On Binscatter*

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

February 25, 2019

Abstract

Binscatter is very popular in applied microeconomics. It provides a flexible, yet parsimonious way of visualizing and summarizing large data sets in regression settings, and it is often used for informal evaluation of substantive hypotheses such as linearity or monotonicity of the regression function. This paper presents a foundational, thorough analysis of binscatter: we give an array of theoretical and practical results that aid both in understanding current practices (i.e., their validity or lack thereof) and in offering theory-based guidance for future applications. Our main results include principled number of bins selection, confidence intervals and bands, hypothesis tests for parametric and shape restrictions of the regression function, and several other new methods, applicable to canonical binscatter as well as higher-order polynomial, covariate-adjusted and smoothness-restricted extensions thereof. In particular, we highlight important methodological problems related to covariate adjustment methods used in current practice. We also discuss extensions to clustered data. Our results are illustrated with simulated and real data throughout. Companion general-purpose software packages for `Stata` and `R` are provided. Finally, from a technical perspective, new theoretical results for partitioning-based series estimation are obtained that may be of independent interest.

Keywords: binned scatter plot, regressogram, piecewise polynomials, splines, partitioning estimators, nonparametric regression, robust bias correction, uniform inference, binning selection.

*We thank Raj Chetty, John Friedman, Andreas Fuster, Paul Goldsmith-Pinkham, Andrew Haughwout, Guido Imbens, David Lucca, Xinwei Ma, Emily Oster, Jesse Shapiro, Rocio Titiunik, Seth Zimmerman, Eric Zwick, and seminar participants at the University of Chicago, Chicago Booth, the Federal Reserve Bank of New York, University of Michigan, and Stanford University for helpful comments and discussion. Oliver Kim and Shahzaib Safi provided excellent research assistance. 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. The latest version of the `Stata` and `R` companion software is available at https://sites.google.com/site/nppackages/binsreg/.

†Department of Economics and Department of Statistics, University of Michigan.

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

§Booth School of Business, University of Chicago.

¶Department of Economics, University of Michigan.

1 Introduction

Binscatter is a flexible, yet parsimonious way of visualizing and summarizing large data sets in regression settings (Chetty and Szeidl, 2006; Chetty, Looney, and Kroft, 2009; Chetty, Friedman, Olsen, and Pistaferri, 2011; Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011). This methodology is also often used for informal (visual) evaluation of substantive hypotheses about shape features of the unknown regression function such as linearity, monotonicity, or concavity. Binscatter has gained immense popularity among empirical researchers and policy makers, and is 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 statistical procedures. Current practice employing binscatter is usually ad-hoc and undisciplined, which not only hampers replicability across studies, but also has the potential of leading to incorrect empirical conclusions.

This paper presents the first foundational study of binscatter. We provide several theoretical and practical results, which 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 particular nonparametric estimator of a regression function employing (possibly restricted) piecewise approximations in a semi-linear regression context. Thus, our first main contribution is to set up an econometrics framework to understand and analyze binscatter formally. This framework allows us to obtain an array of theoretical and practical results for canonical binscatter methods, and also to propose new complementary methods delivering more flexible and smooth approximations of the regression function, which still respect the core features of binscatter. The latter methods are particularly well suited for enhanced graphical presentation of estimated regression functions and confidence bands, and for formal testing of substantive hypotheses about the unknown regression function.

Furthermore, using our econometric framework, we highlight important methodological and theoretical problems with the covariate adjustment methods as commonly employed in practice, and propose a new alternative approach that is more generally valid and principled. To be more specific, we discuss the detrimental effects of the widespread practice of first "residualizing" additional covariates and then constructing the binscatter, and show how our proposed alternative covariate-

adjusted binscatter circumvents those problems.

The proposed econometric framework is then used to offer several new methodological results for binscatter applications. Specifically, our second main contribution is to 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. The third main contribution of this paper is to provide valid confidence intervals, confidence bands, and hypothesis testing procedures of both parametric specifications and nonparametric shape restrictions of the unknown regression function. These results not only give principled guidance for current empirical practice, but also offer new methods encoding informal (visual) assessments commonly found in empirical papers using binscatter.

Our paper offers additional contributions. From an implementation perspective, all the tools we develop are implemented in companion `R` and `Stata` software packages available at `https://sites.google.com/site/nppackages/binsreg/`, and further discussed in Cattaneo, Crump, Farrell, and Feng (2019). This new statistical software improves on current popular implementations in several ways, as we discuss below and in our companion software article. From a technical perspective, we obtain new theoretical results for semi-/non-parametric partitioning-based series regression, which were previously unavailable in the literature but needed to analyze popular implementations of binscatter and generalizations thereof.

The remainder of the paper proceeds as follows. Employing an empirical example throughout, Section 2 gives a gentle introduction to binscatter, overviews and illustrates numerically our main methodological results, and discusses related literature. This relatively long section is meant to be not only heuristic and empirically-driven, but also self-contained in terms of reviewing the methodology and contributions offered by our paper. On the other hand, the next three sections are more technical and precise: Section 3 introduces and formalizes binscatter, starting with its canonical form, then incorporating covariate-adjustment and within-bin higher-order polynomial fitting, and culminating with a smooth version based on imposing continuity restrictions at the boundaries of the bins; Section 4 gives formal results for empirical selection of the number of bins used to implement binscatter; and Section 5 presents our main theoretical results for estimation, inference, and graphical presentation, including shape-related testing procedures of substantive interest. Section

6 briefly discusses an extension to clustered data. Section 7 gives concrete recommendations for practice, which are illustrated empirically. Finally, Section 8 concludes. The supplemental appendix collects more general theoretical results, all technical proofs, and other methodological and numerical results.

2 Overview of Results

In this section we make clear what binned scatter plots are, how they are often used, and how our results can aid empirical practice. The treatment here is informal, but complete, drawing on the formal results presented in the upcoming (more technical) sections. Before detailing our new tools, it is important to define what a binned scatter plot is, and what it is not. See Chetty and Szeidl (2006, Figure 1) for one the very 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), and Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) for other early papers using binscatter methods. In addition, see Stepner (2014) for a widely used software implementation of canonical binscatter methods, and see Starr and Goldfarb (2018) for a very recent heuristic overview of canonical binscatter.

We illustrate our methods using data from the American Community Survey. In order to have full control on different features of the statistical model, this section employs a simulated dataset based on a data generating process constructed using the real survey dataset. But, in Section 7, we return to the original survey dataset to illustrate our main recommendations for practice using the actual data. The supplemental appendix (Section SA-5) details how the simulated data was constructed, and also includes two other distinct empirical illustrations.

The *scatter plot* itself is a classical tool for data visualization. It shows all the data on two variables y and x, and allows a researcher to visualize not only the relationship between y and x but also the variability, areas of high or low mass, and all other features of their joint distribution. However, in the era of large sample sizes, scatter plots are less useful. A scatter plot of a typical administrative data set with millions of observations yields a solid black cloud, and in particular, obscures the underlying relationship between y and x. This is where binning enters.

To construct a binned scatter plot one first divides the support of x into some number of bins,

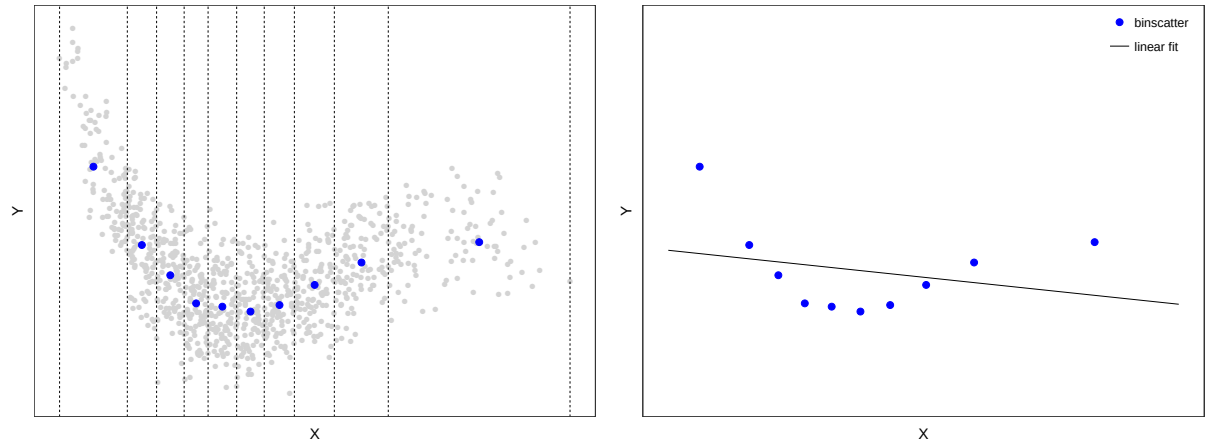
denoted herein by J. In most cases, ad-hoc choices such as $J = 10$ or 20 are most predominant in practice, with bins themselves divided at the empirical quantiles of observations of x (ignoring y), an approach we call quantile-spaced binning (or partitioning) herein. Then a single dot is placed in the plot representing the mean of y for the observations falling in each bin. The final plot consists of only these J dots, usually depicted at the center of each quantile-spaced bin. Often added is the regression line from a OLS fit to the underlying data. See Figure 1. It is typical in applications to "control" for covariates in both the regression line and the plot itself, which as discussed below, requires additional care.

The question is: what aspect of the data is being visualized with a binned scatter plot? This turns out to be not the data itself, but only the conditional expectation of y given x; the regression function. A binned scatter plot is nothing more than the fitted values of a particular nonparametric regression of y on x. This is not a disadvantage, indeed, we view it as the reverse: starting from this insight we can deliver a host of tools and results, both formal and visual, for binned scatter plots.

But it is nonetheless important to point out the limitations of what can be learned from a binned scatter plot. The plot *is not* a visualization of the whole data set in any meaningful way. That is, it is not at all analogous to a traditional scatter plot. Many different data sets can give rise to identical binned scatter plots, as in Figure 2. In particular, the variance (of y given x) is entirely suppressed. Figure 2 shows four different data sets, with different amounts of variance and heteroskedasticity, which nonetheless yield identical plots. This is not a new revelation, but it does seem to be the source of some confusion in practice. Indeed, Chetty, Friedman, and Rockoff (2014, p. 2650) note "that this binned scatter plot provides a nonparametric representation of the conditional expectation function but does not show the underlying variance in the individual-level data." To show the underlying variance from a large data set, one can plot a small random sample of the data. For a large data set, this is perhaps most akin to a traditional scatter plot. However, the sample may need to be small enough as to render the conclusions unreliable, and further, given our results, this is not necessary in most cases.

Turning back to how binscatter plots are constructed, in this paper we analyze these plots from an econometric point of view. This allows us not only to formalize its properties, but also develop new tools for empirical practice. These include a valid and optimal choice for the number of bins,

Figure 1: The basic construction of a binned scatter plot.



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

Notes. The data is divided into $J = 10$ bins according to the observed x. Within each bin a single dot is plotted at the mean of y 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 on x. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

valid confidence intervals and bands reflecting the true uncertainty in the data, and formal (and visual) assessment of substantive hypotheses of interest, such as whether the relationship between y and x is monotonic, or of a particular parametric form such as linear, or different between two groups. Here we give only an introduction to these ideas; formal details are spelled out below and everything is implemented in our companion software (Cattaneo, Crump, Farrell, and Feng, 2019).

