On Binscatter*

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

October 11, 2022

Abstract

Binned scatter plots, or binscatters, have become a popular and convenient tool in applied microeconomics for visualizing bivariate relations and conducting informal specification testing. However, a binscatter, on its own, is very limited in what it can characterize about the conditional mean. We introduce a suite of formal and visualization tools based on binned scatter plots to restore, and in some dimensions surpass, the visualization benefits of the classical scatter plot. We deliver a comprehensive toolkit for applications, including estimation of conditional mean and quantile functions, visualization of variance and precise quantification of uncertainty, and formal tests of substantive hypotheses such as linearity or monotonicity, and an extension to testing differences across groups. To do so we give an extensive theoretical analysis of binscatter and related partition-based methods, accommodating nonlinear and potentially nonsmooth models, which allows us to treat binary, count, and other discrete outcomes as well. We also correct a methodological mistake related to covariate adjustment present in prior implementations, which yields an incorrect shape and support of the conditional mean. All of our results are implemented in publicly available software, and showcased with three substantive empirical illustrations. Our empirical results are dramatically different when compared to those obtained using the prevalent methods in the literature.

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

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

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

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

§Booth School of Business, University of Chicago.

¶School of Economics and Management, Tsinghua University.

1 Introduction

The classical scatter plot is a fundamental visualization tool in data analysis. Given a sample of bivariate data, a scatter plot displays all n data points at their coordinates (x_i, y_i), $i = 1, \ldots, n$. By plotting every data point, one obtains a visualization of the joint distribution of y and x. When used prior to regression analyses, a scatter plot allows researchers to assess the functional form of the regression function, the variability around this conditional mean, and recognize unusual observations, bunching, or other anomalies or irregularities.

Classical scatter plots however have several limitations and have fallen out of favor. First, with the advent of larger data sets, the cloud of points becomes increasingly dense, rendering scatter plots uninformative. Even for moderately sized but noisy samples it can be difficult to asses the shape and other properties of the conditional mean function. Further, with increasing attention paid to privacy concerns, plotting the raw data may be disallowed completely. Second, the classical scatter plot does not naturally allow for a visualization of the relationship of y and x while controlling for other covariates, which is a standard goal in social sciences. Binned scatter plots, or binscatters, have become a popular and convenient alternative tool in applied microeconomics for visualizing bivariate relations (see Starr and Goldfarb, 2020, and references therein, for an overview of the literature). A binscatter is made by partitioning the support of x into a modest number of bins and displaying a single point per bin, showing the average outcome for observations within that bin. While this makes for a simpler, cleaner plot than a classical scatter plot, it, importantly, does not present the same information. While a scatter plot allows one to display the entirety of the data, a binscatter shows only an estimate of the conditional mean function. A binned scatter plot is therefore not an exact substitute for the classical scatter plot, but it can be used to judge functional form, provide a qualitative assessment of features such as monotonicity or concavity, and guide later regression analyses. Handling covariates correctly is a particularly subtle issue.

In this paper we introduce a suite of formal and visual tools based on binned scatter plots to restore, and in some dimensions surpass, the visualization benefits of the classical scatter plot. We deliver a fully featured toolkit for applications, including estimation of conditional mean and quantile functions, visualization of variance and precise quantification of uncertainty, and formal tests of substantive hypotheses such as linearity or monotonicity. Our toolkit allows for characterizing

key features of the data without struggling to parse a dense cloud of a large data set or sharing individual data points or betraying other identifying information. As a foundation for our results we deliver an extensive theoretical analysis of binscatter methods and related partition-based tools. We also correct a prevalent methodological mistake related to covariate adjustment present in prior implementations, which yields an incorrect shape and incorrect support of the conditional mean.

The concept of a binned scatter plot is simple and intuitive: divide the data into $J < n$ bins according to the covariate x, often using empirical ventiles, and then calculate the average outcomes among observations with covariate values lying in each bin. The final plot shows the J points $(\bar{x}_j, \bar{y}_j)$, the sample averages in each bin. Further, by plotting only averages, discrete-valued outcomes are easily accommodated. The result is a figure which shares the conceptual appeal, visual simplicity, and *some* of the utility of a classical scatter plot.

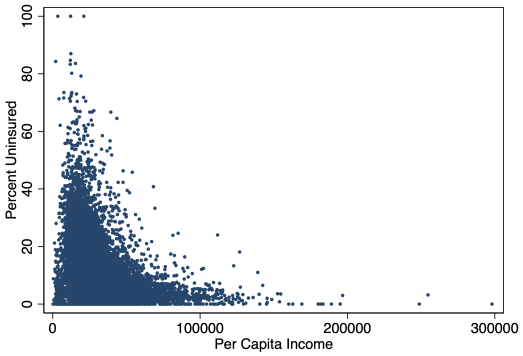
In a binned scatter plot the J points are then used to visually assess the bivariate relation between y and x. Because each of the J points in a binned scatter plot shows a conditional average, i.e. the average outcome given that x_i falls into a specific bin, using the plot to examine the conditional mean is intuitive. The primary use is assessing the shape of this mean function: whether the relationship is linear, monotonic, convex, and so forth. In applications, a roughly linear binscatter often precedes a linear regression analysis. Indeed, we provide formal results which justify such an approach in a principled, valid way.

