On Binscatter*

Matias D. Cattaneo† Richard K. Crump‡ Max H. Farrell§ Yingjie Feng¶

August 6, 2021

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 covariate-adjusted 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.

*We specially thank Jonah Rockoff and Ryan Santos for detailed, invaluable feedback on this project. We also thank Raj Chetty, Michael Droste, John Friedman, Andreas Fuster, Paul Goldsmith-Pinkham, Andrew Haughwout, Guido Imbens, David Lucca, Xinwei Ma, Emily Oster, Jesse Rothstein, Jesse Shapiro, Rocio Titiunik, Seth Zimmerman, Eric Zwick, and seminar participants at various seminars, workshops and conferences for helpful comments and discussions. Oliver Kim and Shahzaib Safi provided excellent research assistance. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805 and SES-2019432. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. Companion general-purpose software and replication files are available at https://nppackages.github.io/binsreg/.

†Department of Operations Research and Financial Engineering, Princeton University.

‡Capital Markets Function, Federal Reserve Bank of New York.

§Booth School of Business, University of Chicago.

¶School of Economics and Management, Tsinghua University.

1 Introduction

Binscatter, or a binned scatter plot, is a flexible, yet parsimonious way of visualizing and summarizing regression functions in large data sets. This methodology is also often used in social and behavioral sciences for informal (visual) evaluation of substantive hypotheses about shape features of the unknown regression function such as linearity, monotonicity, or concavity, as well as to estimate heterogeneous treatment effects and heuristically quantify uncertainty in estimation of mean, quantile and other regression functions. See Chetty and Szeidl (2006, Figure 1) for one of the first explicit appearances of a binned scatter plot in the applied microeconomics literature, and see Chetty, Looney, and Kroft (2009), Chetty, Friedman, Olsen, and Pistaferri (2011), Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011), and Chetty, Friedman, and Rockoff (2014), among many others, for early influential papers using binscatter methods. See Starr and Goldfarb (2020) for an overview of binscatter.

Binscatter methods have gained immense popularity among empirical researchers and policy makers, and are by now a leading component of the standard applied microeconomics toolkit. However, the remarkable proliferation of binscatter in empirical work has not been accompanied by the development of econometric results guiding its correct use and providing valid estimation and inference procedures. Current practice employing canonical binscatter is ad-hoc and undisciplined, which not only hampers replicability across studies, but also can lead to incorrect empirical conclusions.

This paper presents a foundational study of binscatter. We provide several theoretical and practical results that aid both in understanding the validity (or lack thereof) of current practices, and in offering principled guidance for future applications. To give a systematic analysis of binscatter, we first recast it as a nonparametric estimator of a regression function employing (possibly restricted) piecewise approximations in a semi-linear (possibly nonlinear) regression setting. Thus, our first contribution is to introduce an econometrics framework to understand and analyze binscatter formally, covering an array of parameters of interest such as mean, quantile, and other nonlinear regression functions. This framework allows us not only to obtain theoretical and practical results for canonical binscatter methods, but also to propose new methods delivering more flexible and smooth approximations of the regression function of interest, while still retaining the core features

of binscatter. These methodological enhancements are particularly well suited to improve graphical presentation of estimated regression functions and confidence bands, and for formal testing of substantive hypotheses about regression functions.

Furthermore, using our econometric framework, we discuss precisely parameters of interest for applied work and, as a by-product, we highlight important methodological and theoretical problems with the covariate adjustment methods commonly employed in practice. To be more specific, we discuss the detrimental effects of the widespread practice of first "residualizing out" additional covariates and then constructing a canonical binscatter. To solve this problem, we propose a new semi-parametric approach based on semi-linear regression, which is more generally valid and principled, and show how our proposed covariate-adjusted binscatter circumvents the problems introduced by the commonly used covariate-residualized canonical binscatter method. See Remark 1 below for more details.

Our proposed econometric framework is then used to offer several new technical and methodological results for binscatter applications. All our results not only give principled guidance for current empirical practice using binscatter, but also offer new methods encoding informal (visual) assessments commonly found in empirical papers. First, we develop a valid and optimal selector of the number of bins for binscatter implementation, which is constructed based on an integrated mean square error approximation. Our proposed selector intuitively balances the bias and variance of binscatter, and can contrast sharply with ad-hoc choices encountered in practice: always using 10 or 20 bins, which can lead to substantial misspecification bias in applications. Second, we provide valid confidence intervals and confidence bands to quantify uncertainty in the estimation of mean, quantile, and other regression functional parameters of interest. These can be applied, for example, to conduct inference on effects of continuous treatments. Third, we develop valid hypothesis testing procedures for parametric specification and shape restrictions of the unknown regression function, as well as for comparisons across groups. The latter can be used to study treatment effect heterogeneity in experimental and non-experimental settings. We provide general-purpose `Python`, `R` and `Stata` software packages implementing all our methodological results. For further details and numerical illustrations, see our companion software article (Cattaneo, Crump, Farrell, and Feng, 2021) and the software repository (`https://nppackages.github.io/binsreg/`).

1.1 Connections to the Literature

Applying binning to regression problems dates back at least to the regressogram of Tukey (1961), and in nonparametric regression more broadly it is known as partitioning regression (Györfi, Kohler, Krzyżak, and Walk, 2002). Binning has been applied in many other areas due to its intuitive appeal and ease of implementation: in density estimation as the classical histogram; in program evaluation for estimation via subclassification (Cochran, 1968), and for visualization in regression discontinuity designs (Calonico, Cattaneo, and Titiunik, 2015) and bunching designs (Kleven, 2016); in empirical finance as portfolio sorting (Fama, 1976); and in machine learning it is at the heart of regression trees (Hastie, Tibshirani, and Friedman, 2009). Each of these applications have distinct goals and data features, thereby requiring separate technical analyses. Some of our novel theoretical results in this paper could be used to also analyze these other applications, but not without substantial modifications.

Our technical work is most closely related to the literature on nonparametric linear and nonlinear series regression estimation (Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2019; Cattaneo, Farrell, and Feng, 2020, and references therein). However, available technical results have two major shortcomings: (i) not allowing for random partitioning (e.g., quantile-spaced binning generated by estimated quantiles), and (ii) imposing overly stringent conditions on tuning parameters that rule out estimation and inference validity for canonical and other binscatter methods (even with non-random binning). As a consequence, our work necessarily requires new technical results concerning non-/semi-parametric linear and nonlinear partitioning-based series estimation on quantile-spaced random partitions, which we provide in this paper and may be of independent interest.

More specifically, to address the first technical shortcoming, we obtain new theoretical results for quantile-spaced random partitions underlying canonical binscatter and its generalizations, which substantially extend Nobel (1996), together with a general characterization of a linear map linking piecewise polynomials and B-spline bases, properly accounting for quantile-spaced binning. To address the second technical shortcoming, we develop two distinct theoretical results. First, we establish a new strong approximation approach for the supremum of the t-statistic process, building on ideas related to uniform distributional approximations of the supremum of a stochastic process

in Chernozhukov, Chetverikov, and Kato (2014a,b). Second, we establish a novel Bahadur representation for semi-parametric nonlinear partitioning-based series regression estimation, substantially improving on Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019). These results were previously unavailable in the literature but are needed to analyze popular implementations of binscatter and generalizations thereof. Section SA-1.1 of the supplemental appendix provides a full account of our technical contributions, which we also discuss throughout the manuscript.

All our methodological and theoretical contributions are implemented in our newly-created software packages for `Python`, `R` and `Stata`. We offer several important improvements over current software implementations `binscatter` (Stepner, 2017) and `binscatter2` (Droste, 2019), which only implement canonical binscatter and covariate-residualized canonical binscatter. First, we consider estimation and inference for mean, quantile, and other regression functions of interest, as well as transformations thereof (e.g., derivatives for partial effects). Second, we implement covariate adjustments avoiding naive residualization, thus giving principled, interpretable, and valid methods for practice. Third, we implement fully data-driven selections of the number of bins in a principled and objective way, reflecting the features of the underlying data. Fourth, we implement valid (pointwise and uniform) distributional approximations leading to confidence intervals, confidence bands, and hypothesis tests for a wide range of parametric specifications and nonparametric shape restrictions, as well as group-wise comparisons. Fifth, we allow polynomial fits within bins and smoothness restrictions across bins for binscatter in both linear and nonlinear settings, which, rather than being only a technical generalization, is required for implementing nonparametric shape-related hypothesis tests. Cattaneo, Crump, Farrell, and Feng (2021) discusses all the details concerning our accompanying software and further illustrates its use.

1.2 Paper Organization

Section 2 introduces and formalizes least squares binscatter, starting with its canonical form and then incorporating within-bin higher-order polynomial fitting, across-bins smoothness restrictions, and semi-linear covariate-adjustment. That section discusses parameters of interest in the context of least squares binscatter and formalizes the issues with the current practice of first regressing out covariates (Remark 1). Section 3 gives our formal results for least squares binscatter, including estimation, the optimal number of bins, uncertainty quantification (confidence intervals and bands),

inference (shape, specification and multi-sample hypothesis testing), and graphical presentation methods. Section 4 outlines the generalization of all our results to nonlinear binscatter (e.g., Probit, Logit, or quantile regression), focusing in particular on parameters of interest and related estimation and inference methods. Finally, Section 5 concludes. The supplemental appendix collects more general theoretical results, including full, formal details for Section 4, all proofs, and other methodological and numerical results. The paper references the supplement throughout; all sections starting with SA- are located in the supplemental appendix.

We illustrate our methods throughout using synthetic data based on the American Community Survey. In order to have full control on different features of the econometric model, we simulate a dataset based on a data generating process constructed using the real survey dataset. The supplemental appendix (Section SA-6) details how the simulated data was constructed, and also includes several empirical illustrations using real data.

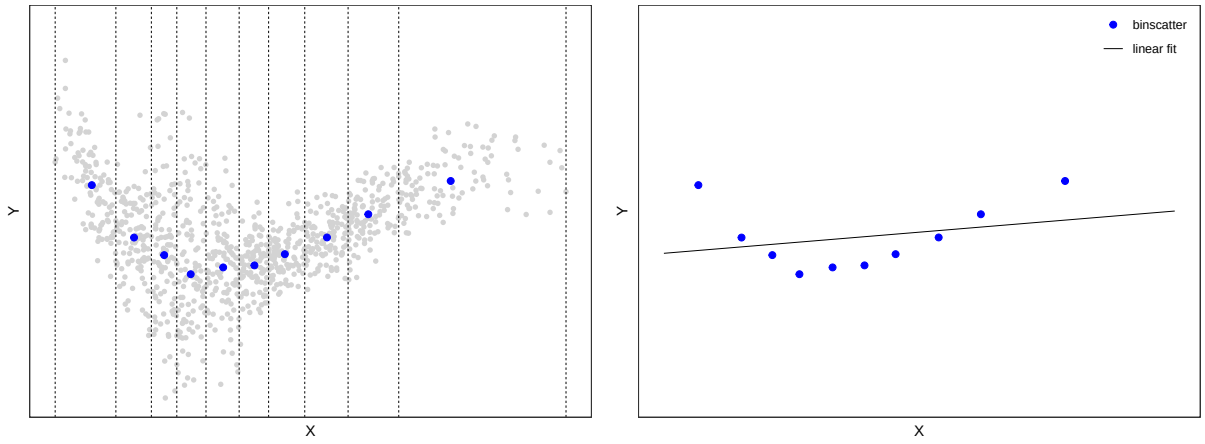
2 Formalizing Least Squares Binscatter

Suppose $(y_i, x_i, \mathbf{w}_i')'$, $i = 1, 2, \ldots, n$, is a random sample, where y_i is a scalar response variable, x_i is the scalar independent variable of interest, and $\mathbf{w}_i$ is a d-dimensional vector of additional covariates. Canonical binscatter aims to approximate the conditional expectation $\mathbb{E}[y_i|x_i]$, without controlling for $\mathbf{w}_i$, and is constructed employing a quantile-spaced, disjoint partitioning of the support of x_i based on the observed data. To be precise, J disjoint intervals are computed using the empirical quantiles of x_i, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, where

$$\widehat{\mathcal{B}}_j = \begin{cases} \left[x_{(1)}, x_{(\lfloor n/J \rfloor)}\right) & \text{if } j = 1 \\ \left[x_{(\lfloor n(j-1)/J \rfloor)}, x_{(\lfloor nj/J \rfloor)}\right) & \text{if } j = 2, 3, \ldots, J-1 \\ \left[x_{(\lfloor n(J-1)/J \rfloor)}, x_{(n)}\right] & \text{if } j = J \end{cases},$$

$x_{(i)}$ denotes the i-th order statistic of the sample $(x_1, x_2, \ldots, x_n)$, $\lfloor \cdot \rfloor$ is the floor operator, and $J < n$. Each estimated bin $\widehat{\mathcal{B}}_j$ contains roughly the same number of observations $N_j = \sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i)$, where $\mathbb{1}_{\mathcal{A}}(x) = \mathbb{1}(x \in \mathcal{A})$, with $\mathbb{1}(\cdot)$ denoting the indicator function. It follows that units are binned according to their rank along the x_i dimension.