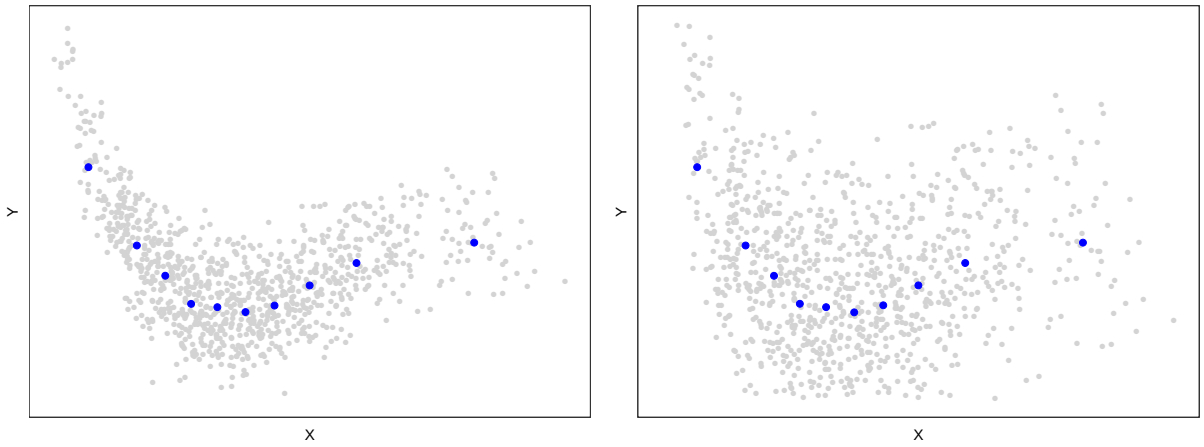
The canonical nonparametric regression model is

$$y_i = \mu(x_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | x_i] = 0, \tag{2.1}$$

where (y_i, x_i), $i = 1, 2, \ldots, n$, is random sample from (y, x). We are interested in the function $\mu(x)$ and its properties. For example, we might like to know if it is (well-approximated by) a linear function, i.e. $\mu(x) = \theta_0 + x\theta_1$, or quadratic, i.e. $\mu(x) = \theta_0 + x\theta_1 + x^2\theta_2$. This is implicitly behind plots of the form of Figure 1: we want to assume that the linear approximation is sound enough that conclusions from a OLS regression of y on x are useful for policy.

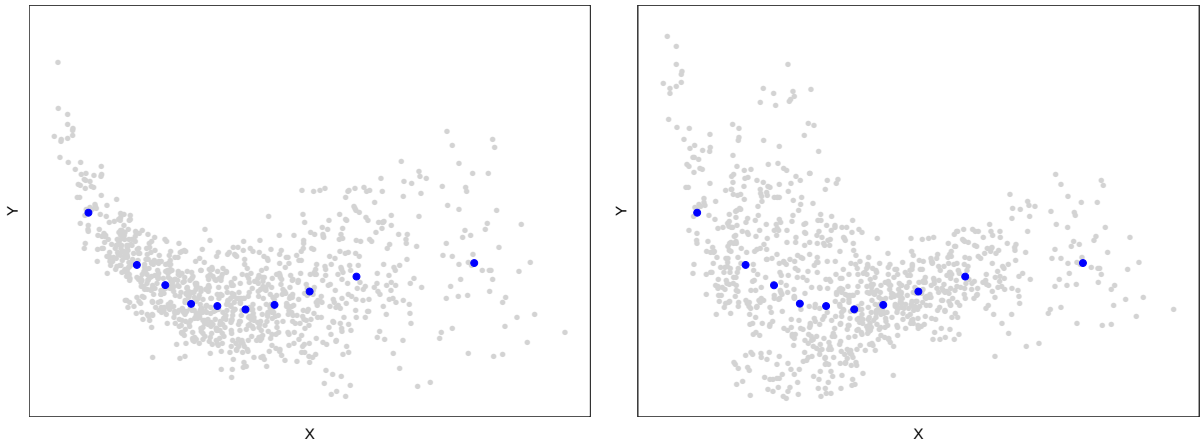
Binned scatter plots estimate the unknown regression function $\mu(x)$ by exploiting the fact that $\mu(x_1) \approx \mu(x_2)$ if $x_1 \approx x_2$. Broadly speaking, all nonparametric regression exploits this same idea. For binscatter regressions, "$x_1 \approx x_2$" is translated as being in the same bin, and then further,

Figure 2: Scatter and Binscatter Plots with Different Variability.



(a) Original Dataset

(b) Dataset with More Variability

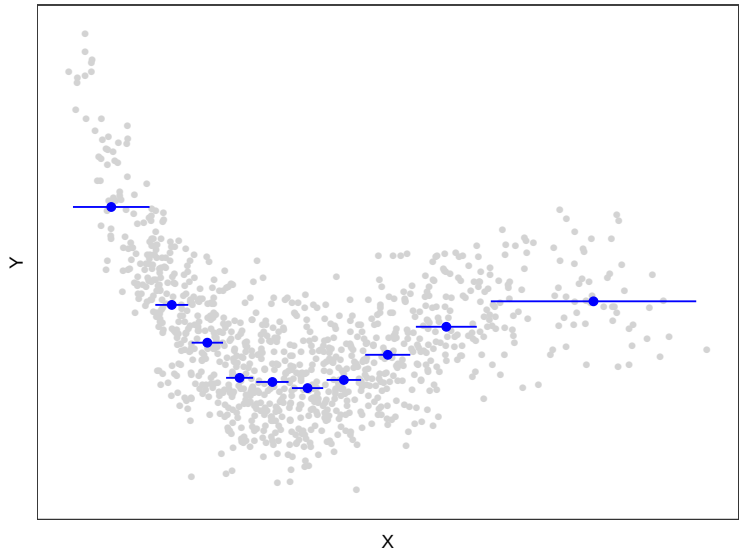


(c) Dataset with Moderate Heteroskedasticity

(d) Dataset with High Heteroskedasticity

Notes. Four simulated different data sets, each with different variance of y given x, but identical binned scatter plots. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

Figure 3: The actual regressogram nonparametric estimator corresponding to a binned scatter plot.



(a) Binned Scatter Plot with Piecewise Constant Fit

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

the estimator sets $\hat{\mu}(x) = \bar{y}_j$ for all x in the j-th bin, $j = 1, 2, \ldots, J$, where $\bar{y}_j$ denotes the sample average of the y_i's with x_i's in that j-th bin. This results in a piecewise constant estimate, as shown in Figure 3. A typical binned scatter plot shows only one point within each bin, but it is important to observe that a binned scatter plot is *equivalent* to this piecewise constant fit, however unfamiliar it may look. As a way of contrast, a traditional kernel regression is distinct almost everywhere from the canonical binscatter, coinciding for a very special implementation and then only at J points: at the center of each bin and employing the uniform kernel with bandwidth equal to half the block length both procedures will yield the same fitted values, but only at these J points. To "fill in" the rest of the regression curve, traditional kernel regression rolls out the window, implying new bandwidths and associated new "bins", distinct almost everywhere from those used to form canonical binscatter.

Despite its appearance, piecewise constant fits over pre-determined quantile-spaced bins is not a "bad" nonparametric estimation method in any sense, when implemented properly it shares many favorable theoretical properties with more familiar methods such as traditional kernel smoothing and, in fact, they are the building block for popular spline approximations. 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; Cattaneo and Farrell, 2013; Cattaneo, Farrell, and Feng, 2018). The use of binned scatter plots in applied economics is most closely related to this strand of literature, and our theory below can be thought of as a generalization and complete formal treatment of the regressogram. 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 subclassification (Cochran, 1968; Cattaneo and Farrell, 2011), and for visualization in regression discontinuity designs (Calonico, Cattaneo, and Titiunik, 2015) and bunching designs (Kleven, 2016); in empirical finance it is related to portfolio sorting (Fama, 1976; Cattaneo, Crump, Farrell, and Schaumburg, 2019); and in machine learning it is at the heart of regression trees and similar methods (Friedman, 1977; Hastie, Tibshirani, and Friedman, 2009). We do not address these other applied contexts directly here, as each is different enough to require a separate analysis, but our results and tools can be adapted and exported to those other settings.

This paper offers three main methodological contributions to the understanding and correct use of binscatter methods in applied microeconomics, which we summarize and illustrate in the remaining of this overview section. In closing this section, we also mention briefly some other contributions related to software and theory.

Contribution 1: Framework and Construction

Understanding binscatter requires formalizing it in a principled way. Thus, our first contribution is to outline a framework that not only correctly incorporates additional covariates, and gives the baseline for further extensions to clustered data, but also permits us to introduce more flexible polynomial regression approximations within bins as well as to incorporate smoothness restrictions across bins. These extensions are particularly useful in applications because it is common for researchers both to control for additional factors in their regression specifications and to prefer more "smooth" global approximations and confidence bands, in combination with canonical binscatter.

In Section 3, we first recast canonical binscatter as a very basic nonparametric least squares regression procedure, and then extend it to incorporate additional covariate adjustments and several other features. Adjusting for additional covariates is standard in applications, and to formalize it

we extend model (2.1) to include a vector of controls, $\mathbf{w}$, as follows:

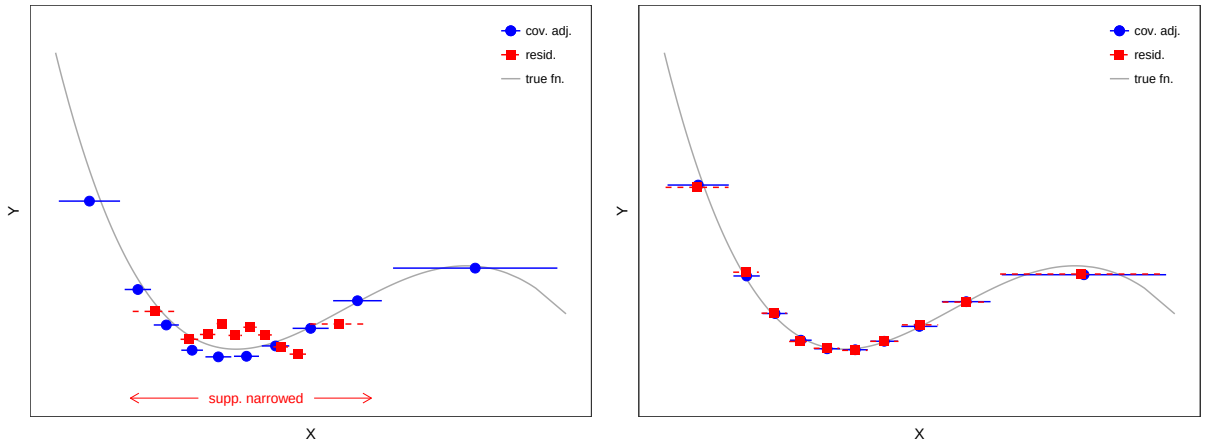
$$y_i = \mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma} + \epsilon_i, \qquad \mathbb{E}[\epsilon_i | x_i, \mathbf{w}_i] = 0, \tag{2.2}$$

where $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is random sample from $(y, x, \mathbf{w})$. The object of interest, both visually and numerically, is still the function $\mu(x)$. This regression model, variously referred to as partially linear, semi-linear, or semiparametric, retains the interpretation familiar from linear models of "the effect of x on y, controlling for $\mathbf{w}$". Here, we control for $\mathbf{w}$ in an additively-separable way under random sampling, but in Section 6 we discuss possible extensions of our work to clustered or grouped data.