Figure 1 shows an example of this construction. Here we study the relationship between the uninsured rate (y) and per capita income (x) at the zip code level. Figure 1(a) shows the classical scatter plot of the raw data. This data set has about 32,000 observations, far from the millions commonly encountered, and already this plot is not useful for assessing functional form: the visualization is dominated by a dense cloud of data with a few outliers. Figure 1(b) shows a binned scatter plot being constructed, with the raw data in the background, and 1(c) isolates the binscatter, and overlays a linear regression. Graphs like 1(c) are often found in empirical papers (further examples are given below). An important note is that although the binned scatter plot invites the viewer to "connect the dots" smoothly, the actual estimator is piecewise constant, as shown explicitly in Figure 1(d). Though graphically distinct, this is formally identical to the dots in Figure 1(c).

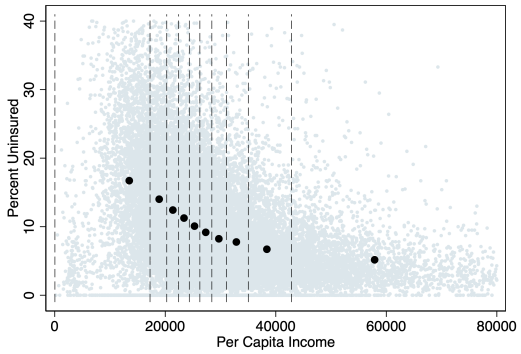
We note that it is common practice to use additional control variables and fixed effects when constructing a binscatter. The standard plots, like 1(c), will often be made after "controlling" for

Figure 1: **Illustration of Binned Scatter plots.** This figure illustrates the core construction of a binned scatter plot along with new features introduced herein. The data are obtained from the American Community Survey (ACS) using the 5-year survey estimates beginning in 2013 and ending in 2017 (available from the Census Bureau website). All analyses are performed at the zip code tabulation area level for the United States (excluding Puerto Rico). The dependent variable is the percentage of individuals without health insurance and the independent variable of interest is per capita income. Shaded regions denote confidence bands for a nominal level of 95%

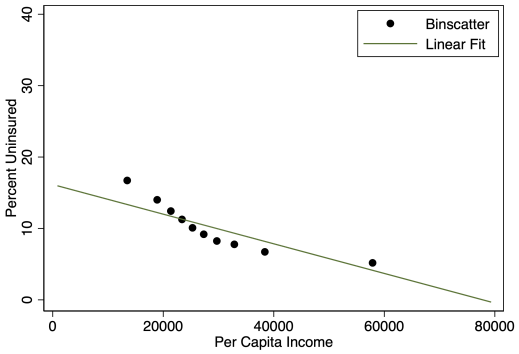
(a) *Raw Scatter Plot*



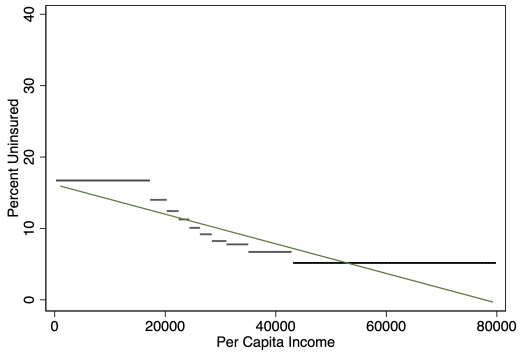
(b) *Bin Cutoffs and Binscatter*



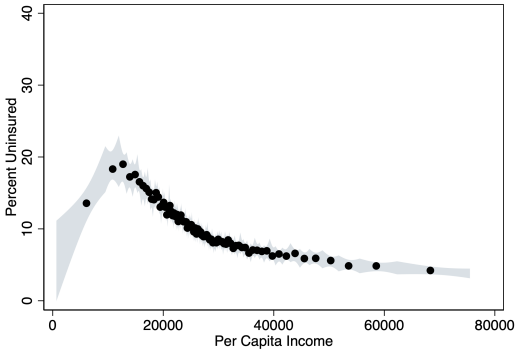
(c) *Canonical Binscatter Plot*



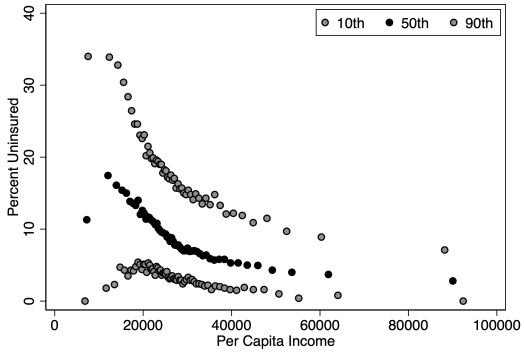
(d) *Conditional Exp. Point Est.*



(e) *Binscatter w/ Uniform Confidence Band*



(f) *Conditional Quantile Point Est.*



a set of covariates. This turns out to be quite a subtle issue, as the controls affect the visualization as well as the degree of uncertainty. Even the common practice of adding a regression line to a binned scatter plot is not straightforward to do correctly. As above, we correct a prevalent mistake

in this context. We discuss these issues in detail in Sections 2.2.1 and 3.2; see also Figure 2 and 3.