Figure 1: The Basic Construction of a Binned Scatter Plot.



(a) Scatter and Canonical Binscatter Plots (b) Canonical Binscatter and Global Linear Fit

Notes. This figure shows the intermediate and final steps to construct a canonical binscatter plot. In plot (a) the data is divided into $J = 10$ bins according to the observed x_i. Within each bin a single blue dot is plotted at the mean of y_i for observations falling in the bin. The final plot (b) consists of only these J dots, and the fit from a least squares linear regression of y_i on x_i. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey dots represent the realized data. The sample size is $n = 1,000$.

Given the quantile-spaced random partitioning scheme $\widehat{\Delta}$, for a choice of number of bins J, the *canonical* binscatter estimator is

$$\widehat{\mathbb{E}}[y_i|x_i] = \widehat{\mathbf{b}}(x_i)'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta})^2, \tag{2.1}$$

where $\widehat{\mathbf{b}}(x) = [\mathbb{1}_{\widehat{\mathcal{B}}_1}(x), \mathbb{1}_{\widehat{\mathcal{B}}_2}(x), \cdots, \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)]'$ is the canonical binscatter basis given by a J-dimensional vector of orthogonal indicator variables, that is, the j-th component of $\widehat{\mathbf{b}}(x)$ records whether the evaluation point x belongs to the j-th bin in the partition $\widehat{\Delta}$. Therefore, canonical binscatter can be expressed as the collection of J sample averages of the response variable: for each bin $\widehat{\mathcal{B}}_j$, $j = 1, 2, \ldots, J$, the estimate (2.1) reduces to $\bar{y}_j = \frac{1}{N_j}\sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i) y_i$ for all points in bin $\widehat{\mathcal{B}}_j$. Empirical work employing canonical binscatter typically plots these binned sample averages along with some other estimate of the regression function $\mathbb{E}[y_i|x_i]$.

Figure 1 illustrates canonical binscatter using the synthetic dataset. Implementation requires selecting the number of bins used. Ad-hoc choices such as $J = 10$ or 20 are most predominant in practice. Then, a single dot is placed in the plot representing the mean of y_i for the observations falling in each bin, $\widehat{\mathcal{B}}_j$, $j = 1, 2, \ldots, J$. The final plot consists of only these J dots, usually depicted at the center of each quantile-spaced bin. The regression line from a linear least squares fit to the

underlying raw data is often added (solid line in Figure 1(b)). It is typical in applications to control for covariates prior to deploying binscatter methods, which requires additional care and may not lead to a valid econometric procedure (Remark 1).

Canonical binscatter can be interpreted as a nonparametric series estimator of $\mathbb{E}[y_i|x_i]$ with a Haar basis (i.e., a zero-degree piecewise polynomial or spline estimator) and tuning parameter $J \to \infty$ as $n \to \infty$. Therefore, this estimation procedure can be naturally extended to allow for a higher-order polynomial fit within bins (piecewise polynomial estimator) and smoothness restrictions across bins (spline estimator). This additional generality helps in estimating derivatives of the function of interest, reducing the smoothing bias of the estimator, and enhancing the binscatter plots, among other useful features. Furthermore, in empirical work it is often important to control for additional covariates, $\mathbf{w}_i$, which can be easily achieved using semi-parametric partially linear regression methods.

Therefore, given the quantile-spaced partitioning scheme, the p-th order polynomial, s-times continuously differentiable, covariate-adjusted least-squares *extended* binscatter estimator is

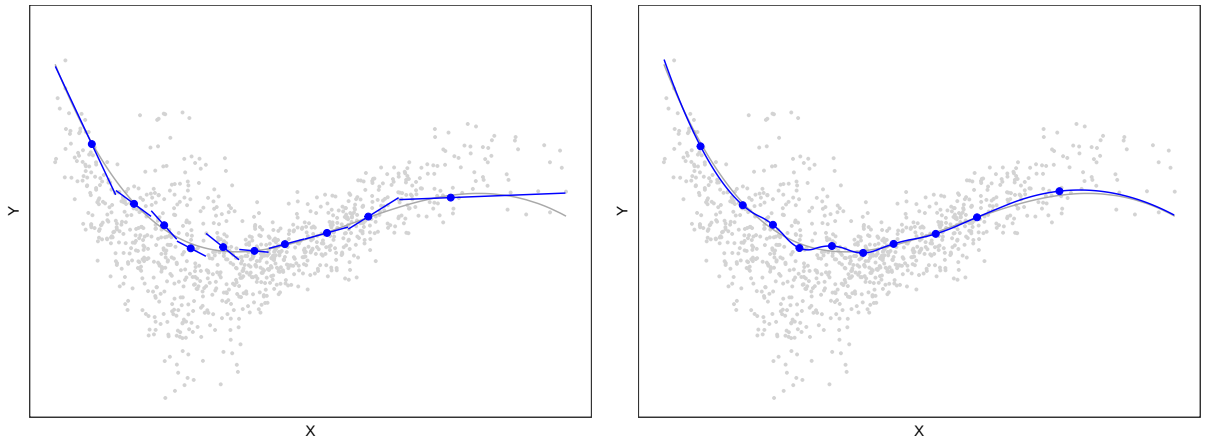
$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad 0 \leq v, s \leq p, \tag{2.2}$$

where $g^{(v)}(x) = \frac{d^v}{dx^v} g(x)$ for any function $g(\cdot)$ whenever well-defined (e.g., one-sided derivative at boundary points), and $\widehat{\mathbf{b}}_s(x) = \widehat{\mathbf{T}}_s \widehat{\mathbf{b}}(x)$ with $\widehat{\mathbf{b}}(x) = [\mathbb{1}_{\widehat{\mathcal{B}}_1}(x), \mathbb{1}_{\widehat{\mathcal{B}}_2}(x), \cdots, \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)]' \otimes [1, x, \cdots, x^p]'$ redefined to be the extended p-th order binscatter basis of dimension $(p+1)J$, and $\widehat{\mathbf{T}}_s$ is a $[(p+1)J - (J-1)s] \times (p+1)J$ matrix of linear restrictions ensuring that the $(s-1)$-th derivative of $\widehat{\mu}(x) = \widehat{\mu}^{(0)}(x)$ is continuous. See Figure 2 for a numerical illustration using the simulated data.

Enforcing smoothness for binscatter boils down to incorporating restrictions on the p-th order binscatter basis. The resulting constrained random basis, $\widehat{\mathbf{b}}_s(x)$, corresponds to a choice of spline basis for approximation, with estimated quantile-spaced knots according to the partition $\widehat{\Delta}$. In this paper, we employ $\widehat{\mathbf{T}}_s$ leading to B-splines, which tend to offer good finite sample properties. When setting $p = s = v = 0$ and excluding the covariates $\mathbf{w}_i$ in (2.2), the extended binscatter estimator reduces to the canonical binscatter, in which case $\widehat{\mu}(x)$ becomes the step function in (2.1).

When $s = 0$, the extended binscatter estimator retains the main features of canonical binscatter: estimation is conducted using only information within each (estimated) bin forming the quantile-

Figure 2: Binscatter Generalizations.



(a) Extended Binscatter: $p = 1$ and $s = 0$ (b) Extended Binscatter: $p = 2$ and $s = 2$

Notes. This figure illustrates the covariate-adjusted least-squares extended binscatter estimator. In plot (a) the binscatter estimator utilizes a linear fit within bins but is otherwise unrestricted. In plot (b) the binscatter estimator utilizes a quadratic fit within bins and imposes continuous differentiability. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey line represents the true function, $\Upsilon_{\bar{\mathbf{w}}}(x) = \mathbb{E}[y_i | x_i = x, \mathbf{w}_i = \bar{\mathbf{w}}]$, and grey dots represent the realized data. The sample size is $n = 1,000$.

spaced partition of the support of x_i. It follows that $\widehat{\mu}(x)$ is discontinuous at the boundaries of the J bins forming the partition $\widehat{\Delta}$. For some empirical analyses, both graphical and formal, researchers would prefer a smoother binscatter of $\mu_0(\cdot)$ and related functions of interest, where the fits within each bin are constrained so that the final estimator exhibits some overall smoothness over the support of x_i. For instance, it is natural to construct confidence bands or conduct hypothesis tests about shape restrictions using $s > 0$.

These extensions of canonical binscatter are exhibited in Figure 2. Increasing the polynomial order p used within bins allows for a more "flexible" local approximation within each bin, while increasing the smoothness order s forces the approximation to be smoother across bins. Thus, the users' choices p and s control flexibility and smoothness from a local and global perspectives, respectively. For example, if $p = 1$, then $s = 0$ corresponds to piecewise linear fits that could be disconnected at the bins' boundaries. This is illustrated in Figure 2(a). An example of within-bin quadratic fit ($p = 2$) with two smoothness constraints ($s = 2$) is given in Figure 2(b), which offers a continuously differentiable estimated regression function.

2.1 Assumptions

The following assumption is imposed throughout when studying the extended least squares binscatter estimator $\widehat{\mu}^{(v)}(x)$ defined in (2.2). Let $\|\cdot\|$ denote the Euclidean norm.

Assumption 1. *The sample $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is i.i.d and satisfies the model $y_i = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0 + \epsilon_i$ with $\mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0$. Further, the covariate x_i has a continuous density function $f_X(x)$ bounded away from zero on the support $\mathcal{X}$, $\mathbb{E}[\mathbb{V}[\mathbf{w}_i|x_i]] > 0$, $\sigma^2(x) = \mathbb{E}[\epsilon_i^2|x_i = x]$ is continuous and bounded away from zero, and $\mathbb{E}[\|\mathbf{w}_i\|^4|x_i = x]$, $\mathbb{E}[|\epsilon_i|^4|x_i = x]$ and $\mathbb{E}[|\epsilon_i|^2|x_i = x, \mathbf{w}_i = \mathbf{w}]$ are uniformly bounded. Finally, $\mu_0(x)$ and $\mathbb{E}[\mathbf{w}_i|x_i = x]$ are $(p+q+1)$-times continuously differentiable form some $q \geq 1$.*

This assumption is not minimal, but is nonetheless mild because it imposes mostly standard conditions from classical regression settings. However, the restriction $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$ is not innocuous: it assumes additive separability between the covariate of interest x_i and the other covariates $\mathbf{w}_i$, which may be restrictive in some settings. We impose that assumption because it follows practice closely, and simplifies the exposition, but our results remain valid under possibly misspecification of $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$, as we discuss in the following subsection. Furthermore, Section 4 studies generalized nonlinear binscatter estimation, thereby providing another relaxation of the additive separability assumption to $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0)$ for some link function $\eta(\cdot)$, in addition to covering a richer class of estimands (e.g., quantile semi-linear regression). When the covariates $\mathbf{w}_i$ are not included in the binscatter procedure, all statements involving $\mathbf{w}_i$ in Assumption 1 can be ignored or simplified by dropping $\mathbf{w}_i$. See the supplemental appendix for precise, weaker regularity conditions implying Assumption 1.

2.2 Estimands and Parameters of Interest

To assess the validity, and usefulness in practice, of covariate adjustment using least squares binscatter methods, it is important to first define the parameter of interest. Unfortunately, the empirical literature employing binscatter methods is usually imprecise on this point. A natural candidate, and one we believe is most often (implicitly) sought for, is the partial mean effect

$$\Upsilon_{\mathbf{w}}^{(v)}(x) = \frac{\partial^v}{\partial x^v}\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}], \tag{2.3}$$

where $\mathbf{w}$ denotes some user-selected evaluation point. For example, $\mathbf{w} = \mathbf{0}$, $\mathbf{w} = \mathbb{E}[\mathbf{w}_i]$ or $\mathbf{w} = \text{median}(\mathbf{w}_i)$, with $\mathbf{0}$ denoting a vector of zeros and $\mathbf{w} = \text{median}(\mathbf{w}_i)$ denoting the population median of each component in $\mathbf{w}_i$. For discrete variables in $\mathbf{w}$ it is natural to set those components to some base category (e.g., zero for binary variables), while for continuous variables it may be preferable to set those components at some other place of their distribution (e.g., mean or some quantile).