The regression model (2.2) justifies a particular way of covariate adjustment, which is not the way encountered in practice: see Section 3.3 for a detail comparison and discussion. In particular, if $\mu(x)$ is not linear, then standard Frisch-Waugh logic does not apply: one cannot estimate (or binned scatter plot) the function $\mu(x)$ using the residuals from least squares regressions of y on $\mathbf{w}$ and x on $\mathbf{w}$. This highlights an important methodological issue with most current applications of covariate-adjusted binscatter, since it is often the case that practitioners first regress out the additional covariates $\mathbf{w}$ and only after construct the binscatter based on the resulting residuals. The latter approach, which differs from our proposed method for covariate-adjustment, can lead to very different empirical findings. In this paper, we refer to the latter approach as (canonical, covariate-adjusted) residualized binscatter.

We illustrate this issue of covariate adjustment numerically in Figure 4. The true regression function, $\mu(x)$, 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). Our recommended method implements binscatter via model (2.2), while residualized binscatter implements binscatter via model (2.1) after replacing y_i and x_i by the residuals obtained from regressing y_i on $\mathbf{w}_i$ and regressing x_i on $\mathbf{w}_i$, respectively. As Figure 4 clearly indicates, the two approaches for covariate adjustment lead to quite different results if x and $\mathbf{w}$ are correlated. The reason is simple: our approach is valid for model (2.2), while residualized binscatter is invalid in general. Figure 4(a) shows that residualized binscatter is unable to correctly approximate the true function of interest $\mu(x)$, while our semi-linear covariate-adjustment approach works well.

Figure 4: Comparison of Covariate Adjustment Approaches.



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

Notes. Two plots comparing semi-linear covariate-adjustment and residualized covariate adjustment for binscatter. Plot (a) illustrates the biases introduced by residualization when there is non-zero correlation between x and the other covariates $\mathbf{w}$ controlled for. Plot (b) shows that the residualization biases are not present when x and $\mathbf{w}$ are independent, and the location of binscatter is adjusted: see Section 3.3 for more details. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

This numerical illustration relies on data generated as in model (2.2), but even when the true regression function of y_i given $(x_i, \mathbf{w}_i')$ does not satisfy the semi-linear structure, our approach to covariate adjustment retains a natural interpretation as a "best" semi-linear approximation in mean square, just as it occurs with simple least squares methods (e.g., Angrist and Pischke, 2008, for more discussion), while residualized binscatter would be fundamentally misspecified and uninterpretable in such case. To put this another way, in the case when the true $\mu(x)$ is nonlinear, then the conclusions reached from the currently dominant binscatter approach are incompatible with the often-presented table of results from a regression of y_i on x_i and $\mathbf{w}_i$. While such dominant approach is not completely "wrong" in all cases, it does not match how the results are usually interpreted. See Section 3.3 for more technical details and discussion on our recommended approach to covariate adjustment vis-á-vis residualized binscatter.

In addition to incorporating covariate adjustments in an appropriate and interpretable way, our proposed framework allows us to introduce new, related binscatter procedures. In particular, we consider two useful extensions for empirical work: fitting a p-th order polynomial regression within each bin and imposing s-th order smoothness restrictions across bins, both with and without covariate adjustments. These generalizations of binscatter are exhibited in Figure 5. 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 user-choices p and s control flexibility and smoothness from a local and global perspectives, respectively. For example, if $p = 1$, then $s = 1$ corresponds to piecewise linear fits that are forced to be connected at the bin's boundaries: a continuous but not differentiable global fit based on binscatter. This is illustrated in Figure 5(b). Of course, removing the smoothness constraint ($s = 0$) leads to piecewise linear fits within bins ($p = 1$) that need not to agree at the bins' boundaries: Figure 5(a). An example of within-bin quadratic fit ($p = 2$) without smoothness constrain ($s = 0$) is given in Figure 5(c), while imposing only continuity at the bins' edges ($s = 1$) for the quadratic case is depicted in Figure 5(d). A within-bin quadratic fit ($p = 2$) with continuously differentiable restrictions at bins' boundaries ($s = 2$) is not depicted to conserve space, but follows the same logic.

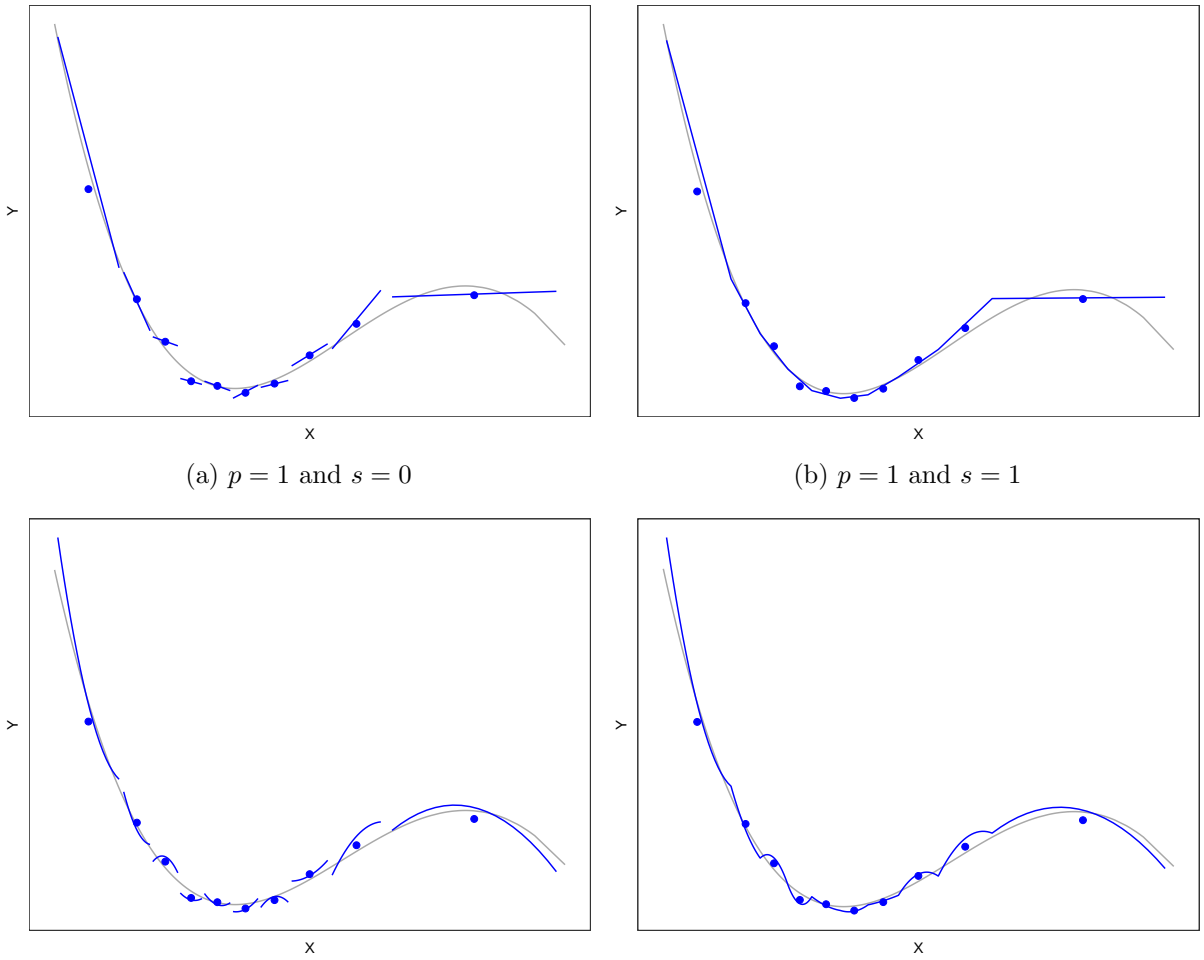
This generalization of binscatter can be implemented with or without covariate adjustment, as discussed in Section 3, and can also be further extended to clustered data, as discussed in Section 6. It should be clear that canonical binscatter corresponds to the specific choice $p = s = 0$ (Figure 3), but one can consider more or less smooth versions thereof by appropriate choice of $s \leq p$. Another advantage of considering $p > 0$ polynomial fits, with or without covariate adjustments and/or smoothness restrictions, is that enables approximating the derivatives $\mu^{(v)}(x) = \frac{d^v}{dx^v}\mu(x)$: estimating derivatives of the regression function $\mu(x)$ is crucial for testing shape features such as monotonicity or concavity, as we discuss further below.

Employing our econometrics framework, we obtain an array of methodological results for canonical binscatter and its generalizations, which we summarize and illustrate next.

Contribution 2: Valid and Optimal Number of Bins Selection

Implementing our standard binned scatter plot requires one choice: the number of bins to be used, J. Given a choice of J, the position of the bins is set by the empirical quantiles of x, via the quantile-spaced binning used in all applications. Because the bins positions are determined by estimated quantiles, the random binning structure underlying binscatter introduces some additional technical issues. Nevertheless, in Section 4, we employ our formalization of binscatter to view the choice of J as that of a tuning parameter in nonparametric estimation, just as a bandwidth is the tuning

Figure 5: Binscatter Generalizations.



(a) $p = 1$ and $s = 0$

(b) $p = 1$ and $s = 1$

(c) $p = 2$ and $s = 0$

(d) $p = 2$ and $s = 1$

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

parameter in kernel regression. As such, it reflects a bias/variance trade-off: as J increases the bias decreases but the variability of the estimator increases. This is depicted in Figure 6.

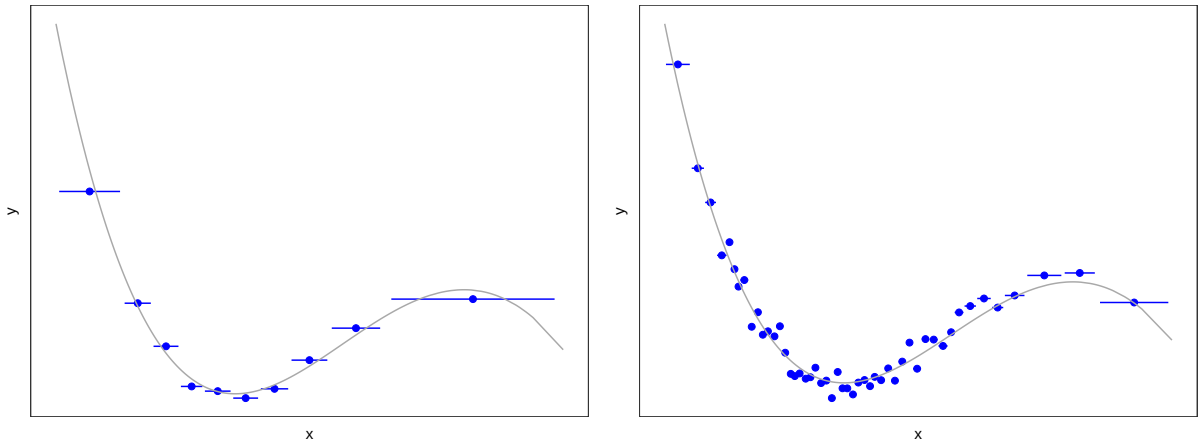
Our second main contribution is to give a precise choice of the number of quantiles J that trades off bias and variance in a valid and optimal way. Specifically, we study the asymptotic properties of the integrated mean square error (IMSE) of binscatter and its generalizations, and show that an IMSE-optimal choice of J is always proportional to $n^{\frac{1}{2p+3}}$, up to rounding to the closest integer, where recall p denotes the order of polynomial fit within each bin. For example, if a constant fit is used (i.e., the canonical binscatter), as in Figure 3, then the optimal choice of number of bins is $J \propto n^{1/3}$. The role of covariate adjustment, smoothness restrictions across bins, and other related features of binscatter, show up only through the constant of proportionality in the optimal rule for J. For implementation, we make the optimal choice of J, including its constant, fully data-driven and automatic, and readily available for empirical work in our companion software.

