Statistics/Data Analysis

help binspwc

Title

binspwc —— Data-Driven Nonparametric Pairwise Group Comparison using Binscatter.

Syntax

binspwc *depvar indvar* [*covars*] [*if*] [*in*] [*weight*] [, **estmethod(***cmdname***) deriv(***v***)**
by(*varname***)**
pwc(*p s***) testtype(***type***) lp(***metric***)**
bins(*p s***) bynbins(***numlist***) binspos(***position***) binsmethod(***method***)**
nbinsrot(*#***) samebinsby**
nsims(*#***) simsgrid(***#***) simsseed(***seed***)**
dfcheck(*n1 n2***) masspoints(***masspointsoption***)**
vce(*vcetype***)**]

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, which can take different values in each case.

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

Description

binspwc implements binscatter-based 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 command binsregselect is used to implement binscatter in a data-driven (optimal) way and inference procedures are based on robust bias correction. Binned scatter plots based on different models can be constructed using the companion commands binsreg, binsqreg, binslogit and binsprobit.

A detailed introduction to this command is given in Cattaneo, Crump, Farrell and Feng (2021b). A companion R package with the same capabilities is available (see website below).

Companion commands: binsreg for binscatter least squares regression with robust inference procedures and plots, binsqreg for binscatter quantile regression with robust inference procedures and plots, binslogit for binscatter logit estimation with robust inference procedures and plots, binsprobit for binscatter probit estimation with robust inference procedures and plots, and binsregselect data-driven (optimal) binning selection.

Related Stata and R packages are available in the following website:

https://nppackages.github.io/

Options

Estimand

estmethod(*cmdname***)** specifies the binscatter model. The default is **estmethod(reg)**, which corresponds to the binscatter least squares regression. Other options are: **estmethod(qreg #)** for binscatter quantile regression where # is the quantile to be estimated, **estmethod(logit)** for binscatter logistic regression and **estmethod(probit)** for binscatter probit regression.

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.

by(*varname***)** specifies the variable containing the group indicator to perform subgroup analysis; both numeric and string variables are supported. When **by(***varname***)** is specified, **binspwc** implements estimation by each subgroup separately and then conduct *all* pairwise comparison tests. 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.

Pairwise Group Comparison Testing

pwc(*p s***)** sets a piecewise polynomial of degree *p* with *s* smoothness constraints for pairwise group comparison. The default is **pwc(3 3)**, which corresponds to a cubic B-spline estimate of the function of interest for each group.

testtype(*type***)** specifies the type of pairwise comparison test. The default is **testtype(***2***)**, which corresponds to a two-sided test of the form H0: *mu_1(x)=mu_2(x)*. Other options are: **testtype(***1***)** for the one-sided test of the form H0: *mu_1(x)<=mu_2(x)* and **testtype(***r***)** for the one-sided test of the form H0: *mu_1(x)>=mu_2(x)*.

lp(*metric***)** specifies a Lp metric used for a (two-sided) test for the difference between two groups. The default is **lp(inf)**, which corresponds to the sup-norm. Other options are Lp(q) for a positive integer q.

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).

bynbins(*numlist***)** sets a numlist of numbers of bins for partitioning/binning of *indvar*, which is applied to the binscatter estimation for each group. The ordering of the group follows the result of the tabulate. If a single number of bins is specified, it applies to the estimation for all groups. If not specified, the number of bins is selected via the companion command binsregselect in a data-driven, optimal way whenever possible.

binspos(*position***)** specifies the position of binning knots. The default is **binspos(qs)**, which corresponds to quantile-spaced binning (canonical binscatter). Other options are: **es** for evenly-spaced binning, or a numlist for manual specification of the positions of inner knots (which must be within the range of *indvar*).

binsmethod(*method***)** specifies the method for data-driven selection of the number of bins via the companion command binsregselect. 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.

samebinsby forces a common partitioning/binning structure across all subgroups specified by the option **by()**. 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.

Simulation

nsims(*#***)** specifies the number of random draws for constructing confidence bands and 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(#) 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.

simsseed(#) sets the seed for simulations.

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 commands regress, logit or qreg. The default is **vce(robust)**.

Examples

Test the difference between two groups

```
. binspwc y x w, by(t)
```

Stored results

Scalars

e(N)	number of observations
e(p)	degree of polynomial for bin selection
e(s)	smoothness of polynomial for bin selection
e(pwc_p)	degree of polynomial for testing
e(pwc_s)	smoothnes of polynomial for testing

Locals

e(byvalue)	name of groups found in **by()**

Matrices

e(N_by)	number of observations for each group
e(Ndist_by)	number of distinct values for each group
e(Nclust_by)	number of clusters for each group
e(nbins_by)	number of bins for each group
e(stat)	test statistics for all pairwise comparisons
e(pval)	p values for all pairwise comparisons

References

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021a. On Binscatter. *arXiv:1902.09608*.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng. 2021b. Binscatter Regressions. *arXiv:1902.09615*.

Authors

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