A plug-in estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$ using least squares binscatter methods (2.2) is

$$\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}, \qquad \text{and} \qquad \widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x) = \widehat{\mu}^{(v)}(x) \quad \text{if } v > 1, \tag{2.4}$$

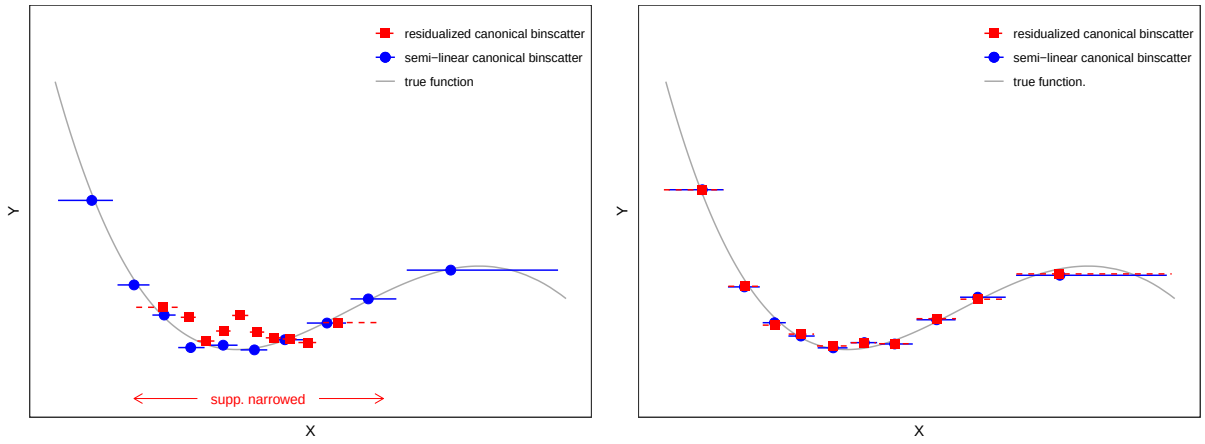
where $\widehat{\mathbf{w}}$ is a consistent estimator of the desired evaluation point $\mathbf{w}$ in (2.3). If $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$, as in Assumption 1, then $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ will be a consistent estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$. This corresponds to the case of correct specification of $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$, which we impose throughout to avoid untidy notation related to (non-random) misspecification errors. In fact, all our results easily extend to the case where $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \mu_0(x_i) + \gamma_0(\mathbf{w}_i)$, for some unknown function $\gamma_0(\cdot)$. In that context, the additional covariates $\mathbf{w}_i$ entering (2.2) would be redefined as increasing-dimension basis functions or transformations constructed out of some underlying fixed-dimension underlying covariates, and able to approximate well the unknown function $\gamma_0(\cdot)$. See Section SA-2 for more details. Another extension we could accommodate is the presence of interaction effects between x_i and $\mathbf{w}_i$ as in

$$\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \zeta_0(x_i) + \zeta_1(x_i)\mathbf{w}_i'\delta + \mathbf{w}_i'\gamma,$$

with appropriate normalizations for identifiability. When $\mathbf{w}_i$ is a scalar binary variable, this model simplifies to the two-sample setup we discuss in Section 3.6. Least squares binscatter methods (2.2) can also be useful when $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$ is left unrestricted, in which case $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ is a misspecified estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$ in general. Nevertheless, our approach to covariate adjustment retains the usual interpretation of standard semi-linear models, where the unknown "long" conditional expectation $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$ is approximated by the closest model of the form $\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$ in a mean square error sense. See, e.g., Angrist and Pischke (2008) for further discussion on the interpretation and use of misspecified (semi-)linear least squares approximations in applied work.

Remark 1 (Residualized Canonical Binscatter)**.** Widespread empirical practice for covariate ad-

Figure 3: Comparison of Covariate Adjustment Approaches.



(a) x_i correlated with $\mathbf{w}_i$ (b) x_i independent of $\mathbf{w}_i$

Notes. This figure compares two approaches to covariate adjustment for binned scatter plots: semi-linear covariate adjustment and residualized covariate adjustment. Plot (a) compares the two approaches when there is non-zero correlation between x_i and the other covariates, $\mathbf{w}_i$, and demonstrates the biases introduced by naive residualization. Plot (b) compares the two approaches when x_i and the other covariates, $\mathbf{w}_i$, are independent. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The sample size is $n = 1,000$.

justment of binscatter proceeds by first regressing out the covariates $\mathbf{w}_i$, and then applying canonical binscatter (2.1) on the residualized variables of interest. This approach is heuristically motivated by the usual Frisch–Waugh–Lovell approach to "regressing/partialling out" other covariates in linear regression settings, and is the default implementation of covariate adjustment in commonly binscatter software used in practice. Under mild assumptions, this covariate-residualized binscatter approximates the conditional expectation $\mathbb{E}[y_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}|x_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}]$, where $\widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}$ and $\widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}$ can be interpreted as the best (in mean square) linear approximation to $\mathbb{E}[y_i|\mathbf{w}_i]$ and $\mathbb{E}[x_i|\mathbf{w}_i]$, respectively, with $\widetilde{\mathbf{w}}_i = (1, \mathbf{w}_i')'$. The conditional expectation $\mathbb{E}[y_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}|x_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}]$ is rather difficult to interpret. Residualized binscatter does not consistently estimate $\mathbb{E}[y_i|x_i]$ in general, and tends to bias the function estimator towards a linear fit, giving the visual appearance of linearity of $\mu_0(x)$ and $\mathbb{E}[y_i|x_i]$. ⌟

The previous remark makes clear that if $\mu_0(x)$ is not linear, then the standard Frisch–Waugh–Lovell logic does not apply: one cannot estimate (or binned scatter plot) the functions $\mathbb{E}[y_i|x_i]$ or $\mu_0(x)$ using the residuals from least squares regressions of y_i on $\mathbf{w}_i$ and x_i on $\mathbf{w}_i$. This highlights an important methodological issue with most current applications using covariates with binscatter, which rely on covariate-residualized canonical binscatter (see Remark 1). The latter approach,

which differs from our proposed method for covariate-adjustment, can lead to different empirical findings, in particular obscuring the presence of nonlinearities in the functions of interest (e.g., $\mathbb{E}[y_i|x_i]$, $\mu_0(x)$, or $\Upsilon_{\mathbf{w}}(x)$).

We illustrate the issue of covariate adjustment numerically in Figure 3, using simulated data. The true regression function is depicted in solid grey, while the two approaches to covariate-adjusted binscatter are presented in solid blue circles (ours) and solid red squares (residualized binscatter). As Figure 3 clearly indicates, the two approaches for covariate adjustment lead to quite different results if x_i and $\mathbf{w}_i$ are correlated. Residualized binscatter is unable to correctly approximate the true function of interest $\Upsilon_{\mathbf{w}}(x)$, say, while our semi-linear covariate-adjustment approach works well.

The least squares binscatter estimator $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ not only extends canonical binscatter by allowing for v-th order derivative estimation, s-th order across-bin smoothness restrictions, and p-th order within-bin polynomial approximations, but also incorporates covariate adjustment in a principled, interpretable way. Such least squares estimation approach is useful in empirical work, but sometimes restrictive in terms of flexible modelling of $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$, or in terms of the parameter of interest in itself. To further improve the scope of our work, Section 4 studies a more general class of binscatter estimators based on nonlinear estimation, and also discusses a richer class of parameters of interest. Because those generalizations are technically more involved, and with cumbersome notation, we relegate them to after we have introduced our main results for least squares binscatter methods.

3 Using Least Squares Binscatter

We present a summary of our main estimation and inference results for the extended binscatter estimator $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ in (2.2), assuming for simplicity that $\widehat{\mathbf{w}}$ is a $\sqrt{n}$-consistent estimator of the evaluation point $\mathbf{w}$ in (2.3). In particular, if $\widehat{\mathbf{w}} = \mathbf{0}$ then $\widehat{\Upsilon}^{(v)}_{\mathbf{0}}(x) = \widehat{\mu}^{(v)}(x)$. Section SA-1.1 gives a full account on how our technical work improves on the non-/semi-parametric least squares series estimation literature. In particular, we account for random partitions ($\widehat{\Delta}$) and obtain a novel strong approximation for uniform inference (with optimal, up to $\log(n)$ terms, rate restrictions). As a consequence, our results are sharp enough to allow for (possibly covariate-adjusted) canonical binscatter ($p = 0$), a result excluded by the prior literature.

3.1 Implementing Binscatter

Canonical binscatter sets $p = s = 0$, with or without covariate-adjustment, but other choices are also natural: for example, a cubic B-spline sets $p = s = 3$. Setting (p, s) when implementing the binscatter estimator $\widehat{\mu}^{(v)}(x)$ is easier relative to selecting the number of bins J, which is a tuning parameter akin to a bandwidth choice in kernel smoothing settings. Therefore, we discuss a valid and optimal choice of J based on minimizing the density-weighted integrated mean square error (IMSE) of $\widehat{\mu}^{(v)}(x)$.

Letting $\approx_{\mathbb{P}}$ denote an approximation in probability, as $J \to \infty$ and $n \to \infty$, we show (Corollary SA-2.4) that

$$\int \left(\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)\right)^2 f_X(x)dx \;\approx_{\mathbb{P}}\; \frac{J^{1+2v}}{n}\mathscr{V}_n(p,s,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,s,v),$$

where these two terms capture the asymptotic variance and (squared) bias of binscatter, respectively, as a function of the polynomial order used (p), the desired derivative to be approximated (v), and the level of smoothness imposed across bins (s). Both quantities are fully characterized in the supplemental appendix (Section SA-2.5), where they are shown to be non-random functions of the sample size n. The variance $\mathscr{V}_n(p,s,v)$, depending on $\sigma^2(x)$ and $f_X(x)$, is bounded and bounded away from zero under minimal assumptions, while the (squared) bias $\mathscr{B}_n(p,s,v)$, depending on $\mu_0^{(p+1)}(x)$ and $f_X(x)$, is generally bounded and bounded away from zero except for a very special case (Remark SA-2.5). Our precise characterization of $\mathscr{V}_n(p,s,v)$ and $\mathscr{B}_n(p,s,v)$ is useful to approximate them in practice for implementation purposes. Furthermore, we show that $\mathscr{V}_n(p,0,v) \to \mathscr{V}(p,0,v)$ and $\mathscr{B}_n(p,0,v) \to \mathscr{B}(p,0,v)$, where $\mathscr{V}(p,0,v)$ and $\mathscr{B}(p,0,v)$ are cumbersome quantities in general (Corollary SA-2.3). However, for special leading cases, the variance and (squared) bias take very simple forms: $\mathscr{V}(0,0,0) = \mathbb{E}[\sigma^2(x_i)]$ and $\mathscr{B}(0,0,0) = \frac{1}{12}\mathbb{E}\big[\big(\frac{\mu_0^{(1)}(x_i)}{f(x_i)}\big)^2\big]$, which corresponds to the canonical binscatter ($p = v = s = 0$).

The main takeaway is that the approximate IMSE of binscatter depends on the squared bias and variance of the estimator, and these factors can be balanced out in order to select an approximate IMSE-optimal number of bins for applications.

Theorem 1 (IMSE-Optimal Binscatter)**.** *Let Assumption 1 hold,* $0 \leq v, s \leq p$*, and* $J\log(J)/n \to 0$

and $nJ^{-4p-5} \to 0$. *Then, the IMSE-optimal number of bins for implementing binscatter is*

$$J_{\mathit{IMSE}} = \left\lceil \left(\frac{2(p-v+1)\mathscr{B}_n(p,s,v)}{(1+2v)\mathscr{V}_n(p,s,v)} \right)^{\frac{1}{2p+3}} n^{\frac{1}{2p+3}} \right\rceil,$$

where $\lceil \cdot \rceil$ *denotes the ceiling operator, and* $\mathscr{B}_n(p,s,v)$ *and* $\mathscr{V}_n(p,s,v)$ *are non-random* n*-varying bounded sequences (see Section SA-2.5 for details).*

This theorem gives an IMSE-optimal choice of J for the general class of covariate-adjusted least squares extended binscatter estimators in (2.2), that is, allowing for higher-order polynomial fits within bins ($p \geq 0$) and possibly imposing smoothness restrictions on the fits across bins ($s \leq p$), with or without covariate adjustment ($\mathbf{w}_i$), when the main object of interest is possibly a derivative ($v \leq p$) of the unknown function $\mu_0(\cdot)$.