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 non-linear relationship $\mu(x)$ (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 "too large", leading to graphical and formal procedures with invalid or at least unreliable statistical properties. Section 4 presents our formal approach to this problem, where we rely on an objective measure (IMSE) to select in a data-driven way the number of bins J to use in each application of binscatter.

Contribution 3: Confidence Bands and Valid Inference

Armed with an IMSE-optimal estimator of the regression function we now turn to inference. Binned scatter plots are often used in applications to guide subsequent regression analyses, essentially as a visual specification check. A second common usage is to visually assess economically meaningful properties such as monotonicity or concavity. Our results allow for a valid assessment, both visually and formally, of these questions, as well as faithful display of the variability in the outcome y in the underlying data set. None of these are possible with a traditional scatter plot nor are currently

Figure 6: Number of Bins (J).



(a) $p = 0$, $s = 0$, $J = 10$ (b) $p = 0$, $s = 0$, IMSE-optimal J

Notes. The plots illustrate the potential effects on binscatter of choosing the number of bins J too small vis-á-vis in an IMSE-optimal way. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

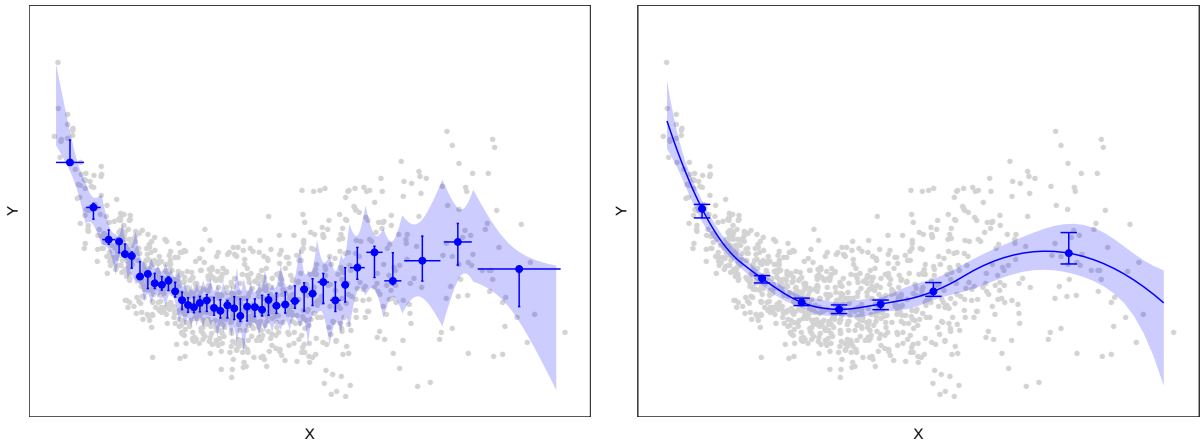
available in the literature for binscatter and its generalizations.

The first, most intuitive display of these results is a confidence *band*. One may, for each bin $j = 1, 2, \ldots, J$, place a standard confidence interval around the sample mean $\hat{\mu}(x) = \bar{y}_j$. However, this is not a correct visualization of the uncertainty about $\mu(x)$ in the data set, and as such, can not be used to assess hypotheses of interest. For example, just because one cannot fit a line through all these intervals does not allow a researcher to conclude that $\mu(x)$ is nonlinear. A confidence band is the tool required here, which naturally generalizes the idea of confidence interval.

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 $\mu(x)$ is linear (or quadratic). Visually, the size of the band reflects the uncertainty in the data, both in terms of overall uncertainty and any heteroskedasticity patterns. Figure 7 shows a confidence bands for the same four data sets as in Figure 2, and we see that the size and shape of the band reflects the underlying data.

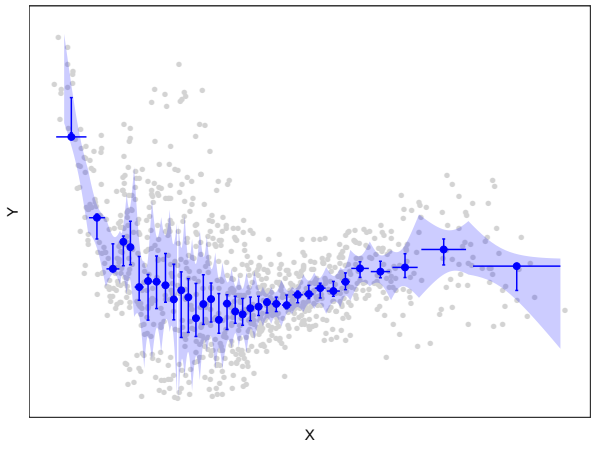
We can use confidence bands, and associated statistical procedures, to test for a variety of substantive hypotheses, both for guiding further analysis and for economic meaning directly. Figure 8 shows two examples: the left plot shows a rejection of linearity while the right plot indicates

Figure 7: Confidence Intervals and Confidence Bands.

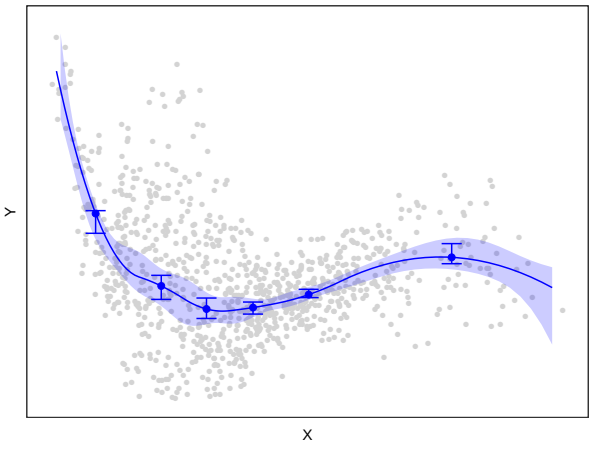


(a) $p = 0$ and $s = 0$

(b) $p = 2$ and $s = 2$



(c) $p = 0$ and $s = 0$



(d) $p = 2$ and $s = 2$

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

statistically significant group differences. Given the left result, a researcher may consider nonlinear regression modeling in the empirical analysis. Given the right plot, we conclude that the different relationship between y and x is different between the two groups shown, which may be of substantive interest in its own right. This is a nonparametric analogue of testing the significance of the interaction between x and a group dummy in a linear model. In this paper, we formalize this kind of test, which we refer to as parametric specification testing because a particular parametric specification for $\mu(x)$, namely the linear-in-parameters model $\mu(x) = \theta_0 + \theta_1 x$ is contrasted against the binscatter approximation of $\mu(x)$. Of course, we can test for any given parametric functional form for $\mu(x)$, including examples such as the Probit model $\mu(x) = \Phi(\theta_0 + \theta_1 x)$ or the log-linear model $\mu(x) = e^{\theta_0 + \theta_1 x}$. Visually, we cannot reject a certain functional form if the confidence band contains a function of that type. Numerically, we can give a precise p-value for such a test.

Heuristically, our formal parametric specification testing approach is based on comparing the maximal empirical deviation between binscatter and the desired parametric specification for $\mu(x)$. If the parametric specification is correct, then there should no deviation beyond what is explained by random sampling for all evaluation points x; hence the connection with the confidence band for $\mu(x)$. The first three rows of Table 1 illustrate our approach numerically.

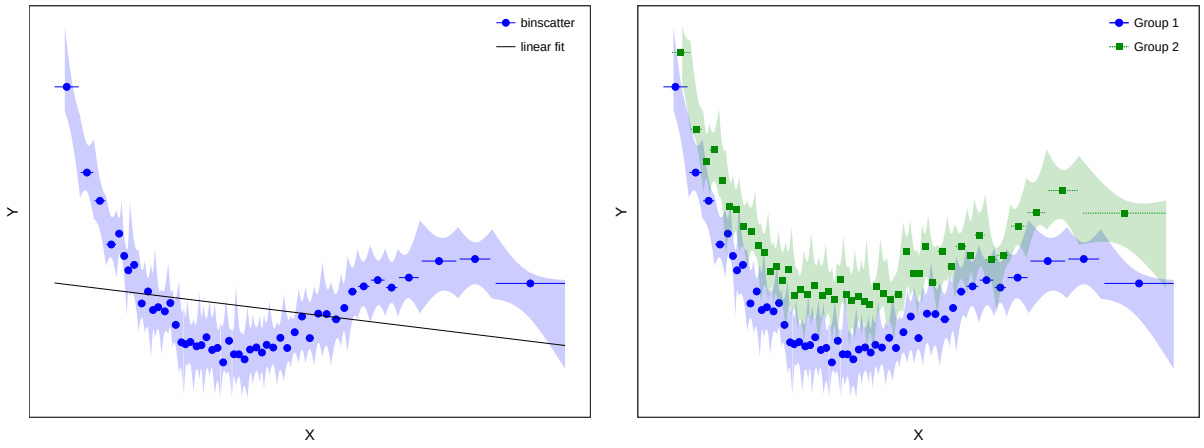
In addition to parametric specification testing, we also develop graphical and formal testing procedures for nonparametric shape restrictions of $\mu(x)$. Prime examples of such tests include negativity, monotonicity or concavity of $\mu(x)$, and their reciprocals positivity and convexity, or course. Graphically, this can be tested as before: using Figure 8 again we can assess whether $\mu(x)$ is "likely" to be monotonic, concave or positive, say. Formally, we can test all these features as a one-sided hypothesis test on $\mu(x)$ or its derivatives. To be more precise, negativity means $\mu(x) \leq 0$, monotonicity means $\mu^{(1)}(x) \leq 0$, and concavity means $\mu^{(2)}(x) \leq 0$. The second three rows of Table 1 illustrate our approach to shape restriction testing numerically.

Precise results for estimation and inference, including regularity conditions and other technicalities, are summarized in Section 5 and discussed in detail in the supplemental appendix.

Other Contributions: Software and Technicalities

Stepner (2014) gives an introduction to a very popular `Stata` software package implementing binscatter. This package implements canonical binscatter and residualized (covariate-adjusted) bin-

Figure 8: Graphical Testing of Substantive Hypotheses.

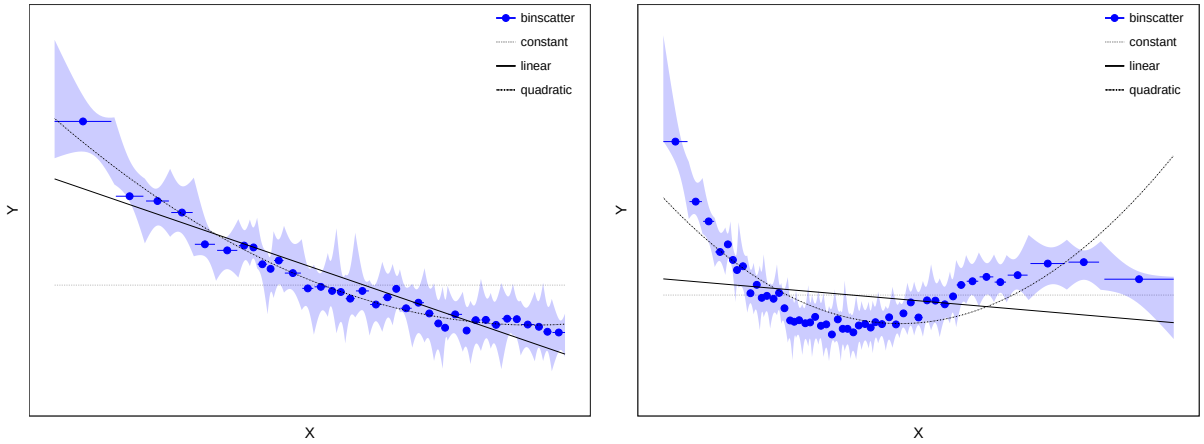


(a) Linear Regression Fit vs. Binscatter

(b) Two-sample Comparison

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

Figure 9: Graphical Representation of Parametric Specification Testing (Table 1)).



