Statistics/Data Analysis

help binsregselect

Title

binsregselect -- Data-driven IMSE-Optimal Partitioning/Binning Selection for Binscatter.

Syntax

binsregselect *depvar* *indvar* [*covars*] [*if*] [*in*] [*weight*] [, **deriv(***v***)** **bins(***p s***)** **binspos(***position***)** **binsmethod(***method***)** **nbinsrot(***#***)** **simsgrid(***#***)** **savegrid(***filename***)** **replace** **dfcheck(***n1 n2***)** **masspoints(***masspointsoption***)** **vce(***vcetype***)** **useeffn(***#***)**]

where *depvar* is the dependent variable, *indvar* is the independent variable for binning, and *covars* are other covariates to be controlled for.

p, s and v are integers satisfying 0 <= s,v <= p.

fweights, **aweight**s and **pweight**s are allowed; see weight.

Description

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

Options

Estimand

deriv(*v***)** specifies the derivative order of the regression function for estimation, testing and plotting. The default is **deriv(0)**, which corresponds to the function itself.

Partitioning/Binning Selection

bins(*p s***)** sets 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(0 0)**, which corresponds to piecewise constant (canonical binscatter).

binspos(*position***)** specifies the position of binning knots. The default is **binspos(qs)**, which corresponds to quantile-spaced binning (canonical binscatter). Other option is **es** for evenly-spaced binning.

binsmethod(*method***)** specifies the 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(*#***)** specifies an 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.

Evaluation Points Grid Generation

simsgrid(*#***)** specifies the number of evaluation points of an evenly-spaced grid within each bin used for evaluation of the supremum (or infimum) 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 (or infimum) operator.

savegrid(*filename***)** specifies a filename for storing the simulation grid of evaluation points. It contains the following variables: *indvar*, which is a sequence of evaluation points used in approximation; all control variables in *covars*, which take values of zero for prediction purpose; *binsreg_isknot*, indicating whether the evaluation point is an inner knot; and *binsreg_bin*, indicating which bin the evaluation point belongs to.

replace overwrites the existing file when saving the grid.

Mass Points and Degrees of Freedom

dfcheck(*n1 n2***)** sets cutoff values for minimum effective sample size checks, which take into account the number of unique values of *indvar* (i.e., adjusting for the number of mass points), number of clusters, and degrees of freedom of the different statistical models considered. The default is **dfcheck(20 30)**. See Cattaneo, Crump, Farrell and Feng (2019b) for more details.

masspoints(*masspointsoption***)** specifies how mass points in *indvar* are handled. By default, all mass point and degrees of freedom checks are implemented. Available options:
masspoints(*noadjust***)** omits mass point checks and the corresponding effective sample size adjustments.
masspoints(*nolocalcheck***)** omits within-bin mass point and degrees of freedom checks.
masspoints(*off***)** sets **masspoints(***noadjust***)** and **masspoints(***nolocalcheck***)** simultaneously.
masspoints(*veryfew***)** forces the command to proceed as if *indvar* has only a few number of mass points (i.e., distinct values). In other words, forces the command to proceed as if the mass point and degrees of freedom checks were failed.

Other Options

vce(*vcetype***)** specifies the *vcetype* for variance estimation used by the command regress. The default is **vce(robust)**.

useeffn(*#***)** specifies the 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.

Examples

Select IMSE-optimal number of bins using DPI-procedure
. binsregselect y x w

Stored results

Scalars

e(N)	number of observations
e(Ndist)	number of distince values
e(Nclust)	number of clusters
e(p)	degree of piecewise polynomial
e(s)	smoothness of piecewise polynomial
e(deriv)	order of derivative
e(nbinsrot_poly)	ROT number of bins, unregularized
e(nbinsrot_regul)	ROT number of bins, regularized or user-specified
e(nbinsrot_uknot)	ROT number of bins, unique knots
e(nbinsdpi)	DPI number of bins
e(nbinsdpi_uknot)	DPI number of bins, unique knots

Matrices

e(knot)	numlist of knots

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2019a. On Binscatter. *arXiv:1902.09608*.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2019b. Binscatter Regressions. *arXiv:1902.09615*.

Authors

Matias D. Cattaneo, University of Michigan, Ann Arbor, MI. cattaneo@umich.edu.

Richard K. Crump, Federal Reserve Band 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, University of Michigan, Ann Arbor, MI. yjfeng@umich.edu.