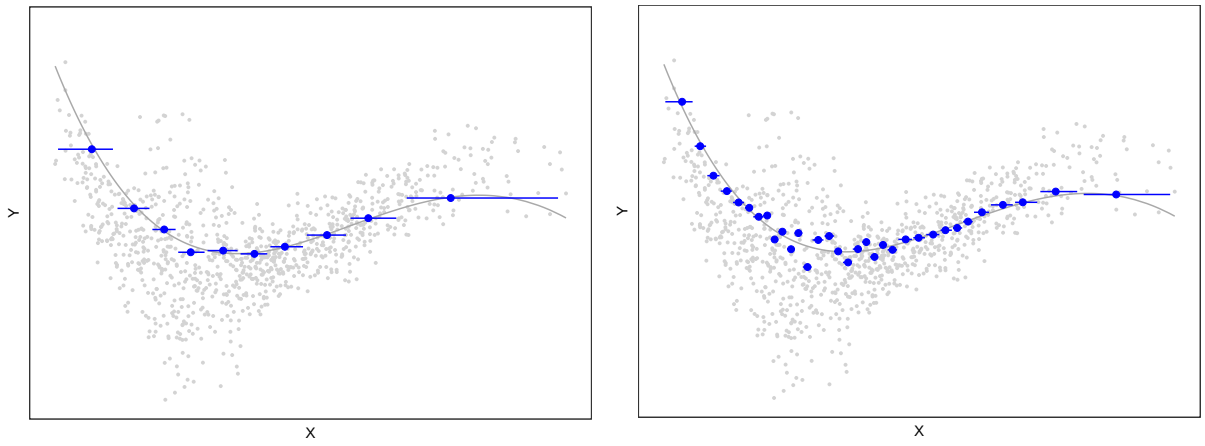
Section SA-5 of the supplemental appendix discusses implementation issues for selecting J in applications. Under mild additional regularity conditions, the estimator $\widehat{J}_{\mathtt{IMSE}}$ presented there satisfies $\widehat{J}_{\mathtt{IMSE}}/J_{\mathtt{IMSE}} \to_{\mathbb{P}} 1$, where $\to_{\mathbb{P}}$ denotes convergence in probability. See Section SA-5 and Cattaneo, Crump, Farrell, and Feng (2021) for more implementation details.

Figure 4 gives an illustration of our data-driven choice of number of bins, $\widehat{J}_{\mathtt{IMSE}}$, which trades off bias and variance in a valid and optimal way. Most of the current binscatter applications employ an ad-hoc number of bins, usually $J = 10$ or $J = 20$. There is no a priori reason why these choices would be valid: these ad-hoc choices can be too "small" to account for a nonlinear relationship $\mu_0(x)$ (i.e., too much misspecification bias), leading to incorrect empirical conclusions. Furthermore, even when too "large", there is no a priori reason why they would be optimal in terms of the usual bias-variance trade-off. Depending on the underlying unknown features of the data, such an arbitrary choice of J could be too "small" or "large", leading to graphical and formal procedures with invalid or at least unreliable statistical properties.

3.2 Pointwise Inference and Confidence Intervals

Binscatter can be implemented using $J_{\mathtt{IMSE}}$, or an estimator thereof, which leads to an IMSE-optimal point estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$. Thus, we turn next to confidence interval estimators to assess

Figure 4: Number of Bins (J) for Binscatter Estimators.



(a) $p = 0$, $s = 0$, $J = 10$ (b) $p = 0$, $s = 0$, $\widehat{J}_{\texttt{IMSE}} = 33$

Notes. This figure illustrates the difference between the covariate-adjusted, least-squares binscatter estimator implemented with an *ad hoc* number of bins ($J = 10$) as compared to a principled choice ($\widehat{J}_{\texttt{IMSE}} = 33$) based on the density-weighted integrated mean square error. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey line represents the true function, $\mu_0(\cdot)$, and grey dots represent the realized data. The sample size is $n = 1,000$.

the pointwise (in x) uncertainty of binscatter. The Studentized t-statistic is

$$\widehat{T}_p(x) = \frac{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad \widehat{\Omega}(x) = \widehat{\mathbf{b}}^{(v)}_s(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}^{(v)}_s(x),$$

where $\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'$ and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2$. The variance formula $\widehat{\Omega}(x)$ follows directly from the least squares structure of $\widehat{\mu}^{(v)}(x)$ because

$$\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x) \;\approx_{\mathbb{P}}\; \widehat{\mathbf{b}}^{(v)}_s(x)'\widehat{\mathbf{Q}}^{-1}\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{b}}_s(x_i)\epsilon_i, \tag{3.1}$$

where the approximation ignores the smoothing/misspecification bias of the estimator. This stochastic linearization motivates the following result.

Lemma 1 (Distributional Approximation: Pointwise)**.** *Let Assumption 1 hold,* $0 \leq v, s \leq p$*, and* $J^2\log^2(J)/n \to 0$ *and* $nJ^{-2p-3} \to 0$*. Then,*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\big[\widehat{T}_p(x) \leq u\big] - \Phi(u)\right| \to 0, \qquad \textit{for each } x \in \mathcal{X},$$

where $\Phi(u)$ *denotes the distribution function of a normal random variable.*

Lemma 1 can be used to form asymptotically valid confidence intervals for $\Upsilon_{\mathbf{w}}^{(v)}(x)$, pointwise in $x \in \mathcal{X}$, provided the misspecification error introduced by binscatter is removed from the distributional approximation. Specifically, for a choice p, the confidence intervals take the form:

$$\widehat{I}_p(x) = \left[\ \widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x) \pm \mathfrak{c} \cdot \sqrt{\widehat{\Omega}(x)/n}\ \right], \qquad 0 \leq v, s \leq p, \tag{3.2}$$

where $\mathfrak{c}$ denotes a choice of quantile (e.g., $\mathfrak{c} \approx 1.96$ for a 95% confidence intervals). However, employing an IMSE-optimal binscatter (Theorem 1) introduces a first-order misspecification error leading to invalidity of these confidence intervals. To address this problem, we rely on robust bias-correction (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018) to form valid confidence intervals based on IMSE-optimal binscatter, that is, without altering the partitioning scheme $\widehat{\Delta}$ used.

Our proposed robust bias-corrected binscatter confidence intervals are constructed as follows. First, for a given choice of p, we select the number of bins in $\widehat{\Delta}$ via Theorem 1, and construct the binscatter accordingly. Then, we employ the confidence interval $\widehat{I}_{p+q}(x)$ with $\mathfrak{c} = \Phi^{-1}(1-\alpha/2)$ and $q \geq 1$. The following Theorem summarizes this approach.

Theorem 2 (Confidence Intervals)**.** *Suppose the conditions in Lemma 1 hold and* $J = J_{\texttt{IMSE}}$*. If* $\mathfrak{c} = \Phi^{-1}(1-\alpha/2)$*, then*

$$\mathbb{P}\Big[\Upsilon_{\mathbf{w}}^{(v)}(x) \in \widehat{I}_{p+q}(x)\Big] \to 1-\alpha, \qquad \textit{for all } x \in \mathcal{X}.$$

The confidence intervals constructed in the above theorem are based on an IMSE-optimal implementation of binscatter and robust bias correction.

3.3 Uniform Inference and Confidence Bands

In many empirical applications of binscatter, the goal is to conduct inference uniformly over $x \in \mathcal{X}$ as opposed to pointwise in $x \in \mathcal{X}$. Examples include reporting confidence bands for $\Upsilon_{\mathbf{w}}^{(v)}(x)$ and its derivatives, as well as testing for linearity, monotonicity, concavity, or other shape features of $\Upsilon_{\mathbf{w}}^{(v)}(x)$. This section develops a formal approach for uniform inference employing binscatter and its generalizations, and constructs valid confidence bands. In the following subsections, we employ

these uniform inference results to develop asymptotically valid testing procedures for parametric model specification, nonparametric shape restrictions, and multi-sample comparisons.

Our approach to uniform inference extends recent work on strong approximations to allow for estimated quantile-spaced partitioning $\widehat{\Delta}$. In fact, it is not possible to obtain a valid strong approximation for the entire stochastic process $\{\widehat{T}_p(x) : x \in \mathcal{X}\}$ because uniformity fails when the partitioning scheme is random. As an alternative, we establish a conditional Gaussian strong approximation as the key building block for uniform inference. Heuristically, our strong approximation is constructed as follows: letting $\mathbf{N}_K$ denote a K-dimensional standard Gaussian random vector, independent of the data,

$$\begin{aligned}\sqrt{n}\big(\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)\big) \;&\approx_{\mathbb{P}}\; \widehat{\mathbf{b}}^{(v)}_s(x)'\widehat{\mathbf{Q}}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\mathbf{b}}_s(x_i)\epsilon_i \\ &\approx_{\mathbb{P}}\; \widehat{\mathbf{b}}^{(v)}_s(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}^{1/2}\mathbf{N}_{(p+1)J-(J-1)s},\end{aligned}$$

where the first approximation is analogous to (3.1) but now established uniformly over $x \in \mathcal{X}$ (Theorem SA-2.1), and the second approximation corresponds to a conditional coupling (Theorems SA-2.4 and SA-2.5). It is not difficult to show that $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{\Sigma}}$ are sufficiently close in probability to well-defined non-random matrices in the necessary norm (Lemma SA-2.1 and Theorem SA-2.2). However, $\widehat{\mathbf{b}}^{(v)}_s(x)$ fails to be close in probability to its non-random counterpart *uniformly* in $x \in \mathcal{X}$ due to the sharp discontinuity introduced by the indicator functions entering the binning procedure. Nevertheless, inspired by the work in Chernozhukov, Chetverikov, and Kato (2014a,b), our approach circumvents that technical hurdle by developing a strong approximation that is conditionally Gaussian first, retaining some of the randomness introduced by $\widehat{\Delta}$, and then using such coupling to deduce a distributional approximation for the specific functional of interest (e.g., suprema); see Section SA-4.1 for details.

Given the technical nature of our strong approximation results for quantile-spaced estimated partitions and binscatter, we relegate further details to the supplemental appendix (Section SA-2.4). Here we discuss how those results are used to construct valid robust bias-corrected confidence bands for $\mu^{(v)}_0(x)$; the two upcoming subsections discuss applications to functional hypothesis testing.

By construction, a confidence band for the entire function $\mu^{(v)}_0(x)$ with $1-\alpha$ coverage probability

is

$$\big\{\widehat{I}_{p+q}(x) : x \in \mathcal{X}\big\} \quad \text{with} \quad \mathfrak{c} = \inf\Big\{c \in \mathbb{R}_+ : \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|\widehat{T}_{p+q}(x)\big| \leq c\Big] \geq 1-\alpha\Big\},$$

where $\widehat{I}_p(x)$ is defined in (3.2). The quantile $\mathfrak{c}$ is unknown because the finite sample distribution of $\sup_{x\in\mathcal{X}} \big|\widehat{T}_{p+q}(x)\big|$ is unknown. Our strong approximation results allow us to approximate this distribution by resampling from a Gaussian vector of length $(p+q+1)J-(J-1)s$. To be more precise, define the following (conditional) Gaussian process:

$$\widehat{Z}_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}}{\sqrt{\widehat{\Omega}(x)/n}}\mathbf{N}_{(p+1)J-(J-1)s}, \qquad x \in \mathcal{X}. \tag{3.3}$$

The following result shows that the distribution of $\sup_{x\in\mathcal{X}} \big|\widehat{T}_p(x)\big|$ is well approximated by that of $\sup_{x\in\mathcal{X}} \big|\widehat{Z}_p(x)\big|$, conditional on the data $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : i = 1, 2, \ldots, n\}$.