(a) Half Support of x

(b) Full Support of x

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

Table 1: Formal Testing of Substantive Hypothesis.

	Half Support ($n = 482$)			**Full Support** ($n = 1000$)		
	Test Statistic	P-value	J	Test Statistic	P-value	J
Parametric Specification						
Constant	21.761	0.000	37	22.680	0	50
Linear	8.968	0.000	37	20.433	0	50
Quadratic	4.478	0.000	37	44.650	0	50
Shape Restrictions						
Negativity	187.185	0.000	37	188.414	0	50
Decreasingness	0.339	0.996	6	6.149	0	11
Concavity	6.009	0.000	3	8.976	0	5

Notes. Constructed using simulated data described in Section SA-5 of the supplemental appendix.

scatter. Accompanying this paper, we provide new software packages in `Stata` and `R`, which improve on current software implementations in several directions. First, we implement polynomial fits within bins and smoothness restrictions across bins for binscatter, and hence consider estimation and inference for both the regression function and its derivatives. Second, we implement fully data-driven selections of J, the number of bins, reflecting the features of the underlying data. Third, we implement covariate adjustments as discussed above, avoiding residualization, which leads to valid and interpretable methods for practice. Fourth, we implement valid distributional approximations leading to confindence intervals, confidence bands, and a wide range of parametric specification and nonparametric shape restriction hypothesis tests. Cattaneo, Crump, Farrell, and Feng (2019) discusses all the details concerning our accompanying software and further illustrates it.

Finally, while not the focus on our paper, it is fair to underscore that studying in full generality standard empirical practice using binscatter forced us to develop new technical results that may be of independent interest. Our theoretical work is connected to the literature on nonparametric series estimation because binscatter is a partitioning-based nonparametric least squares estimator (e.g., Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2019; Cattaneo and Farrell, 2013; Cattaneo, Farrell, and Feng, 2018, and references therein), and to the literature on partially linear semiparametric regression because of the way we incorporate covariate adjustments (e.g., Cattaneo, Jansson, and Newey, 2018a,b, and references therein). However, available technical results can not be used to analyze binscatter because it

is implemented with a quantile-spaced binning, an example of random partitioning, generated by estimated quantiles.

As a consequence, our theoretical work necessarily relies on new results concerning non-/semi-parametric partitioning-based estimation on quantile-spaced (data-driven) partitions, which may be of independent interest. To be specific, we establish three main set of new theoretical results. First, we formally handle quantile-spaced (random) partitions underlying binscatter by resorting to appropriate empirical process techniques, substantially extending the results in Nobel (1996). Second, we obtain a general characterization of a linear map between piecewise polynomials and B-splines and give several technical results for it, properly accounting for quantile-spaced binning. Third, we develop 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 stochastic process in Chernozhukov, Chetverikov, and Kato (2014a,b) and on a conditional coupling lemma in Cattaneo, Farrell, and Feng (2018, Lemma 8.2). Since this paper is purposely practical, we relegate most discussions on our underlying technical work to the supplemental appendix, unless it is strictly necessary for practical implementation or methodological interpretation of binscatter.

3 Formalizing Binscatter

We now begin our formal econometric-theoretical treatment of binscatter. Canonical binscatter builds on the standard regression model (2.1), 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 constructed employing the empirical quantiles of x_i, leading to the partitioning scheme $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \ldots, \widehat{\mathcal{B}}_J\}$, where

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

$x_{(i)}$ 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 in the x_i dimension.

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

$$\widehat{\mu}(x) = \widehat{\mathbf{b}}(x)'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta})^2, \tag{3.1}$$

where

$$\widehat{\mathbf{b}}(x) = \begin{bmatrix} \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) & \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) & \cdots & \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \end{bmatrix}',$$

is the binscatter basis given by a J-dimensional vector of orthogonal dummy 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 y_i, one for each bin $\widehat{\mathcal{B}}_j$: $\bar{y}_j = \frac{1}{N_j}\sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i) y_i$ for $j = 1, 2, \ldots, J$. As illustrated in Section 2, empirical work employing canonical binscatter typically plots these binned sample averages along with some other estimate of the regression function $\mu(x)$.

3.1 Polynomial and Covariate Adjustments

We investigate the properties of binscatter in more generality. First, we allow for a more flexible polynomial regression approximation within each bin $\widehat{\mathcal{B}}_j$ forming the partitioning scheme $\widehat{\Delta}$, and thus expand the binscatter basis to allow for p-th order polynomial fits within each bin. For a choice of $p = 0, 1, 2, \ldots$, we redefine

$$\widehat{\mathbf{b}}(x) = \begin{bmatrix} \mathbb{1}_{\widehat{\mathcal{B}}_1}(x) & \mathbb{1}_{\widehat{\mathcal{B}}_2}(x) & \cdots & \mathbb{1}_{\widehat{\mathcal{B}}_J}(x) \end{bmatrix}' \otimes \begin{bmatrix} 1 & x & \cdots & x^p \end{bmatrix}',$$

where now the binscatter basis is of dimension $(p+1)J$. Setting $p = 0$ restores canonical binscatter. This generalization allows us to consider two important extensions of binscatter: (i) estimating derivatives of $\mu(\cdot)$, and (ii) incorporating smoothness restrictions across bins. Both will be very useful in Section 5 when we develop novel smooth confidence band estimators and formal hypothesis tests for shape restrictions.

Our second generalization of binscatter concerns covariate adjustment. As discussed in Section

2, we allow for additive separable covariate regression-based adjustment. Given the quantile-spaced partitioning scheme already introduced and a choice of p-th order polynomial approximation within bin, our proposed covariate-adjusted binscatter estimator is

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad v \leq p, \tag{3.2}$$

using the standard notation $\mathbf{g}^{(v)}(x) = d^v\mathbf{g}(x)/dx^v$ for a function $\mathbf{g}(x)$ and $\mathbf{g}(x) = \mathbf{g}^{(0)}(x)$. The partially linear structure of model (2.2) naturally justifies our way of covariate adjustment, and sharply contrasts with the most common approach based on least squares residualization. See Section 3.3 below for more details.

The generalized binscatter (3.2) reduces to the canonical binscatter (3.1) when $p = 0 = v$ and $\boldsymbol{\gamma} = \mathbf{0}_d$, in which case $\widehat{\mu}(x) = \widehat{\mu}^{(0)}(x)$ becomes an step function (Figure 3) reporting the sample averages $\bar{y}_j$ according to whether $x \in \widehat{\mathcal{B}}_j$, $j = 1, 2, \ldots, J$. The generalized binscatter $\widehat{\mu}^{(v)}(x)$ is useful to formalize commonly used empirical procedures based on binscatter, and to develop new binscatter-based estimation and inference procedures with better theoretical and practical properties.

3.2 Smoothness Restrictions

The binscatter estimator $\widehat{\mu}^{(v)}(x)$ retains the main features of canonical binscatter: estimation is conducted using only information within each (estimated) bin forming the quantile-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 $\widehat{\Delta}$; see Figure 5. For some empirical analyses, both graphical and formal, researchers prefer a more smooth binscatter of $\mu(\cdot)$, where the fits within each bin are constrained so that the final estimator exhibits some overall smoothness over the support of x_i. In this section we further generalize binscatter to provide such an alternative.

Given the quantile-spaced partitioning scheme, a smooth binscatter is the p-th order polynomial, s-times continuously differentiable, covariate-adjusted estimator given by

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}_s(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad s \leq p, \tag{3.3}$$

where $\widehat{\mathbf{b}}_s(x) = \widehat{\mathbf{T}}_s\widehat{\mathbf{b}}(x)$ with $\widehat{\mathbf{T}}_s$ being 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. When $s=0$, $\widehat{\mathbf{T}}_0 = \mathbf{I}_{(p+1)J}$, the identity matrix of dimension $(p+1)J$, and therefore no restrictions are imposed: $\widehat{\mathbf{b}}(x) = \widehat{\mathbf{b}}_0(x)$, as given in the previous subsection. Consequently, if $s=0$, we obtain the binscatter $\widehat{\mu}(x)$, which is not a continuous function estimate. On the other hand, $p \geq s$ implies that a least squares p-th order polynomial fit is constructed within each bin $\widehat{\mathcal{B}}_j$, in which case setting $s=1$ forces these fits to be connected at the boundaries of adjacent bins, $s=2$ forces these fits to be connected and continuously differentiable at the boundaries of adjacent bins, and so on, for $s=3,4,\ldots,p$. This is the formalization leading to Figure 5.

Enforcing smoothness for binscatter boils down to incorporating restrictions on the binscatter basis. The resulting constrained basis, $\widehat{\mathbf{b}}_s(x)$, corresponds to a choice of spline basis for approximation of $\mu(\cdot)$, 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 have very good finite sample properties, but other choices are of course possible. Smooth binscatter (3.3) reduces to binscatter (3.2) when $s=0$, and therefore the former is a strict generalization of latter and hence, in particular, of the canonical binscatter (3.1).

3.3 Comparison to the Canonical Residualized Binscatter

Current widespread empirical practice for covariate adjustment of binscatter proceeds by first regressing out the covariates $\mathbf{w}_i$, and then applying canonical binscatter on the residualized variables of interest. To be precise, standard practice applies (3.1) upon replacing y_i by $y_i - \widetilde{\mathbf{w}}_i'\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and x_i by $x_i - \widetilde{\mathbf{w}}_i'\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$, where $\widetilde{\mathbf{w}}_i = (1,\mathbf{w}_i')'$, and $\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and $\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$ denote the OLS coefficients obtained from regressing y on $\mathbf{w}$ and x on $\mathbf{w}$, respectively, with each regression including a constant term. This is the default (and only) implementation of covariate adjustment in standard binscatter software widely used in practice (Stepner, 2014).

Under mild assumptions, the estimators $\widehat{\boldsymbol{\delta}}_{y.\widetilde{w}}$ and $\widehat{\boldsymbol{\delta}}_{x.\widetilde{w}}$ are consistent for $\boldsymbol{\delta}_{y.\widetilde{w}} = \mathbb{E}[\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i']^{-1}\mathbb{E}[\widetilde{\mathbf{w}}_i y_i]$ and $\boldsymbol{\delta}_{x.\widetilde{w}} = \mathbb{E}[\widetilde{\mathbf{w}}_i\widetilde{\mathbf{w}}_i']^{-1}\mathbb{E}[\widetilde{\mathbf{w}}_i x_i]$, respectively. As it is customary in applied work, $\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}$ and $\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{x.\widetilde{w}}$ can be interpreted as a "best" linear approximation to $\mathbb{E}[y|\mathbf{w}]$ and $\mathbb{E}[x|\mathbf{w}]$, respectively. It can be argued that, under non-trivial assumptions, the residualized binscatter approximates the conditional expectation $\mathbb{E}[y-\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}|x-\widetilde{\mathbf{w}}'\boldsymbol{\delta}_{x.\widetilde{w}}]$, a parameter that is quite difficult to interpret. Consequently,

as illustrated in Figure 4, residualized binscatter does not consistently estimate $\mu(x)$ in model (2.2), nor $\mathbb{E}[y_i|x_i]$ in general. Under additional restrictive assumptions, the probability limit of residualized binscatter does have some interpretation when model (2.2) holds: if x and $\mathbf{w}$ are uncorrelated, then $\boldsymbol{\delta}_{x.\widetilde{w}} = (\mathbb{E}[x], \mathbf{0}')'$, and the residualized binscatter procedure consistently estimates

$$\mathbb{E}[y - \widetilde{\mathbf{w}}'\boldsymbol{\delta}_{y.\widetilde{w}}|x - \mathbb{E}[x]] = \mu(x) - \mathbb{E}[y] + \mathbb{E}\big[\mathbf{w}|x - \mathbb{E}[x]\big]'(\boldsymbol{\gamma} - \check{\boldsymbol{\delta}}_{y.\widetilde{w}}),$$

where $\check{\boldsymbol{\delta}}_{y.\widetilde{w}} = \mathbb{E}[(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i])\mathbf{w}_i']^{-1}\mathbb{E}[(\mathbf{w}_i - \mathbb{E}[\mathbf{w}_i])y_i]$. This estimand is clearly not equal to $\mu(x)$ unless additional assumptions hold.

