized linear model when estmethod="glm". (See family for details of family functions.)
quantile	the quantile to be estimated. A number strictly between 0 and 1.
deriv	derivative order of the regression function for estimation, testing and plotting. The default is deriv=0, which corresponds to the function itself.
at	value of w at which the estimated function is evaluated. The default is at="mean", which corresponds to the mean of w. Other options are: at="median" for the median of w, at="zero" for a vector of zeros. at can also be a vector of the same length as the number of columns of w (if w is a matrix) or a data frame containing the same variables as specified in w (when data is specified). Note that when at="mean" or at="median", all factor variables (if specified) are excluded from the evaluation (set as zero).
nolink	if true, the function within the inverse link function is reported instead of the conditional mean function for the outcome.
testmodel	a vector. testmodel=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for parametric model specification testing. The default is testmodel=c(3,3), which corresponds to a cubic B-spline estimate of the regression function of interest for testing against the fitting from a parametric model specification.
testmodelparfit	a data frame or matrix which contains the evaluation grid and fitted values of the model(s) to be tested against. The column contains a series of evaluation points at which the binscatter model and the parametric model of interest are compared with each other. Each parametric model is represented by other columns, which must contain the fitted values at the corresponding evaluation points.
testmodelpoly	degree of a global polynomial model to be tested against.
testshape	a vector. testshape=c(p,s) sets a piecewise polynomial of degree p with s smoothness constraints for nonparametric shape restriction testing. The default is testshape=c(3,3), which corresponds to a cubic B-spline estimate of the regression function of interest for one-sided or two-sided testing.
testshapel	a vector of null boundary values for hypothesis testing. Each number a in the vector corresponds to one boundary of a one-sided hypothesis test to the left of the form H0: sup_x mu(x)<=a.
testshaper	a vector of null boundary values for hypothesis testing. Each number a in the vector corresponds to one boundary of a one-sided hypothesis test to the right of the form H0: inf_x mu(x)>=a.
testshape2	a vector of null boundary values for hypothesis testing. Each number a in the vector corresponds to one boundary of a two-sided hypothesis test ofthe form H0: sup_x |mu(x)-a|=0.
lp	an Lp metric used for (two-sided) parametric model specification testing and/or shape restriction testing. The default is lp=Inf, which corresponds to the sup-norm of the t-statistic. Other options are lp=q for a positive integer q.
bins	Degree and smoothness for bin selection.
nbins	number of bins for partitioning/binning of x. If not specified, the number of bins is selected via the companion function binsregselect in a data-driven, optimal way whenever possible.

binspos	position of binning knots. The default is binspos="qs", which corresponds to quantile-spaced binning (canonical binscatter). The other options are "es" for evenly-spaced binning, or a vector for manual specification of the positions of inner knots (which must be within the range of x).
binsmethod	method for data-driven selection of the number of bins. The default is binsmethod="dpi", which corresponds to the IMSE-optimal direct plug-in rule. The other option is: "rot" for rule of thumb implementation.
nbinsrot	initial number of bins value used to construct the DPI number of bins selector. If not specified, the data-driven ROT selector is used instead.
randcut	upper bound on a uniformly distributed variable used to draw a subsample for bins selection. Observations for which runif()<=# are used. # must be between 0 and 1.
nsims	number of random draws for hypothesis testing. The default is nsims=500, which corresponds to 500 draws from a standard Gaussian random vector of size [(p+1)*J-(J-1)*s].
simsgrid	number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum (infimum or Lp metric) operation needed to construct hypothesis testing procedures. The default is simsgrid=20, which corresponds to 20 evenly-spaced evaluation points within each bin for approximating the supremum (infimum or Lp metric) operator.
simsseed	seed for simulation.
vce	procedure to compute the variance-covariance matrix estimator. For least squares regression and generalized linear regression, the allowed options are the same as that for binsreg or binsqreg. For quantile regression, the allowed options are the same as that for binsqreg.
cluster	cluster ID. Used for compute cluster-robust standard errors.
asyvar	If true, the standard error of the nonparametric component is computed and the uncertainty related to control variables is omitted. Default is asyvar=FALSE, that is, the uncertainty related to control variables is taken into account.
dfcheck	adjustments for minimum effective sample size checks, which take into account number of unique values of x (i.e., number of mass points), number of clusters, and degrees of freedom of the different stat models considered. The default is dfcheck=c(20,30). See Cattaneo, Crump, Farrell and Feng (2021b) for more details.
masspoints	how mass points in x are handled. Available options: • "on" all mass point and degrees of freedom checks are implemented. Default. • "noadjust" mass point checks and the corresponding effective sample size adjustments are omitted. • "nolocalcheck" within-bin mass point and degrees of freedom checks are omitted. • "off" "noadjust" and "nolocalcheck" are set simultaneously. • "veryfew" forces the function to proceed as if x has only a few number of mass points (i.e., distinct values). In other words, forces the function to proceed as if the mass point and degrees of freedom checks were failed.
weights	an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. For more details, see lm.
subset	optional rule specifying a subset of observations to be used.

numdist	Number of distinct for selection. Used to speed up computation.
numclust	Number of clusters for selection. Used to speed up computation.
...	optional arguments to control bootstrapping if estmethod="qreg" and vce="boot". See boot.rq.

Value

testshapeL	Results for testshapel, including: testvalL, null boundary values; stat.shapeL, test statistics; and pval.shapeL, p-value.
testshapeR	Results for testshaper, including: testvalR, null boundary values; stat.shapeR, test statistics; and pval.shapeR, p-value.
testshape2	Results for testshape2, including: testval2, null boundary values; stat.shape2, test statistics; and pval.shape2, p-value.
testpoly	Results for testmodelpoly, including: testpoly, the degree of global polynomial; stat.poly, test statistic; pval.poly, p-value.
testmodel	Results for testmodelparfit, including: stat.model, test statistics; pval.model, p-values.
opt	A list containing options passed to the function, as well as total sample size n, number of distinct values Ndist in x, number of clusters Nclust, and number of bins nbins.

Author(s)

Matias D. Cattaneo, Princeton University, Princeton, NJ. <cattaneo@princeton.edu>.

Richard K. Crump, Federal Reserve Bank of New York, New York, NY. <richard.crump@ny.frb.org>.

Max H. Farrell, University of Chicago, Chicago, IL. <max.farrell@chicagobooth.edu>.

Yingjie Feng (maintainer), Tsinghua University, Beijing, China. <fengyingjiepku@gmail.com>.

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a: On Binscatter. Working Paper.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b: Binscatter Regressions. Working Paper.

See Also

binsreg, binsregselect.

Examples

```
x <- runif(500); y <- sin(x)+rnorm(500)
est <- binstest(y,x, testmodelpoly=1)
summary(est)
```

Index

_PACKAGE *(*binsreg-package*)*, 2

binsglm, *2*, 2, *7*, *23*
binspwc, *2*, 7
binsqreg, *2*, *7*, *9*, 11, *23*, *25*
binsreg, *2*, *7*, *9*, 16, *23*, *25*, *26*
binsreg-package, 2
binsregselect, *2*, *3*, *5*, *7*, *9–11*, *13*, *15*, *16*, *18*, *20*, 20, *23*, *26*
binstest, *2*, *3*, *7*, *10*, *11*, *15*, *16*, *20*, *23*, 23
boot.rq, *10*, *14*, *26*

family, *3*, *8*, *24*
formula, *3*, *8*, *11*, *16*, *21*, *23*

lm, *6*, *10*, *14*, *19*, *22*, *25*

summary.rq, *13*