Lemma 2 (Distributional Approximation: Supremum)**.** *Let Assumption 1 hold,* $0 \leq v, s \leq p$*, and* $J^2 \log^6(J)/n \to 0$ *and* $nJ^{-2p-3}\log J \to 0$*. Then,*

$$\sup_{u\in\mathbb{R}} \Big|\mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\widehat{T}_p(x)| \leq u\Big] - \mathbb{P}\Big[\sup_{x\in\mathcal{X}} |\widehat{Z}_p(x)| \leq u \Big| \mathbf{D}\Big]\Big| \to_{\mathbb{P}} 0.$$

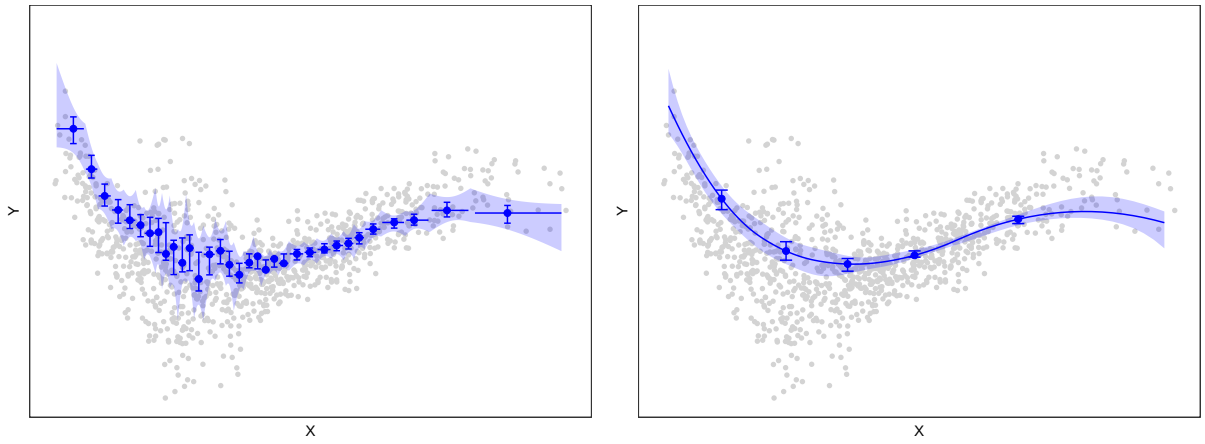
The proof of this Lemma employs a subtle technical argument to handle the random partitioning entering the (conditional) Gaussian process $\widehat{Z}_p(\cdot)$. See Theorem SA-4.1 and its proof for details. Building on that result, we are able to establish the large sample validity of the robust bias-corrected confidence bands for $\Upsilon_{\mathbf{w}}^{(v)}(x)$.

Theorem 3 (Confidence Bands)**.** *Suppose the conditions in Lemma 2 hold and* $J = J_{\texttt{IMSE}}$*. If* $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} \big|\widehat{Z}_{p+q}(x)\big| \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$*, then*

$$\mathbb{P}\Big[\Upsilon_{\mathbf{w}}^{(v)}(x) \in \widehat{I}_{p+q}(x), \text{ for all } x \in \mathcal{X}\Big] \to 1-\alpha.$$

In practice, the supremum in $\sup_{x\in\mathcal{X}} \big|\widehat{Z}_{p+q}(x)\big|$ is replaced by a maximum over a fine grid of evaluation points and, conditional on the original data, each realization of $\widehat{Z}_{p+q}(\cdot)$ is obtained by resampling from the standard Gaussian random vector $\mathbf{N}_{(p+q+1)J-(J-1)s}$. Further details on implementation are given in Section SA-5 and in Cattaneo, Crump, Farrell, and Feng (2021).

Figure 5: Robust Bias Corrected Confidence Intervals and Confidence Bands.



(a) Canonical Binscatter: $p = 0$ and $s = 0$ (b) Extended Binscatter: $p = 2$ and $s = 2$

Notes. This figure reports robust, bias-corrected confidence intervals and confidence bands for the covariate-adjusted, least-squared extended binscatter estimator. Plot (a) reports a binscatter estimator using $p = 0$ and $s = 0$ with associated confidence intervals (vertical capped spikes) and confidence bands (shaded region) using $p = 1$ and $s = 1$. Plot (b) reports a binscatter estimator using $p = 2$ and $s = 2$ with associated confidence intervals (vertical capped spikes) and confidence bands (shaded region) using $p = 3$ and $s = 3$. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey dots represent the realized data. Nominal coverage for confidence intervals and confidence bands is 95%. The sample size is $n = 1,000$.

Loosely speaking, a band is simply a confidence "interval" for a function, and like a traditional confidence interval, it is given by the area between two "endpoint" functions, say $\hat{\mu}_{\mathtt{U}}(x)$ and $\hat{\mu}_{\mathtt{L}}(x)$. We may then make statements analogous to those for usual confidence intervals. For example, if this band does not contain a line (or quadratic function), then we say that at level α we reject the null hypothesis that $\mu_0(x)$ is linear (or quadratic). Visually, the size of the band reflects the uncertainty in the data, both in terms of overall sampling uncertainty and any heteroskedasticity patterns. Figure 5 shows pointwise confidence intervals and uniform confidence bands for the simulated data, and we see that the size and shape of the band reflects the underlying data.

3.4 Testing Parametric Specifications

Binscatter is often used to heuristically assess different types of shape features of the unknown regression function and its derivatives. In this section, we provide a rigorous formalization of one such kind of hypothesis test: parametric specifications of $\Upsilon_{\mathbf{w}}^{(v)}(x)$. In the next section, we discuss another type of shape-related hypothesis tests: testing for nonparametric features such as monotonicity or concavity of $\Upsilon_{\mathbf{w}}^{(v)}(x)$.

One type of informal analysis commonly encountered in empirical work concerns comparing the binscatter $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ relative to some parametric fit. For example, setting $v = 0$ and abstracting from covariate adjustment for simplicity, $\widehat{\mu}(x)$ can be compared to $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ to assess whether there is a relationship between y_i and x_i or, more formally, whether $\mu_0(x)$ is a constant function. Similarly, it is common to see binscatter used to assess whether there is a linear or perhaps quadratic relationship, that is, whether $\Upsilon_{\mathbf{w}}(x) = \theta_0 + x\theta_1 + \mathbf{w}'\boldsymbol{\gamma}$ or perhaps $\Upsilon_{\mathbf{w}}(x) = \theta_0 + x\theta_1 + x^2\theta_2 + \mathbf{w}'\boldsymbol{\gamma}$ for some unknown coefficients $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$ and $\boldsymbol{\gamma}$. More generally, researchers are often interested in assessing formally whether $\Upsilon_{\mathbf{w}}(x) = M_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma})$ when $M_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma}) = m(x; \boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}$ for some $m(\cdot)$ known up to a finite parameter $\boldsymbol{\theta}$, which can be estimated using the available data. We formalize this class of hypothesis tests as follows: for a choice of v and function $M^{(v)}_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma}) = \partial^v M_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma})/\partial x^v$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$, the null and alternative hypotheses are

$$\begin{aligned}
\ddot{\mathsf{H}}_0 :&\quad \sup_{x\in\mathcal{X}} \left| \Upsilon^{(v)}_{\mathbf{w}}(x) - M^{(v)}_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma}) \right| = 0, \quad \text{for some } \boldsymbol{\theta} \text{ and } \boldsymbol{\gamma}, \qquad vs. \\
\ddot{\mathsf{H}}_{\mathrm{A}} :&\quad \sup_{x\in\mathcal{X}} \left| \Upsilon^{(v)}_{\mathbf{w}}(x) - M^{(v)}_{\mathbf{w}}(x; \boldsymbol{\theta}, \boldsymbol{\gamma}) \right| > 0, \quad \text{for all } \boldsymbol{\theta} \text{ and } \boldsymbol{\gamma}.
\end{aligned}$$

This hypothesis testing problem can be implemented using test statistics involving the supremum of (derivatives of) binscatter, possibly employing higher-order polynomials, imposing smoothness restrictions, or adjusting for additional covariates. In all cases it is required to approximate the quantiles of the finite sample distribution of such statistics, which can be done in a similar fashion as discussed above for constructing confidence bands.

Since $\boldsymbol{\theta}$ is unknown and not set by the null hypothesis, we construct a feasible testing procedure by assuming that there exists an estimator $(\widetilde{\boldsymbol{\theta}}', \widetilde{\boldsymbol{\gamma}}')'$ that consistently estimates $(\boldsymbol{\theta}', \boldsymbol{\gamma}')'$ under the null hypothesis (correct parametric specification), and that is "well behaved" under the alternative hypothesis (parametric misspecification). See Theorem 4 below for precise restrictions. Then, we consider the following test statistic

$$\ddot{T}_p(x) = \frac{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - M^{(v)}_{\widehat{\mathbf{w}}}(x; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad 0 \leq v, s \leq p,$$

leading to the hypothesis test:

$$\text{Reject } \ddot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} |\ddot{T}_p(x)| \geq \mathfrak{c}, \tag{3.4}$$

for an appropriate choice of critical value $\mathfrak{c}$ to control false rejections (Type I error).

The following theorem gives the remaining details, and makes the hypothesis testing procedure (3.4) feasible.

Theorem 4 (Hypothesis Testing: Parametric Specification)**.** *Let Assumption 1 hold, and set* $J = J_{\texttt{IMSE}}$ *and* $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+q}(x)| \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$.

Under $\ddot{\mathsf{H}}_0$, *if* $\sup_{x\in\mathcal{X}} |M^{(v)}_{\widehat{\mathbf{w}}}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - \Upsilon^{(v)}_{\mathbf{w}}(x)| = O_{\mathbb{P}}(n^{-1/2})$, *then*

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|\ddot{T}_{p+q}(x)\big| > \mathfrak{c}\Big] = \alpha.$$

Under $\ddot{\mathsf{H}}_A$, *if there exists* $(\bar{\boldsymbol{\theta}}', \bar{\boldsymbol{\gamma}}')'$ *such that* $\sup_{x\in\mathcal{X}} |M^{(v)}_{\widehat{\mathbf{w}}}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - M^{(v)}_{\widehat{\mathbf{w}}}(x;\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\gamma}})| = O_{\mathbb{P}}(n^{-1/2})$, *then*

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|\ddot{T}_{p+q}(x)\big| > \mathfrak{c}\Big] = 1.$$

This theorem formalizes an intuitive idea: if the confidence band for $\Upsilon^{(v)}_{\mathbf{w}}(x)$ does not entirely contain the parametric fit considered, then such parametric fit is incompatible with the data, i.e., should be rejected. Formally, this leads to the hypothesis testing procedure (3.4), which relies on a proper (simulated) critical value. The condition $\sup_{x\in\mathcal{X}} |M^{(v)}_{\widehat{\mathbf{w}}}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - \Upsilon^{(v)}_{\mathbf{w}}(x)| = O_{\mathbb{P}}(n^{-1/2})$ under the null hypothesis is mild: it essentially states that the unknown parameters entering the parametric specification $\mu_0(x) = m(x,\boldsymbol{\theta})$ is $\sqrt{n}$-estimable, provided some mild regularity holds for the known regression function $m(x,\boldsymbol{\theta})$. For example, a simple sufficient condition is $\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_{\mathbb{P}}(1)$ and $m^{(v)}(x,\boldsymbol{\theta})$ is continuous in x and continuously differentiable in $\boldsymbol{\theta}$. Most standard parametric models in applied microeconomics satisfy such conditions, including linear and nonlinear regression models, discrete choice models (Probit or Logit), censored and truncation models, and many more.