When model (2.2) is misspecified for $\mathbb{E}[y|x, \mathbf{w}]$, the probability limit of both residualized binscatter and our recommended covariate-adjusted binscatter changes. In the case of residualized binscatter, the probability limit becomes quite difficult to interpret and relate to any meaningful "partial effect" of x on y. On the other hand, our approach to covariate adjustment retains the usual interpretation of standard semiparametric semi-linear models, where the true unknown "long" conditional expectation $\mathbb{E}[y|x, \mathbf{w}]$ is approximated by the closest model $\mu(x)+\mathbf{w}'\boldsymbol{\gamma}$ in a mean square error sense. See Angrist and Pischke (2008) for further discussion on the interpretation of (semi-)linear least squares approximations, and its uses in applied work.

For the reasons above, we recommend to covariate-adjust binscatter by incorporating covariates in an additively separable way, as in (3.3), and not via residualization as currently done in most empirical applications.

4 Implementing Binscatter

Binscatter is readily implementable once the number of bins J is chosen, for any polynomial order p, level of smoothness constrain $s \leq p$, and derivative of interest $v \leq p$. Therefore, for implementation purposes, we discuss first a valid and optimal choice of J based on minimizing the IMSE of binscatter as a point estimator of $\mu^{(v)}(x)$ in model (2.2), given the researchers' choice of p, s, and v.

The following basic assumption is the only one used throughout our paper.

Assumption 1. *The sample $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is i.i.d and satisfies model (2.2). Further, the covariate x_i has a continuous density function $f(x)$ 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,* $\mu(x)$ *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 standard conditions in classical regression settings. Precise regularity conditions for our theoretical results, implied by Assumption 1, are given and discussed in the supplemental appendix. When the covariates $\mathbf{w}_i$ are not adjusted for in the binscatter, all statements involving these covariates in Assumption 1 can be ignored.

To select the number of bins J forming the quantile-spaced partition $\widehat{\Delta}$ used by binscatter, we proposed to minimize an approximation to the density-weighted integrated mean square error of the estimator $\widehat{\mu}^{(v)}(x)$. Letting $\approx_{\mathbb{P}}$ denote an approximation in probability, we show that

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

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, where they are shown to be non-random functions of the sample size n, at this level of generality. The variance $\mathscr{V}_n(p,s,v)$, depending on $\sigma^2(x)$ and $f(x)$, is bounded and bounded away from zero under minimal assumptions, while the (squared) bias $\mathscr{B}_n(p,s,v)$, depending on $\mu^{(p+1)}(x)$ and $f(x)$, is generally bounded and bounded away from zero except for a very special case (see Remark SA-3.1 in the supplemental appendix for more details). 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. 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 $\mathscr{B}(0,0,0) = \frac{1}{12}\mathbb{E}\big[\big(\frac{\mu^{(1)}(x_i)}{f(x_i)}\big)^2\big]$, which corresponds to canonical binscatter ($p = v = s = 0$).

The main takeaway is that the IMSE of binscatter naturally depends on the squared bias and variance of the estimator, and these factors can be balanced out in order to select the IMSE-optimal number of bins to use in applications. The following theorem summarizes this result.

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.*

This theorem gives the optimal choice of J for the general class of binscatter considered in this paper, that is, allowing for higher-order polynomial fits within bins and imposing smooth restrictions on the fits across bins, with or without covariate adjustment, when the main object of interest is possibly a derivative of the unknown function $\mu(\cdot)$. This additional versatility will be useful in upcoming sections when constructing formal statistical testing procedures based on binscatter derivative estimates. In particular, the optimal number of bins for the canonical binscatter is obtained when $p = v = s = 0$.

As discussed in Section 2, most common practice set $s = 0$ first, in which the size of the partition is chosen without smoothness restrictions, even if later those restrictions are imposed and used for constructing smoother regression estimates and confidence bands. An important result emerging from Theorem 1 is that this approach is justified in large samples because the optimal number of bins for any $0 \leq s \leq p$ is proportional to $n^{\frac{1}{2p+3}}$, and therefore choosing J with or without imposing smoothness restrictions leads to an IMSE rate optimal binscatter —only the the constant of proportionality changes slightly depending on the s chosen.

Finally, the supplemental appendix discusses implementation issues for selecting J in applications for any binscatter: any polynomial order fit $p \geq 0$, any smoothness restriction $s = 0, 1, \ldots, p$, and any derivative level $v = 0, 1, \ldots, p$. Our implementations are readily available in `Stata` and `R`, and further discussed in Cattaneo, Crump, Farrell, and Feng (2019).

5 Using Binscatter

Our generalized binscatter estimator $\widehat{\mu}^{(v)}(x)$, with $0 \leq p$ and $0 \leq v, s \leq p$, is constructed to approximate the function $\mu^{(v)}(x)$ in model (2.2), which captures the v-th derivative partial effect of x on y, after controlling for $\mathbf{w}$. Viewed as a semi-/non-parametric estimator, binscatter can be

implemented in a valid and IMSE-optimal way by setting $J = J_{\texttt{IMSE}}$ (Theorem 1) when forming the bins partitioning the support of x.

In this section we employ binscatter for three main purposes. First, we discuss valid and optimal graphical presentation of the regression function and its derivatives. Second, we offer valid confidence intervals and bands for $\mu^{(v)}(x)$. Finally, we develop valid hypethesis test for parametric specification and nonparametric shape restrictions of $\mu^{(v)}(x)$. All the results discussed in this section are formalizations of the procedures already illustrated in Section 2.

5.1 Graphical Presentation

We proved that the binscatter estimator $\widehat{\mu}^{(v)}(x)$, implemented with $J = J_{\texttt{IMSE}}$ as in Theorem 1, is an IMSE-optimal point estimator of $\mu^{(v)}(x)$. Furthermore, we also show there that binscatter can achieve the fastest uniform rate of convergence. These results highlight some of the good statistical properties of binscatter, and justify its use for depicting an approximation to the unknown function $\mu(x)$.

In Section 2, we illustrated several of the resulting binned scatter plots, all constructed using $\widehat{\mu}^{(v)}(x)$ for appropriate choice of polynomial order within bin (p), smoothness level (s), and derivative of interest (v). To describe how these plots are constructed, let $\bar{b}_j$ denote the center of the j-th quantile-spaced bin $\widehat{\mathcal{B}}_j$, where $j = 1, 2, \ldots, J$. Then, the dots in the binned scatter plot correspond to $\widehat{\mu}(\bar{b}_j)$ for any choice of $0 \leq v, s \leq p$. In Figure 5 we illustrated the effects of varying s and p by plotting $\widehat{\mu}(x)$ as a function of x. When $s = 0$, the resulting estimator $\widehat{\mu}(x)$ is discontinuous at the bins' edges, while when $s > 0$ it is at least continuous.

In constructing a binned scatter plot, it may be convenient to report more than one estimate of $\mu(x)$ over the same quantile-spaced bins. For example, researchers can report a collection of "dots" using $\widehat{\mu}(\bar{b}_j)$, $j = 1, 2, \ldots, J$, with $p = s = 0$ (canonical binscatter), and a "line" representing a smoother estimate such as $\widehat{\mu}(x)$, $x \in \mathcal{X}$, with $p = s = 3$ (cubic B-spline). We will return to this discussion in Section 7 when we provide concrete recommendations for practice.

Finally, while not graphically illustrated in Section 2, derivative estimates can also lead to powerful and useful binned scatter plots. Specifically, in some applications researchers may be interested in an "average marginal effect" of x on y, possibly controlling for other factors $\mathbf{w}$, which is naturally captured by $\mu^{(1)}(x)$. Such a quantity is of interest in many different setups, ranging from

reduced form latent variable models to structural non-separable models. Furthermore, derivatives of $\mu(x)$ are of interest in testing for substantive hypotheses such as marginal diminishing returns. We formalize these latter ideas further below.

5.2 Pointwise Inference and Confidence Intervals

We turn now to confidence interval and confidence band estimators based on binscatter. The Studentized t-statistic is

$$\widehat{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - \mu^{(v)}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad 0 \leq v, s \leq p,$$

where the binscatter variance estimator is

$$\widehat{\Omega}(x) = \widehat{\mathbf{b}}_s^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}_s^{(v)}(x),$$

$$\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)', \qquad \widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}_s(x_i)\widehat{\mathbf{b}}_s(x_i)'(y_i - \widehat{\mathbf{b}}_s(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2.$$

Omitted technical details are given in the supplemental appendix to conserve space.

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 $\mu^{(v)}(x)$, 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:

$$\widehat{I}_p(x) = \Big[\ \widehat{\mu}^{(v)}(x) \pm \mathfrak{c}\cdot\sqrt{\widehat{\Omega}(x)/n}\ \Big], \qquad 0 \leq v, s \leq p,$$

where $\mathfrak{c}$ denotes a choice of quantile (e.g., $\mathfrak{c} \approx 1.96$ for a 95% Gaussian confidence intervals). However, employing an IMSE-optimal binscatter (e.g., Theorem 1) introduces a first-order mis-

specification error leading to invalidity of these confidence intervals. To address this problem, we rely on a simple application of robust bias-correction (Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018; Cattaneo, Farrell, and Feng, 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 $\widehat{\Delta}$ according to Theorem 1, and construct the binscatter accordingly. Then, we employ the confidence interval $\widehat{I}_{p+q}(x)$ with $\mathfrak{c} = \Phi^{-1}(1-\alpha/2)$. The following Corollary summarizes this approach.

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

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

The confidence intervals constructed in the above theorem are based on an IMSE-optimal implementation of binscatter and robust bias correction. They were illustrated in Figure 7 as individual vertical segments inside the shaded bands, which are discussed in the next subsection.

5.3 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 as in the previous section. Examples include reporting confidence bands for $\mu(x)$ and its derivatives, as well as testing for linearity, monotonicity, concavity, or other shape features of $\mu^{(v)}(x)$. This section applies a formal approach for uniform inference employing binscatter and its generalizations, and constructs valid confidence bands based on binscatter and its generalizations. In the following subsections, we employ these uniform inference results to develop asymptotically valid testing procedures for parametric model specification and nonparametric shape restrictions.

Our approach to uniform inference extends the recent work on strong approximations in Cattaneo, Farrell, and Feng (2018) to allow for estimated quantile-spaced partitioning $\widehat{\Delta}$, as commonly used in binscatter settings, which requires non-trivial additional technical work. In fact, it is not

possible to obtain a valid strong approximation for the entire stochastic process $\{\widehat{T}_p(x) : x \in \mathcal{X}\}$, as done in Cattaneo, Farrell, and Feng (2018), because uniformity fundamentally fails when the partitioning scheme is random: see the supplemental appendix for details. Inspired by the work in Chernozhukov, Chetverikov, and Kato (2014a,b), our approach circumvents this technical hurdle by retaining the randomness introduced by $\widehat{\Delta}$, and focusing instead on the specific functional of interest (i.e., suprema).

