Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
DOI: 10.11159/icsta22.005

On Binscatter*

Matias D. Cattaneo[1], Richard K. Crump, Max H. Farrell, Yingjie Feng
[1]Princeton University, USA

Abstract

Binscatter, or a binned scatter plot, is a very popular tool in applied microeconomics. It provides a flexible, yet parsimonious way of visualizing and summarizing mean, quantile, and other nonparametric regression functions in large data sets. It is also often used for informal evaluation of substantive hypotheses such as linearity or monotonicity of the unknown function. This paper presents a foundational econometric analysis of binscatter, offering an array of theoretical and practical results that aid both understanding current practices (i.e., their validity or lack thereof) as well as guiding future applications. In particular, we highlight important methodological problems related to covariate adjustment methods used in current practice, and provide a simple, valid approach. Our results include a principled choice for the number of bins, confidence intervals and bands, hypothesis tests for parametric and shape restrictions for mean, quantile, and other functions of interest, among other new methods, all applicable to canonical binscatter as well as to nonlinear, higher-order polynomial, smoothness-restricted and covariateadjusted extensions thereof. Companion general-purpose software packages for Python, R, and Stata are provided. From a technical perspective, we present novel theoretical results for possibly nonlinear semi-parametric partitioning-based series estimation with random partitions that are of independent interest.

Keywords: binned scatter plot, regressogram, piecewise polynomials, splines, partitioning estimators, nonparametric regression, nonparametric quantile regression, nonparametric nonlinear semilinear quasi-maximum likelihood, robust bias correction, uniform inference, binning selection.