Remark 2 (Other Metrics)**.** The parametric specification testing methods are based on the maximum (uniform) discrepancy between one binscatter fit and possibly some parametric fit. Some

practitioners, however, may prefer to assess the discrepancy by means of an alternative metric. For instance, one can construct a testing procedure using the mean squared difference between the parametric and nonparametric fits:

$$\text{Reject } \ddot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \int_{\mathcal{X}} |\ddot{T}_p(x)|^2 dx \geq \mathfrak{c},$$

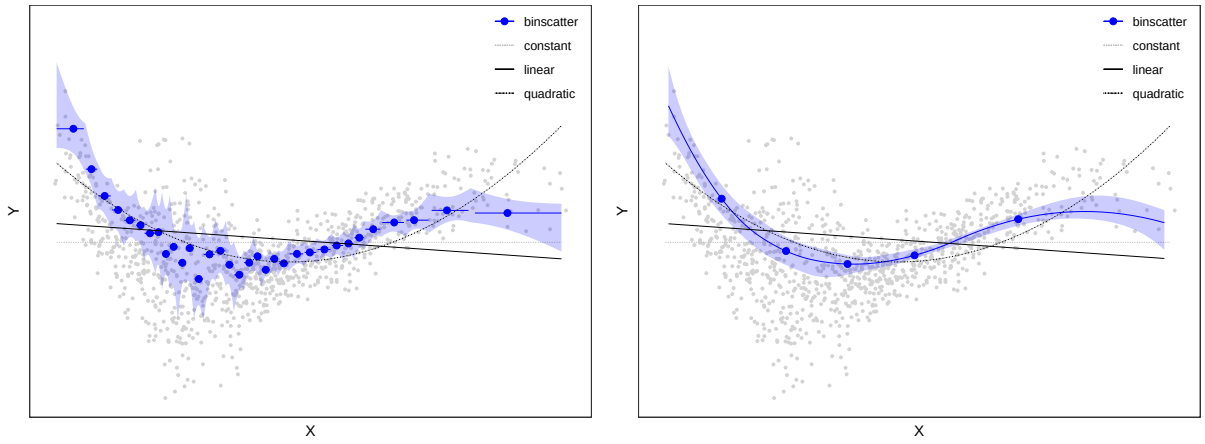
for $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\int_{\mathcal{X}} |\widehat{Z}_{p+q}(x)|^2 dx \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$. Our theoretical results are general enough to accommodate such alternative comparisons, which are also implemented in our companion software (Cattaneo, Crump, Farrell, and Feng, 2021). ⌟

In practice, it is natural to combine the formal hypothesis test emerging from Theorem 4 with a binned scatter plot that includes a binscatter confidence band and a line representing the parametric fit. We illustrate our parametric specification testing approach using the simulated data. Heuristically, our formal parametric specification testing approach is based on comparing the maximal empirical deviation between the binscatter and the desired parametric specification for $\mu_0(x)$. If the parametric specification is correct, then there should be no deviation beyond what is explained by random sampling for all evaluation points x, and hence the parametric fit should be contained in the confidence band for $\mu_0(x)$. Figure 6 depicts this testing approach. Also, one may construct testing procedures using other metrics, as discussed in Remark 2, or perhaps over restricted portions of the support of x_i. In Table 1 we report formal testing results based on sup norm and L_2 norm. The comparison between parametric fits and confidence bands in Figure 6 corresponds to the sup-norm-based tests reported in the first two columns.

3.5 Testing Shape Restrictions

The hypothesis test (3.4) concerns parametric specification testing for a choice of $m(x, \boldsymbol{\theta})$, but it can also be used to conduct certain nonparametric shape restriction testing. For example, if the function $\mu_0(x)$ is constant, then $\mu_0^{(1)}(x) = 0$ for all $x \in \mathcal{X}$, which can be implemented using Theorem 4 upon setting $m(\cdot) = 0$ and $v = 1$, for any $p \geq 1$. Similarly, linearity or other related nonparametric shape restrictions can be tested via Theorem 4, for appropriate choice of v. The common feature in all cases is that the null hypothesis of interest is two-sided.

Figure 6: Graphical Representation of Parametric Specification Testing (Table 1).



(a) Canonical Binscatter: $p = 0$ and $s = 0$ (b) Extended Binscatter: $p = 2$ and $s = 2$

Notes. This figure reports robust, bias-corrected confidence bands for the covariate-adjusted, least-squared extended binscatter estimator. In addition, these figures report constant, linear, and quadratic regression fits to the data. Plot (a) reports a binscatter estimator using $p = 0$ and $s = 0$ with an associated confidence band (shaded region) using $p = 1$ and $s = 1$. Plot (b) reports a binscatter estimator using $p = 2$ and $s = 2$ with an associated confidence band (shaded region) using $p = 3$ and $s = 3$. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey dots represent the realized data. Nominal coverage for confidence bands is 95%. The sample size is $n = 1,000$.

There are, however, other shape restriction hypotheses about $\Upsilon_{\mathbf{w}}^{(v)}(x)$ or $\mu_0(x)$ that correspond to one-sided null hypotheses, and thus cannot be implemented using Theorem 4. For example, negativity, monotonicity and concavity of $\Upsilon_{\mathbf{w}}^{(v)}(x)$ all correspond to formal statements of the form $\Upsilon_{\mathbf{w}}(x) \leq 0$, $\Upsilon_{\mathbf{w}}^{(1)}(x) \leq 0$ and $\Upsilon_{\mathbf{w}}^{(2)}(x) \leq 0$, respectively. Thus, in this section we also study the following class of hypothesis tests: for a choice of v, the null and alternative hypotheses are

$$\dot{\mathsf{H}}_0 : \quad \sup_{x \in \mathcal{X}} \Upsilon_{\mathbf{w}}^{(v)}(x) \leq 0, \qquad vs. \qquad \dot{\mathsf{H}}_{\mathrm{A}} : \quad \sup_{x \in \mathcal{X}} \Upsilon_{\mathbf{w}}^{(v)}(x) > 0.$$

These hypotheses highlight the importance of extending binscatter to derivative estimation, which necessarily requires considering $p \geq v > 0$, with or without smoothness restrictions or covariate adjustments. In other words, considering higher-order polynomial fits within bins is not a spurious generalization of binscatter, but rather a fundamental input for implementing the above nonparametric shape-related hypothesis tests.

To make our hypothesis testing procedures precise, we employ the following feasible, Studentized

Table 1: Formal Tests for Parametric Specifications.

Binscatter	**Sup norm**		**L_2 norm**		
	Test Statistic	P-value	Test Statistic	P-value	$\hat{J}_{\texttt{IMSE}}$
Canonical: $p = s = 0$					
Constant	16.72	0.00	5.38	0.00	33
Linear	14.05	0.00	5.61	0.00	33
Quadratic	14.51	0.00	3.37	0.00	33
Extended: $p = s = 2$					
Constant	15.70	0.00	7.62	0.00	5
Linear	15.12	0.00	8.43	0.00	5
Quadratic	14.93	0.00	5.45	0.00	5

Notes. This table reports the test statistics and associated p-values from hypothesis tests of parametric specifications. The top panel reports test results under the null of a constant, linear, or quadratic regression model using $p = 1$ and $s = 1$ to construct critical values. The bottom panel reports test results under the null of a constant, linear, or quadratic regression model using $p = 3$ and $s = 3$ to construct critical values. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The sample size is $n = 1,000$.

statistic:

$$\dot{T}_p(x) = \frac{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad x \in \mathcal{X}, \quad 0 \leq v, s \leq p$$

leading to the hypothesis test:

$$\text{Reject } \dot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x \in \mathcal{X}} \dot{T}_p(x) \geq \mathfrak{c}, \tag{3.5}$$

for an appropriate choice of critical value $\mathfrak{c}$ to control false rejections (Type I error). Of course, the other one-sided hypothesis tests are constructed in the obvious symmetric way.

Theorem 5 (Hypothesis Testing: nonparametric Shape Restriction)**.** *Let Assumption 1 hold, and set* $J = J_{\texttt{IMSE}}$ *and* $\mathfrak{c} = \inf\left\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x \in \mathcal{X}} \widehat{Z}_{p+q}(x) \leq c \;\big|\; \mathbf{D}\big] \geq 1 - \alpha\right\}$.

Under $\dot{\mathsf{H}}_0$*, then*

$$\lim_{n \to \infty} \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \dot{T}_{p+q}(x) > \mathfrak{c}\Big] \leq \alpha.$$

Under $\dot{\mathsf{H}}_A$*, then*

$$\lim_{n \to \infty} \mathbb{P}\Big[\sup_{x \in \mathcal{X}} \dot{T}_{p+q}(x) > \mathfrak{c}\Big] = 1.$$

This theorem shows that the hypothesis testing procedure (3.5) is valid. Because of its one-sided nature, the test is conservative in general. Further, because it relies on a supremum-type statistic, this nonparametric shape restriction test also employs a simulated critical value, just like those

used in the previous sections to construct confidence bands or to conduct parametric specification testing. Theorem 5 corresponds to the one-sided "left" hypothesis test, but of course the analogous theorem "to the right" also holds. Our software implementation allows for all three possibilities: one-sided (left or right) and two-sided hypothesis testing. See Cattaneo, Crump, Farrell, and Feng (2021) for more details.

We illustrate our testing approach using the simulated data. The results are reported in Table 2. The "Negativity" test corresponds to the test of $\sup_{x\in\mathcal{X}} \Upsilon_{\mathbf{w}}(x) \leq 0$, the "Decreasingness" test corresponds to the test of $\sup_{x\in\mathcal{X}} \Upsilon^{(1)}_{\mathbf{w}}(x) \leq 0$, and the "Concavity" test corresponds to the test of $\sup_{x\in\mathcal{X}} \Upsilon^{(2)}_{\mathbf{w}}(x) \leq 0$. In the first row of Table 2, we use canonical binscatter ($p = s = 0$) to select the number of bins J and then test for negativity based on the linear spline ($p = s = 1$). In the second and third rows, we select the number of bins J based on the linear spline ($p = s = 1$) and then test for both negativity and decreasingness based on the quadratic spline ($p = s = 2$). In the last three rows, we select the number of bins J based on the quadratic spline ($p = s = 2$) and then test for negativity, decreasingness, and concavity based on the cubic spline ($p = s = 3$). Note that using higher-order polynomials, i.e., increasing p, allows us to test for properties of higher-order derivatives of the unknown function.

Table 2: Formal Tests for Shape Restrictions.

Binscatter	Test Statistic	P-value	$\hat{J}_{\texttt{IMSE}}$
Canonical: $p = s = 0$			
Negativity	249.58	0.00	33
Extended: $p = s = 1$			
Negativity	395.03	0.00	10
Decreasingness	10.23	0.00	9
Extended: $p = s = 2$			
Negativity	455.01	0.00	5
Decreasingness	11.80	0.00	5
Concavity	8.78	0.00	4

Notes. This table reports the test statistics and associated p-values from hypothesis tests of shape restrictions on $\mu_0(\cdot)$. The top panel reports test results under the null hypothesis that the function only takes on negative values, using $p = 1$ and $s = 1$ to construct critical values. The middle panel reports test results under the null hypothesis that the function only takes on negative values or is monotonically decreasing, using $p = 2$ and $s = 2$ to construct critical values. The bottom panel reports test results under the null hypothesis that the function only takes on negative values, is monotonically decreasing, or is concave, using $p = 3$ and $s = 3$ to construct critical values. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The sample size is $n = 1,000$.

Finally, in SA-4.3 we include a more general version of Theorem 5, where we introduce one-sided

tests against a parametric fit. For example, setting $v = 0$ and abstracting from covariate adjustment for simplicity, this more involved testing procedure might be useful to assess whether $\mu_0(x)$ always resides "below" the line defined by $m(x, \bar{\boldsymbol{\theta}}; a) = a + x\bar{\boldsymbol{\theta}}$, when $\widetilde{\boldsymbol{\theta}}$ is the OLS estimator based on y_i and x_i, $\bar{\boldsymbol{\theta}}$ is its probability limit (i.e., $x\bar{\boldsymbol{\theta}}$ denotes the best linear predictor of y_i and x_i), and a is a user-chosen positive constant.

3.6 Multi-Sample Estimation and Testing

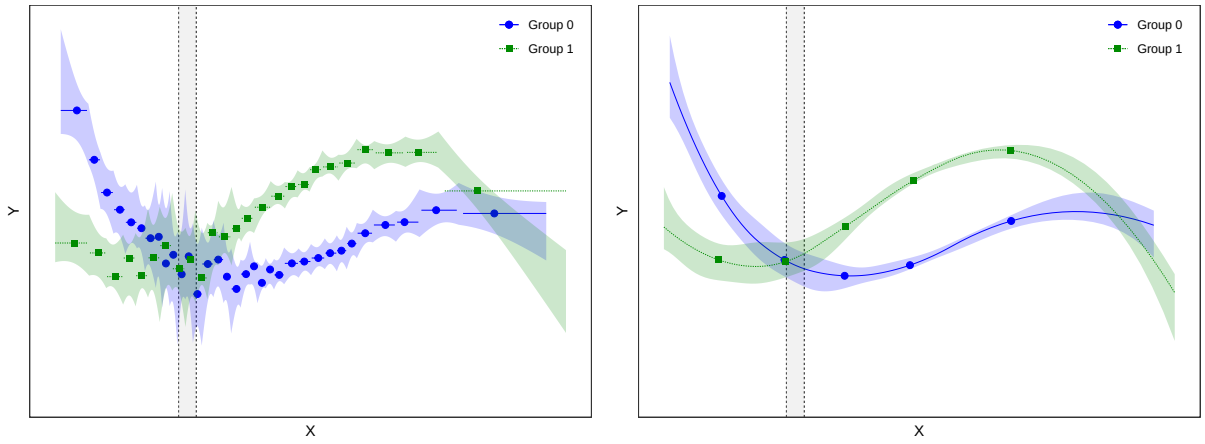
In program evaluation and causal inference settings, researchers often want to compare mean, quantile, and other regression functions across different groups (or treatment arms), capturing heterogeneous in x_i treatment effects, possibly after controlling for $\mathbf{w}_i$. Therefore, for each subsample defined by a group indicator, the parameter of interest can be defined as $\Upsilon^{(v)}_{\mathbf{w},\ell}(x)$ denoting the parameter in (2.3) for subsample $\ell = 0, 1, 2, \ldots, L$. For example, in a randomized experiment with a binary treatment, $\Upsilon^{(v)}_{\mathbf{w},0}(x)$ and $\Upsilon^{(v)}_{\mathbf{w},1}(x)$ may denote the partial mean effect for control and treatment units, respectively, and thus $\Upsilon^{(v)}_{\mathbf{w},1}(x) - \Upsilon^{(v)}_{\mathbf{w},0}(x)$ can be interpreted as the average treatment effect conditional on $x_i = x$. The latter parameter naturally captures treatment effect heterogeneity along the x_i dimension.