Given the technical nature of our strong approximation results for quantile-spaced estimated partitions and binscatter, we relegate further discussion to the supplemental appendix. In this section we apply these results to construct valid robust bias-corrected confidence bands for $\mu^{(v)}(x)$, while in the next two upcoming sections we employ them to develop valid testing procedures. For a choice of p, $0 \leq v, s \leq p$, and quantile-spaced partition size J, we define

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

By construction, this band covers the entire function $\mu^{(v)}(x)$ with probability $1-\alpha$.

The main drawback in the construction above, however, is that the quantiles $\mathfrak{c}$ are 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, let $\mathbf{N}_K$ be a K-dimensional standard normal random vector, and define the following (conditional) Gaussian process:

$$\widehat{Z}_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}^{-1/2}}{\sqrt{\widehat{\Omega}(x)/n}}\mathbf{N}_{(p+1)J-(J-1)s}, \qquad x \in \mathcal{X}, \quad 0 \leq v, s \leq p. \tag{5.1}$$

We show 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 $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : i = 1, 2, \ldots, n\}$. This result implies that the quantiles used to construct confidence bands can be approximated by resampling from the normal random vectors $\mathbf{N}_{(p+1)J-(J-1)s}$, keeping the data $\mathbf{D}$ fixed. We make this approach precise in the following theorem.

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,*

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

where $\to_{\mathbb{P}}$ *denotes convergence in probability.*

Putting the above together, we have the following main result for robust bias-corrected confidence bands.

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

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

This corollary offers a valid confidence bands construction for $\mu^{(v)}(\cdot)$, which relies on resampling from a particular random variable: $\sup_{x\mathcal{X}} |\widehat{Z}_{p+q}(x)|$, conditional on the original data. In practice, this supremum is replaced by a maximum over a fine grid of evaluation points, and each realization of $\widehat{Z}_{p+q}(x)$ is obtained by resampling from the standard normal random vector $\mathbf{N}_{(p+q+1)J-(J-1)s}$ and then computing $\widehat{Z}_{p+q}(x)$ as in (5.1), where all other quantities are fixed and known given the original data. As a consequence, the quantile $\mathfrak{c}$ is actually estimated conditional on $\mathbf{D}$. Further details on implementation are given in our companion software Cattaneo, Crump, Farrell, and Feng (2019).

5.4 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 tests: parametric specifications of $\mu^{(v)}(x)$. In the next section, we discuss another type of shape-related hypothesis test: testing for nonparametric features such as monotonicity or concavity of $\mu^{(v)}(x)$.

One type of informal analysis commonly encountered in empirical work concerns comparing the binscatter $\widehat{\mu}^{(v)}(x)$ relative to some parametric fit. For example, $\widehat{\mu}(x)$ 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 $\mu(x)$ 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 $\mu(x) = \theta_0 + x\theta_1$ or perhaps $\mu(x) = \theta_0 + x\theta_1 + x^2\theta_2$ for some unknown coefficients $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$. More generally, researchers are often interested in assessing formally whether $\mu(x) = m(x, \boldsymbol{\theta})$ 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 $m^{(v)}(x, \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$, the null and alternative hypotheses are

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

As is evident from its formulation, this testing problem can be implemented using test statistics involving the supremum of (derivatives of) binscatter, with or without employing higher-order polynomials, imposing smoothness restrictions, or adjusting for additional covariates. Crucially, 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 $\widehat{\boldsymbol{\theta}}$ that consistently estimates $\boldsymbol{\theta}$ 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

$$\ddot{T}_p(x) = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x, \widehat{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}, \qquad 0 \leq v, s \leq p,$$

leading to the hypothesis test:

$$\text{Reject } \ddot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} |\ddot{T}_p(x)| \geq \mathfrak{c}, \tag{5.2}$$

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 (5.2) feasible.

Theorem 4 (Hypothesis Testing: Parametric Specification)**.** *Let Assumption 1 hold. For given p, and $0 \leq v, s \leq p$, 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 $\sup_{x\in\mathcal{X}} |m^{(v)}(x,\widehat{\boldsymbol{\theta}}) - \mu^{(v)}(x)| = O_{\mathbb{P}}(n^{-1/2})$, 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 some $\bar{\boldsymbol{\theta}}$ such that $\sup_{x\in\mathcal{X}} |m^{(v)}(x,\widehat{\boldsymbol{\theta}}) - m^{(v)}(x,\bar{\boldsymbol{\theta}})| = O_{\mathbb{P}}(n^{-1/2})$, 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 a very intuitive idea: if the confidence band for $\mu^{(v)}(x)$ does not contain entirely the parametric fit considered, then such parametric fit is inconsistent with the data, i.e., should be rejected. Formally, this leads to the hypothesis testing procedure (5.2), which relies on a proper (simulated) critical value. The condition $\sup_{x\in\mathcal{X}} |m^{(v)}(x,\widehat{\boldsymbol{\theta}}) - \mu^{(v)}(x)| = O_{\mathbb{P}}(n^{-1/2})$ under the null hypothesis is very mild: it states that the unknown parameters entering the parametric specification $\mu(x) = m(x,\boldsymbol{\theta})$ 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 $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_{\mathbb{P}}(1)$ and $m(x,\boldsymbol{\theta})$ continuous in x and continuously differentiable in $\boldsymbol{\theta}$. Most standard parametric models in applied microeconomics satisfy such conditions, including linear and non-linear regression models, discrete choice models (Probit or Logit), censored and truncation models, and many more.

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. Section 2 illustrated this with Table 1 and Figure 8. See Cattaneo, Crump, Farrell, and Feng (2019) for more implementation ideas and details.

Remark 1 (Other Metrics)**.** The parametric specification test in (5.2) is based on the maximum discrepancy between the fit of the hypothesized parametric model for $\mu(x)$ and the nonparametric binscatter approximation. Some practitioners, however, may prefer to assess the discrepancy by means of an alternative metric, such as the mean square difference between the parametric and

nonparametric fits. Our theoretical results given in the supplemental appendix are general enough to accommodate such alternative comparisons, but we do not discuss them here only to conserve space. ⌟

5.5 Testing Shape Restrictions

The hypothesis test (5.2) 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 $\mu(x)$ is constant, then $\mu^{(1)}(x)=0$ for all $x\in\mathcal{X}$, which can be implemented using Theorem 4 upon setting $m(\cdot)=0$ and $v=1$, for any $p\geq 1$ and $0\leq s\leq p$. Similarly, linearity or other related nonparametric shape restrictions can be tested for via the results in Theorem 4, for appropriate choice of v. The common feature in all cases is that the null hypothesis of interest is two-sided.

There are, however, other important nonparametric shape restriction hypotheses about $\mu(x)$ that correspond to one-sided null hypothesis, and thus cannot be implemented using Theorem 4. For example, negativity, monotonicity and concavity of $\mu(x)$ all correspond to formal statements of the form $\mu(x)\leq 0$, $\mu^{(1)}(x)\leq 0$ and $\mu^{(2)}(x)\leq 0$, 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

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

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 statistic:

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

leading to the hypothesis test:

$$\text{Reject } \dot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} \dot{T}_p(x) \geq \mathfrak{c}, \tag{5.3}$$

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. For given p, and $0 \leq v, s \leq p$, 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 $\dot{\mathsf{H}}_0$, 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 (5.3) 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 (2019) for more details.

Finally, the supplemental appendix includes a more general version of Theorem 5, where we test against a parametric fit of the form $m(x,\bar{\boldsymbol{\theta}})$ where $\bar{\boldsymbol{\theta}}$ is such that $\sup_{x\in\mathcal{X}} |m(x,\widehat{\boldsymbol{\theta}}) - m(x,\bar{\boldsymbol{\theta}})| = O_{\mathbb{P}}(n^{-1/2})$ for some estimator $\widehat{\boldsymbol{\theta}}$. For example, this more involved testing procedure might be useful to assess whether $\mu(x)$ is "below" the line defined by $m(x,\bar{\boldsymbol{\theta}}) = x'\bar{\boldsymbol{\theta}}$, when $\widehat{\boldsymbol{\theta}}$ is the OLS estimator based on y_i and x_i, and $\bar{\boldsymbol{\theta}}$ is its probability limit (i.e., $x'\bar{\boldsymbol{\theta}}$ denotes the best linear predictor of y based on x).

Remark 2 (Two-Sample Nonparametric Testing)**.** Our results can also be extended to handle nonparametric testing about features of $\mu(x)$ for two or more groups. For example, assuming that two (sub-)samples are available, our methods can be used to test the null hypothesis: H_0 :

$\mu_1(x) = \mu_2(x)$ for all $x \in \mathcal{X}$, where $\mu_1(x)$ and $\mu_2(x)$ denote the $\mu(x)$ function in our framework for two distinct (sub-)samples. Such a hypothesis test can be formally implemented using a uniform measure of discrepancy, as we used above, or some other metric (see Remark 1). Our theoretical results given in the supplemental appendix are general enough to accomodate this extension, which we plan to undertake in upcoming work. ⌟

6 Extension: Clustered Data

Our main methodological and theoretical results can be extended to standard clustered data settings, allowing for fixed effects and clustered variance estimation. This extension is mostly straightforward, although notationally more cumbersome. In this section we briefly discuss how such an extension would proceed, focusing only on the minimum necessary material for tackling new technical issues specific to the clustered data context. Cattaneo, Crump, Farrell, and Feng (2019) discusses implementation issues related to binscatter based on clustered data.

In the standard setting with a "large" number of clusters or groups, under random sampling, $\{(y_{ig}, x_{ig}, \mathbf{w}'_{ig}) : 1 \leq i \leq n_g, 1 \leq g \leq G\}$ are i.i.d. over $1 \leq g \leq G$, where $n_1, n_2, \ldots, n_G$ are fixed within-group sample sizes and $G \to \infty$ is the number of groups. In many applications researchers may hope to control for a number of group fixed effects, thus allowing for potential heterogeneity across groups, or to rely on clustered variance estimation for inference. We discuss briefly each of these cases next.

6.1 Fixed Effects

Given the cluster/group notation and sampling assumption above, model (2.2) can be generalized as follows:

$$y_{ig} = \mu(x_{ig}) + \mathbf{w}'_{ig}\boldsymbol{\gamma} + \lambda_g + \epsilon_{ig}, \qquad \mathbb{E}[\epsilon_{ig}|x_{ig}, \mathbf{w}_{ig}, \lambda_g] = 0, \tag{6.1}$$

where the two subindexes account for within and between grouping of observations, and λ_g denotes an unobserved group-specific random component that does not vary within group. In this context, a pooled cross-sectional regression can be used directly, as in the previous sections, whenever λ_g is uncorrelated with $(x_i, \mathbf{w}'_i)$.

On the other hand, when λ_g is correlated with $(x_i, \mathbf{w}_i')$, a group-fixed-effect model should be used instead. Intuitively, this amounts to simply adding a collection of dummy variables (group indicators) to the binscatter regression (3.3) for covariate-adjustment, just as for the covariate vector $\mathbf{w}_i$. From a practical perspective, semi-linear binscatter estimation based on the model (6.1) remains the same as that for the classical linear models with one-way fixed effects, once the unknown function $\mu(x_{ig})$ is replaced by the binscatter approximation basis.

However, there are two important theoretical issues worth mentioning when fixed effects are added to the binscatter regression (3.3). First, while $\boldsymbol{\gamma}$ was assumed to be consistently estimable, at a sufficiently fast rate, the fixed effects need not (and usually will not) be consistently estimable. Alternatively, as it is well known, a within-group transformation or first differences may be employed to purge the group fixed effects. This issue can be resolved easily with a careful asymptotic analysis of the fixed effect binscatter estimator.

Second, the unknown function $\mu(x)$ is only identifiable up to a "shift" in the level when saturated fixed effects are included in the estimation. In other words, the "constant term" in $\mu(x)$ is not identifiable without further restriction. A simple identifying condition is $\mathbb{E}[\mu(x_{ig})] = 0$. Accordingly, the binscatter basis $\widehat{\mathbf{b}}_s(x)$ must be rotated such that the new basis is centered at zero. Recall that $\widehat{\mathbf{b}}_s(x)$ is a $[(p+1)J - (J-1)s]$-dimensional basis that reproduces any constant by construction. Therefore, following the identifying condition $\mathbb{E}[\mu(x_{ig})] = 0$, the new basis centered at zero needs to be $[(p+1)J - (J-1)s - 1]$-dimensional with one less degree of freedom. Such rotated bases have been developed in the literature before: see Burman (1991) for one example.

6.2 Clustering

When λ_g in (6.1) is uncorrelated with $(x_i, \mathbf{w}_i')$, it can be absorbed into the unobserved error term, leading to the simple model $y_{ig} = \mu(x_{ig}) + \mathbf{w}_{ig}'\boldsymbol{\gamma} + \upsilon_{ig}$ with $\upsilon_{ig} = \lambda_g + \epsilon_{ig}$. Clearly, the new error term satisfies $\mathbb{E}[\upsilon_{ig}|x_{ig}, \mathbf{w}_{ig}] = 0$, but $\mathbb{E}[\upsilon_{ig}\upsilon_{i'g}] \neq 0$ for $i \neq i'$ in general due to the existence of the common group feature λ_g. This corresponds to the classical "random effect" linear regression model.

More generally, in analyses of clustered or grouped data, a typical assumption is that "individuals" within the same group share some unobserved common characteristic or shock, thus rendering potential within-group correlation between unobservables. Specifically, the canonical clustering

model is:

$$y_{ig} = \mu(x_{ig}) + \mathbf{w}_{ig}'\boldsymbol{\gamma} + \upsilon_{ig}, \qquad \begin{aligned} &\mathbb{E}[\upsilon_{ig}|x_{ig}, \mathbf{w}_{ig}] = 0, \\ &\mathbb{E}[\upsilon_{ig}\upsilon_{i'g}] \neq 0 \text{ for } i \neq i', \\ &\mathbb{E}[\upsilon_{ig}\upsilon_{jg'}] = 0 \text{ for } g \neq g'. \end{aligned} \tag{6.2}$$

The unobserved error terms are uncorrelated over $1 \leq g \leq G$, but correlated within each group. Given such structure, the effective sample size is now G, rather than the total sample size $n = \sum_{g=1}^{G} n_g$. When the cluster sizes are fixed (non-increasing as $n \to \infty$), the difference in the asymptotic analysis of the binscatter estimator based on model (6.2) is minor, since G is proportional to n up to some constant. Finite sample performance, however, may exhibit more pronounced differences, particularly when G is relatively small.

To be more precise, allowing for clustering in the unobserved error component as in model (6.2) requires two main changes in the results presented in the previous sections for the binscatter regression estimator (3.3). First, the asymptotic variance of binscatter changes due to the clustering structure, and therefore a cluster-robust consistent estimator thereof must be used instead. Second, the IMSE-optimal choice of J developed in Theorem 1 now depends on G, the effective sample size, as opposed to n as before, and of course the variance formula entering in the denominator of $J_{\texttt{IMSE}}$ also changes to reflect the clustering structure underlying the data.

In conclusion, while some specific aspects of our results need to be adjusted to account for the within-group correlation present in the data (e.g., cluster-robust variance estimation or convergence rate of $J_{\texttt{IMSE}}$), all the main theoretical and methodological conclusions presented in the previous sections remain conceptually unchanged.

7 Recommendations for Practice

This paper offered a thorough treatment of canonical binscatter and its generalizations to within-bins polynomial fitting, across-bins smoothness restrictions, and covariate adjustments. Our main results can be used to guide practice in a principle way. In this section, we offer a list of recommendations for empirical work employing binscatter, and illustrate them using data from the American

Community Survey (ACS). See Section SA-5 in the supplemental appendix for details on data, and for two other empirical illustrations.

For graphical presentation we recommend the following:

Step 1. Use IMSE-optimal canonical binscatter to depict data, with covariate adjustment and accounting for clustered data as appropriate. Formally, set $p = s = v = 0$ and $J = J_{\texttt{IMSE}}$ (Theorem 1), and then plot $\widehat{\mu}(\bar{b}_j)$ as "dots", where $\bar{b}_j$ denotes the center of the j-th quantile-spaced bin (Section 5), $j = 1, 2, \ldots, J$.

Step 2. On the same quantile-space partition determined by $J_{\texttt{IMSE}}$ in Step 1, construct a cubic B-spline for flexible approximation of $\mu(x)$, with covariate adjustment and accounting for clustered data as appropriate. Formally, set $p = 3$, $s = 3$ and $v = 0$, and then plot $\widehat{\mu}(x)$ as a solid line.

Step 3. Under the baseline configuration in Steps 1 and 2, confidence bands can be constructed on the same quantile-space partitioning and for the same cubic B-spline choices using the simulated quantiles (Lemma 2). These bands can be plotted directly on top of the "dots" from Step 1 and the solid line from Step 2.

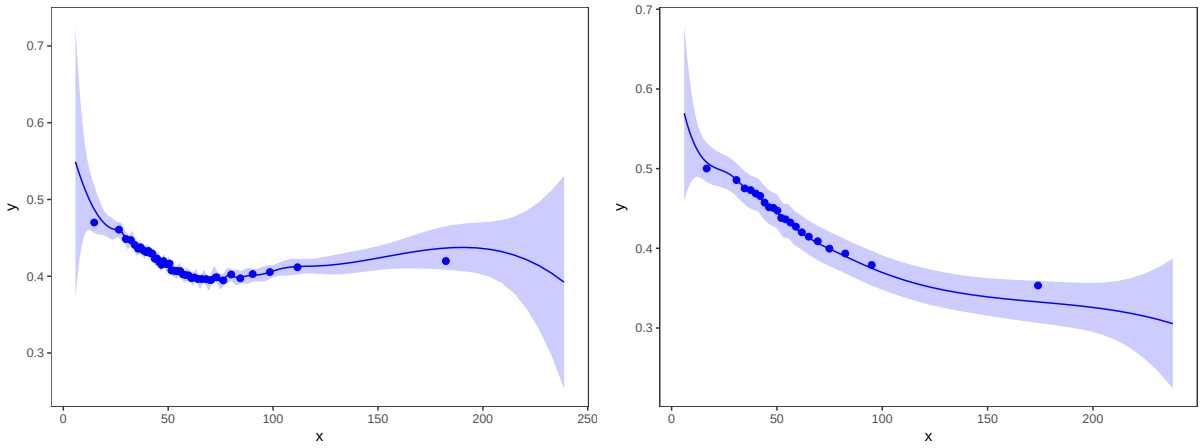
Step 4. As a way of contrast, add a parametric fit as well, if desired.

This approach is illustrated in Figure 10. The dependent variable is the Gini index observed at the zip code tabulation area level based on 5-year ACS estimates from 2013 to 2017 (excluding observations for Puerto Rico). Our variable of interest is median household income (dollars in thousands). When we apply controls, the control variables are (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, (iv) percentage of residents without health insurance, and (v) the local unemployment rate. All control variables are also observed at the zip code tabulation area level. Plot (a) shows that their relationship is nonlinear and unstable, while it becomes monotonically decreasing when the control variables are added, as shown in Plot (b).

For formal testing of substantive features of $\mu(x)$ we recommend the following:

Step 1. Use IMSE-optimal cubic B-spline binscatter to approximate the function $\mu(x)$, with covariate adjustment and accounting for clustered data as appropriate. Specifically, set $p = 3$ and

Figure 10: Gini Index versus Household Income.



(a) No Controls (b) With Controls

Notes. See Section SA-5 in the Supplemental Appendix for further details on data sources.

$s = 3$, and $J = J_{\texttt{IMSE}}$ (Theorem 1), for $v = 0, 1, 2$ (see Section 5 for details). For $v > 2$, use $p = 3 + v$ and $s = 3 + v$, and $J = J_{\texttt{IMSE}}$ (Theorem 1).

Step 2. Conduct formal hypothesis testing procedures using Theorem 4 and Theorem 5, as appropriate, with $q = 1$.

This approach is illustrated in Table 2. Specifically, the results indicate that the relation between the Gini index and household income, controlling for covariates, is nonlinear, but can be modelled well by a quadratic polynomial. Moreover, the hypotheses of monotonicity and convexity are also supported. In this empirical example, positivity holds by construction, but the test is nonetheless included for completeness.

Finally, all the recommendations discussed above are the default choices in our companion `R` and `Stata` software implementations (see Cattaneo, Crump, Farrell, and Feng, 2019, for more details).

8 Conclusion

We introduced a general econometrics framework to understand binscatter, a very popular methodology for approximating the conditional expectation function in applied microeconomics. Our framework leads to a variety of new methodological (and theoretical) results for the canonical binscatter, including novel smooth and/or polynomial approximation approaches, principled covariate

Table 2: Testing of Substantive Hypothesis.

	Test Statistic	P-value	J
Parametric Specification			
Constant	17.665	0.000	21
Linear	6.152	0.000	21
Quadratic	1.947	0.268	21
Shape Restrictions			
Positivity	9.666	1.000	21
Decreasingness	−0.634	1.000	9
Convexity	−1.678	0.420	5

Notes. A set of control variables are added. The number of bins is IMSE-optimal, selected based on a fully data-driven procedure. See Section SA-5 in the Supplemental Appendix for further details.

adjustment implementation, and valid inference and hypothesis testing methods. In particular, we highlight important problems with the way covariate adjustment is currently done in practice via current popular software implementations (Stepner, 2014).

In addition to providing the first foundational methodological study of binscatter and extensions thereof, which helps both in understanding the validity (or lack thereof) of current practices and in offering principled guidance for future applications, we also offer new accompanying `Stata` and `R` software implementing all our results (Cattaneo, Crump, Farrell, and Feng, 2019).

References

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*, forthcoming.

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.

Burman, P. (1991): "Rates of Convergence for the Estimates of the Optimal Transformations of Variables," *Annals of Statistics*, pp. 702–723.

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 (2019): "Binscatter Regressions," in preparation for the *Stata Journal*.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and E. Schaumburg (2019): "Characteristic-Sorted Portfolios: Estimation and Inference," arXiv:1809.03584.

Cattaneo, M. D., and M. H. Farrell (2011): "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates," in *Advances in Econometrics: Missing Data Methods*, ed. by D. Drukker, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

——— (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174(2), 127–143.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2018): "Large Sample Properties of Partitioning-Based Estimators," arXiv:1804.04916.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34(2), 277–301.

——— (2018b): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113(523), 1350–1361.

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.

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

Friedman, J. H. (1977): "A Recursive Partitioning Decision Rule for Nonparametric Classification," *IEEE Transactions on Computers*, C-26(4), 404–408.

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 (2018): "A Binned Scatterplot is Worth a Hundred Regressions: Diffusing a Simple Tool to Make Empirical Research Easier and Better," SSRN Working paper No. 3257345.

Stepner, M. (2014): "Binned Scatterplots: Introducing -binscatter- and Exploring its Applications," 2014 Stata Conference 4, Stata Users Group.

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.