Our theoretical results can also handle nonparametric testing about features of $\Upsilon^{(v)}_{\mathbf{w},\ell}(x)$, and transformations thereof for two or more groups. For example, assuming that two sub-samples are available ($L = 1$), our methods can be used to formally test for the null hypothesis: $\mathsf{H}_0 : \Upsilon^{(v)}_{\mathbf{w},0}(x) = \Upsilon^{(v)}_{\mathbf{w},1}(x)$ for all $x \in \mathcal{X}$, which captures the idea of no (heterogeneous) treatment effect. As a second example, our theory can be used to quantify uncertainty for the largest heterogeneous treatment effect in a binary treatment setting:

$$\widehat{x} = \arg\sup_{x \in \mathcal{X}} \left| \widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}},1}(x) - \widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}},0}(x) \right|.$$

These and many other problems of interest in applied microeconometrics concern the uniform discrepancy of two or more binscatter function estimators, which can be analyzed using our strong approximation and related theoretical results in the supplemental appendix. We do not provide further details here to conserve space, but our companion software implements several multi-sample estimation, uncertainty quantification and hypothesis testing procedures (Cattaneo, Crump, Farrell,

Figure 7: Graphical Testing of Two-Sample Comparison.



(a) Canonical Binscatter: $p = 0$ and $s = 0$ (b) Extended Binscatter: $p = 2$ and $s = 2$

Notes. This figure reports robust, bias-corrected confidence bands for the covariate-adjusted, least-squared extended binscatter estimator for each of two subgroups. Plot (a) reports a binscatter estimator using $p = 0$ and $s = 0$ with an associated confidence band (shaded region) using $p = 1$ and $s = 1$. Plot (b) reports a binscatter estimator using $p = 2$ and $s = 2$ with an associated confidence band (shaded region) using $p = 3$ and $s = 3$. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey shaded regions between vertical dashed lines correspond to the subsample with $x \in [0.22, 0.25]$ used in Table 3. Nominal coverage for confidence bands is 95%. The total sample size is $n = 2,000$ with $n = 1,000$ observations in each subgroup.

and Feng, 2021). To illustrate those results, we consider again the simulated data. Figure 7 plots the binscatter fits for two groups and their corresponding confidence bands. We can see the two bands intersect at some values of x, indicating a heterogeneous treatment effect along the x_i dimension. Moreover, we report the formal testing results (based on sup norm) for the two-group comparison in Table 3. The null hypothesis is $\Upsilon_{\mathbf{w},0}(x) = \Upsilon_{\mathbf{w},1}(x)$ for all $x \in \mathcal{X}$, i.e., there is no treatment effect. Given the p-values of our tests, for the full sample, this null hypothesis is rejected, while for the subsample with x between 0.22 and 0.25 (the grey shaded region in Figure 7), it is not, which are in line with the graphical illustration in Figure 7.

4 Generalized nonlinear Binscatter

The discussion so far has focused on least squares covariate-adjusted binscatter. This section considers a more general class of binscatter methods, including examples such as nonlinear least squares regression, quantile regression, and MLE methods. More precisely, we consider the following

Table 3: Formal Tests for Two-Sample Comparison.

Binscatter	Test Statistic	P-value	$\hat{J}_{\texttt{IMSE},0}$	$\hat{J}_{\texttt{IMSE},1}$
Canonical: $p = s = 0$				
$0.22 \leq x \leq 0.25$	1.63	0.43	4	5
full sample	22.15	0.00	33	25
Extended: $p = s = 2$				
$0.22 \leq x \leq 0.25$	1.87	0.21	2	3
full sample	39.69	0.00	5	5

Notes. This table reports the test statistics and associated p-values from hypothesis tests based on two-sample comparisons. The top panel reports test results for $x \in [0.22, 0.25]$ along with the full sample, using $p = 1$ and $s = 1$ to construct critical values. The bottom panel reports test results for $x \in [0.22, 0.25]$ along with the full sample, using $p = 3$ and $s = 3$ to construct critical values. $\widehat{J}_{\texttt{IMSE},0}$ and $\widehat{J}_{\texttt{IMSE},1}$ are the estimates of IMSE-optimal numbers of bins for the two groups. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The total sample size is 2,000 with 1,000 observations in each subgroup.

generalized nonlinear binscatter estimator:

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma}\big)\Big), \tag{4.1}$$

where the loss function $\eta \mapsto \rho(\cdot;\eta)$ is assumed to be absolutely continuous with a piecewise Lipschitz continuous weak derivative $\psi(y;\eta) = \psi(y-\eta)$ exhibiting at most a finite number of discontinuity points, and the (inverse) link function $\eta(\cdot)$ is assumed to be strictly monotonic and twice continuously differentiable. Furthermore, we assume $u \mapsto \rho(\cdot;\eta(u))$ is convex to simplify our technical results, but this restriction can be dropped with additional technical work.

Generalized nonlinear binscatter gives rise to a richer class of parameters of interest. Specifically, for some user-chosen evaluation point $\mathbf{w}$, we consider

$$\vartheta_{\mathbf{w}}^{(v)}(x) = \frac{\partial^v}{\partial x^v}\eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0), \tag{4.2}$$

where now the underlying parameters $\mu_0(\cdot)$ and $\boldsymbol{\gamma}_0$ are defined by

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \underset{\mu\in\mathcal{M},\boldsymbol{\gamma}\in\mathbb{R}^d}{\arg\min}\ \mathbb{E}[\rho(y_i;\eta(\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}))], \tag{4.3}$$

with $\mathcal{M}$ an appropriate space of functions satisfying certain (smoothness) conditions made precise in the supplemental appendix.

Under regularity conditions, (4.3) is the probability limit of (4.1), and therefore a natural plug-in binscatter estimator for $\vartheta_{\mathbf{w}}^{(v)}(x)$ is

$$\widehat{\vartheta}_{\widehat{\mathbf{w}}}^{(v)}(x) = \frac{\partial^v}{\partial x^v}\eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}), \tag{4.4}$$

where $\widehat{\mathbf{w}}$ is a consistent estimator of the desired evaluation point $\mathbf{w}$ in (4.2). For example, if $v = 0$, then $\widehat{\vartheta}_{\widehat{\mathbf{w}}}(x) = \eta(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})$ is an estimator of the relationship between x and y evaluated at level $\mathbf{w}$. If $v = 1$, then $\widehat{\vartheta}_{\widehat{\mathbf{w}}}^{(1)}(x) = \eta^{(1)}(\widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}})\widehat{\mu}^{(1)}(x)$ is an estimator of the marginal partial effect of x on y at level $\mathbf{w}$.

The least squares binscatter estimator (2.2) studied in previous sections corresponds to the choice $\rho(y;\eta) = (y-\eta)^2$ and $\eta(u) = u$, in which case $\widehat{\vartheta}_{\widehat{\mathbf{w}}}^{(v)}(x) = \widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ by construction. Furthermore, if $\mathbb{E}[y_i|x_i,\mathbf{w}_i] = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$, then $\vartheta_{\mathbf{w}}^{(v)}(x) = \Upsilon_{\mathbf{w}}^{(v)}(x)$, but otherwise $\vartheta_{\mathbf{w}}^{(v)}(x)$ would correspond to the best mean square approximation of $\mathbb{E}[y_i|x_i,\mathbf{w}_i]$ based on functions of the form $\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}$, as previously mentioned, and as defined in (4.3). The theoretical results presented in previous sections for least squares semi-linear binscatter were obtained under weaker conditions (Section SA-2) than those imposed for the general binscatter estimator (4.1), presented in Section SA-3, because the former admits closed-form expressions that can be leveraged substantially.

Many interesting problems do not admit a closed form solution due to the nonlinearity of $\eta(u)$ or the non-differentiability of $\rho(\cdot;\cdot)$ in (4.1). For example, nonlinear least squares binscatter estimators also employ a quadratic loss function $\rho(y;\eta) = (y-\eta)^2$, but the link function $\eta(u)$ is nonlinear (e.g., Logist, Probit or Poisson regression). As a consequence, under standard assumptions, $\widehat{\vartheta}_{\widehat{\mathbf{w}}}^{(v)}(x)$ can be interpreted as the generalized nonlinear binscatter estimator of (derivatives of) the conditional expectation $\mathbb{E}[y_i|x_i,\mathbf{w}_i] = \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)$, under correct specification. If misspecified, then the generalized nonlinear binscatter estimator would be consistent for the best mean square approximation of $\mathbb{E}[y_i|x_i,\mathbf{w}_i]$ based on functions of the form $\eta(\mu(x) + \mathbf{w}'\boldsymbol{\gamma})$ for some $\mu(\cdot)$ and $\boldsymbol{\gamma}$.

Another important class of binscatter estimators covered by (4.1) is semi-linear quantile regression. For example, the $\mathfrak{q}$-th quantile binscatter regression estimator sets $\rho(y;\eta) = (\mathfrak{q} - \mathbb{1}(y < \eta))(y-\eta)$ for some $0 < \mathfrak{q} < 1$ and $\eta(u) = u$. This case is of empirical interest even under misspecification: see Angrist, Chernozhukov, and Fernández-Val (2006) for further discussion.

Many other estimation procedures commonly used in empirical work are covered by (4.1), and

therefore by our theoretical and methodological results.

4.1 Technical Results

The theoretical results for generalized nonlinear binscatter are substantially more involved than those for least squares binscatter. The supplemental appendix (Sections SA-3 and SA-4) provide all the details. In a nutshell, we are able to extend all the results presented in previous sections for $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ in (2.4) to now accommodate $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$ in (4.4).

The crux of our theoretical work is a novel uniform Bahadur representation (Theorem SA-3.1) for the generalized nonlinear binscatter estimator $\widehat{\mu}^{(v)}(x)$ in (4.1). See Section SA-3 for details. To give a summary of that result, under standard but cumbersome assumptions (Assumptions SA-DGP and SA-GL), we show that if $J^2\log^2(J)/n \to 0$ and $\log(n)J^{-1} \to 0$, then

$$\widehat{\mu}^{(v)}(x) - \mu_0^{(v)}(x) \;\approx_{\mathbb{P}}\; \widehat{\mathbf{b}}_s^{(v)}(x)'\bar{\mathbf{Q}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}_s(x_i)\eta_{i,1}\psi(y_i-\eta_{i,0}), \tag{4.5}$$

uniformly in $x \in \mathcal{X}$, where $\bar{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'\Psi_{i,1}\eta_{i,1}^2$ with $\Psi_{i,1} = \frac{\partial}{\partial\eta}\mathbb{E}[\psi(y_i;\eta)|x_i,\mathbf{w}_i]\big|_{\eta=\eta_{i,0}}$, and $\eta_{i,v} = \eta^{(v)}(\mu_0(x_i)+\mathbf{w}_i'\boldsymbol{\gamma}_0)$. This result substantially generalizes (3.1) under essentially the same rate restrictions previously imposed, with an error of approximation that is optimal up to $\log(n)$ terms.

Taking the uniform Bahadur representation (4.5) as the starting point, we establish analogous results to those discussed in Section 3 but now for generalized nonlinear binscatter.

1. Theorem SA-3.6 establishes a Nagar-type approximate IMSE expansion for $\widehat{\mu}^{(v)}(x)$ in (4.1), and shows that the resulting optimal choice of J is equivalent to $J_{\texttt{IMSE}}$ in Theorem 1 but with different bias and variance constants $\mathscr{B}_n(p,s,v)$ and $\mathscr{V}_n(p,s,v)$, which are substantially more complicated (as expected). Fortunately, Theorem SA-3.6 shows that $J_{\texttt{IMSE}}$ in Theorem 1 can always be used as a valid rule-of-thumb IMSE-optimal estimator of the number of bins for generalized nonlinear binscatter.

2. Theorem SA-3.3 establishes an analogue of Lemma 1, from which validity of pointwise confidence intervals follows (as in Theorem 2 for the case of least squares binscatter). As discussed in the supplemental appendix, estimation of the asymptotic variance is more complicated, and

our results there provide general high-level conditions justifying several alternatives commonly used in practice.

3. Theorems SA-3.4 and SA-3.5 establish strong approximations and valid uniform inference for $\widehat{\mu}^{(v)}(x)$ in (4.1). Those results provide all the necessary tools to establish analogues of Lemma 2 and Theorem 2 (confidence bands), Theorem 4 (parametric specification testing) and Theorem 5 (shape restriction testing). For both least squares and generalized nonlinear bisncatter, the latter three theorems are established formally in Sections SA-4.1, SA-4.2, SA-4.3, respectively, in more generality and under somewhat weaker assumptions.

4. Finally, Section SA-4.4 establishes strong approximation results for $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$, from which validity of confidence bands and hypothesis testing about parametric specification and shape restrictions follow directly. More precisely, Theorem SA-4.4 gives the foundational building block to analyze the uniform (in $x \in \mathcal{X}$) distributional features of t-statistics based on the estimator $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$, and therefore construct confidence bands for the estimand $\vartheta^{(v)}_{\mathbf{w}}(x)$ as well as conduct hypothesis tests about parametric specifications and/or shape restrictions of $\vartheta^{(v)}_{\mathbf{w}}(x)$, following the same methodology presented in Section 3.

4.2 Numerical Illustration

We illustrate the generalized nonlinear binscatter using the simulated data. In Figure 8(a) and (b), we plot two binscatter fits corresponding to the 0.25- and 0.75-conditional-quantile functions of y_i given x_i and the confidence bands thereof. To illustrate the binscatter logit estimation, we define a binary response variable $\mathbb{1}(y_i \leq y_{(\lfloor n/3 \rfloor)})$ where $y_{(\lfloor n/3 \rfloor)}$ is the 1/3-empirical-quantile of y_i for the full sample. Let $\rho(\cdot)$ and $\eta(\cdot)$ be the loss function and (inverse) link function corresponding to logistic regression. Then, we can estimate the conditional probability of the event $\{y_i \leq y_{(\lfloor n/3 \rfloor)}\}$ given x_i using the binscatter method (4.1). We restrict our attention to the subsample with $x \in [0.12, 0.5]$ (the grey shaded regions in Figure 8(a) and (b)), thus excluding the two tails where y_i is above the cutoff $y_{(\lfloor n/3 \rfloor)}$ with extremely high probability. Figure 8(c) and (d) plot these binscatter-based logistic fits and the corresponding confindence bands. Finally, in Figure 8(e) and (f), we plot the binscatter-based estimates of 0.75-conditional-quantile functions of y_i given x_i for two groups and the confidence bands thereof. Such comparison can be used to examine the heterogeneity of the

quantile treatment effect across the x_i dimension.

5 Conclusion

We presented a foundational study of binscatter, a very popular methodology for approximating the conditional expectation and other functions in applied microeconomics. Our results not only cover the canonical binscatter method commonly used in application, but also several extensions and generalizations including mean, quantile and other regression function estimation problems. Our theory allows for within-bin polynomial fits and across-bins smoothness restrictions, in both linear and nonlinear settings with and without principled semi-linear covariate adjustments. In particular, we highlighted important problems with the way covariate adjustment is currently done in practice via current popular software implementations (Stepner, 2017; Droste, 2019), and provide an alternative solution. More generally, we also clarify and discuss issues related to parameters of interest and associated hypothesis tests for practice, for both least squares and generalized nonlinear binscatter methods.

Our econometric methods for estimation, uncertainty quantification and hypothesis testing using binscatter methods offer an array of new tools for practice. For example, we offer valid confidence intervals and confidence bands, hypothesis testing about parametric specifications, shape constraints, and comparisons of multiple samples (treatment groups). All our methods are readily available in accompanying software (Cattaneo, Crump, Farrell, and Feng, 2021).

References

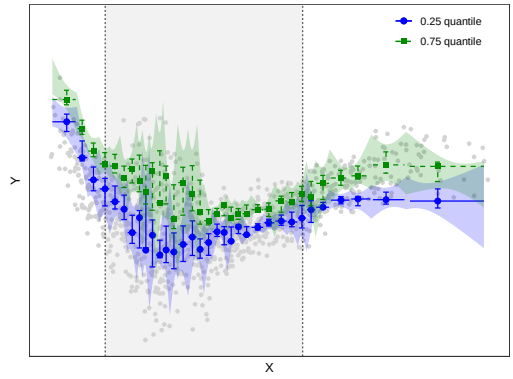
Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile Regression under Misspecification, with an Application to the US Wage Structure," *Econometrica*, 74(2), 539–563.

Angrist, J. D., and J. S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

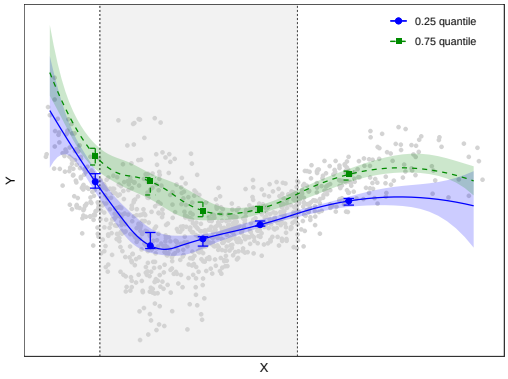
Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernandez-Val (2019): "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, 213(1), 4–29.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186(2), 345–366.

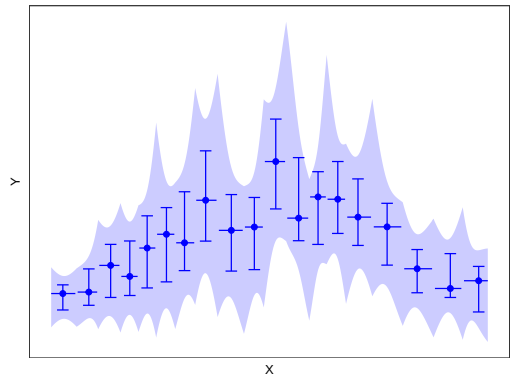
Figure 8: Generalized Nonlinear Binscatter Methods.



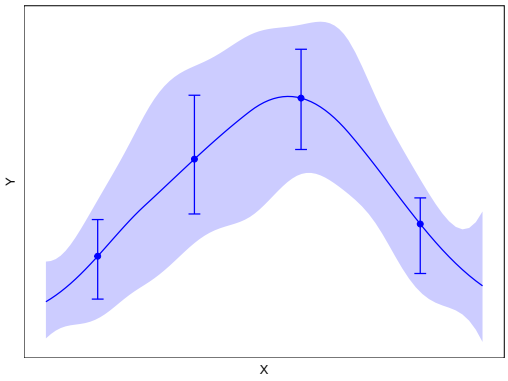
(a) Canonical Binscatter: $p = 0$ and $s = 0$



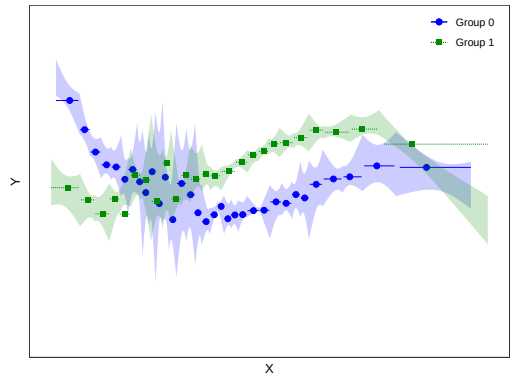
(b) Extended Binscatter: $p = 2$ and $s = 2$



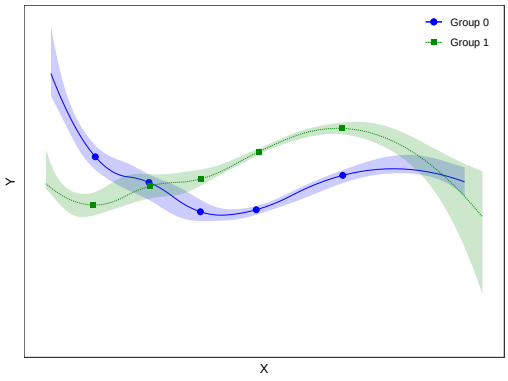
(c) Canonical Binscatter: $p = 0$ and $s = 0$



(d) Extended Binscatter: $p = 2$ and $s = 2$



(e) Canonical Binscatter: $p = 0$ and $s = 0$



(f) Extended Binscatter: $p = 2$ and $s = 2$

Notes. This figure reports robust, bias-corrected confidence intervals and confidence bands for the generalized nonlinear binscatter estimator. The left column reports binscatter estimators using $p = 0$ and $s = 0$ with associated confidence intervals (vertical capped spikes) and confidence bands (shaded region) using $p = 1$ and $s = 1$. The right column reports binscatter estimators using $p = 2$ and $s = 2$ with associated confidence intervals (vertical capped spikes) and confidence bands (shaded region) using $p = 3$ and $s = 3$. The top row displays the covariate-adjusted, quantile binscatter estimator evaluated at $\mathtt{q} = 0.25$ and $\mathtt{q} = 0.75$. The middle row displays the covariate-adjusted, logit binscatter estimator for $x \in [0.12, 0.50]$ (grey shaded regions between vertical dashed lines in (a) and (b)) and dependent variable $\mathbb{1}(y_i \leq y_{(\lfloor n/3 \rfloor)})$ where $y_{(\lfloor n/3 \rfloor)}$ is the 1/3-empirical-quantile of y_i for the full sample. The bottom row displays the covariate-adjusted, quantile binscatter estimator evaluated at $\mathtt{q} = 0.75$ for each of two subgroups. Constructed using simulated data described in Section SA-6 of the supplemental appendix. The grey dots represent the realized data. Nominal coverage for confidence intervals and confidence bands is 95%. The sample size is $n = 1,000$ for the one-sample results (top two rows) and $n = 2,000$ for the two-sample results (bottom row) with $n = 1,000$ observation in each subgroup.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113(522), 767–779.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

——— (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110(512), 1753–1769.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2021): "Binscatter Regressions," in preparation for the *Stata Journal*.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020): "Large sample properties of partitioning-based series estimators," *Annals of Statistics*, 48(3), 1718–1741.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014a): "Gaussian Approximation of Suprema of Empirical Processes," *Annals of Statistics*, 42(4), 1564–1597.

——— (2014b): "Anti-Concentration and Honest Adaptive Confidence Bands," *Annals of Statistics*, 42(5), 1787–1818.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011): "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics*, 126(2), 749–804.

Chetty, R., J. N. Friedman, and J. Rockoff (2014): "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 104(9), 2633–2679.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99(4), 1145–1177.

Chetty, R., and A. Szeidl (2006): "Marriage, Housing, and Portfolio Choice: A Test of Grossman-Laroque," Working Paper, UC-Berkeley.

Cochran, W. G. (1968): "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24(2), 295–313.

Droste, M. (2019): "`Stata` Module `Binscatter2`," https://github.com/mdroste/stata-binscatter2/.

Fama, E. F. (1976): *Foundations of Finance: Portfolio Decisions and Securities Prices*. Basic Books, New York, NY.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning*, Springer Series in Statistics. Springer-Verlag, New York.

Kleven, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

Nobel, A. (1996): "Histogram Regression Estimation Using Data-Dependent Partitions," *Annals of Statistics*, 24(3), 1084–1105.

Starr, E., and B. Goldfarb (2020): "Binned Scatterplots: A Simple Tool to Make Research Easier and Better," *Strategic Management Journal*, 41(12), 2261–2274.

Stepner, M. (2017): "`Stata` Module `Binscatter`," https://github.com/michaelstepner/binscatter/.

Tukey, J. W. (1961): "Curves As Parameters, and Touch Estimation," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 1, pp. 681–694.