Figures 1(b) and 1(c) also highlight the fact that although the averaging is useful for evaluating the conditional mean, it masks other features of the conditional distribution which may be important to the subsequent analysis, including the variance, potential outliers, or the behavior of different quantiles. This presents a clear limitation to the usefulness of canonical binscatter methods for visualization and analysis. Note how much information is lost in moving from Figure 1(b) to 1(c).

We provide array of results and tools for binned scatter plots aimed at restoring this lost information and improving their empirical application. We improve on the estimation of conditional mean functions and also provide tools for quantifying uncertainty and capturing other features of the conditional distribution, such as variability and quantiles. To facilitate our analysis, we first demonstrate that a binscatter is a nonparametric estimator and we provide a modeling framework that enables formal analysis, allowing us to deliver new, more powerful methods and to resolve conceptual and implementation issues. We clarify precisely the parameters of interest in applications, both for visualization and formal inference. In so doing, we highlight important methodological and theoretical problems with the commonly used practice of first "residualizing out" additional covariates before constructing a binscatter. Our framework centers around a partially linear model, wherein we show how to control for additional variables in a principled and interpretable way and discuss how prior implementations are neither correct nor interpretable.

Within our framework, we first discuss the choice of the number of bins, J. Here we provide two methods. The first mimics common practice by taking $J = J_{\text{FIXED}}$ as a fixed, pre-set value, often set as $J_{\text{FIXED}} = 20$. This yields a simple and appealing plot for visualization. Second, we give an integrated mean squared error expansion and use this to select J_{IMSE}. Given a choice of J, we can then use a binscatter to estimate the conditional mean, which is the canonical usage.

We then turn to uncertainty quantification. For visualization, we provide confidence bands that capture the uncertainty in estimating the conditional mean or other functional parameters of interest. A confidence band is a region that contains the entire function with some pre-set probability, just as a confidence interval covers a single value, and is thus the proper tool for assessing uncertainty about the regression function. Confidence bands can be used to visually assess the plausibility of parametric functional forms, such as linearity. Further, our confidence

bands are explicitly functions of the conditional heteroskedasticity in the underlying data. Figure 1(e) shows an example of a valid confidence band. Notice that the seemingly good fit of linearity in Figure 1(c) is, in fact, summarily rejected by the data. The findings of Figure 1(e) are furthermore in line with the institutional structure of health care in the U.S. due to the presence of Medicaid programs, which cause the uninsured rate to fall for the lowest income areas. These conclusions are only possible with (i) a proper choice of J and (ii) valid uncertainty quantification. Confidence bands partly restore uncertainty visualization capability of the classical scatter plot by capturing how certain we are about the functional form of the conditional mean.

Delivering a valid confidence band requires novel theoretical results, which represent the main technical contribution of our work. This theory underpins numerous formal inference results, beyond visualization. We immediately obtain tests for parametric specifications (such as linearity or quadraticity) and shape restrictions (monotonicity, convexity, and so forth), thus formalizing a major use case of binned scatter plots. Furthermore, we also study group-wise comparisons which includes treatment effect heterogeneity in experimental and non-experimental settings. Our results allow for the valid discovery and visualization of treatment effect heterogeneity patterns across all types of covariates, for all types of outcomes.

Further, our results highlight important conceptual issues that arise when endeavoring to quantify uncertainty in the presence of control variables. We show how the results of these tests are sensitive to the way in which control variables are coded. This result may be somewhat counter-intuitive, as one would expect if the additional controls are modeled as additively linear they should not impact conclusions about the nonparametric relationship between y and x. We show that this may be circumvented by focusing instead on the derivative of the mean function, highlighting the importance of our theoretical contributions which can accommodate the estimation of derivatives. For example, we explain and formalize how testing linearity of the mean function can be different than testing if the first derivative is constant. We show why the latter approach provides more appealing and robust inference.

All of our results cover general nonlinear and potentially nonsmooth models, which allows us to discrete outcomes (such as logit/probit or Poisson regression), as well as conditional quantiles. Beyond formal estimation and inference, these extensions are important for visualization. For example, a classical scatter plot for binary outcomes is an ineffectual visualization tool, but a

binscatter with accompanying confidence bands allows researchers to inspect many of the same features they would expect with a continuous outcome variable. Conditional quantiles bring the visualization functionality of binscatter still even closer to the features of a classical scatter plot by capturing the variance in the data; this augments the uncertainty visualization of a confidence band. Observe that Figure 1(f) restores the visualization of the variability in the data that is present in Figure 1(a) but hidden by the averaging in 1(c).

Although binning as a nonparametric procedure has been studied in the past, existing theory is insufficient for our purposes for two main reasons. First, the extant literature cannot generally accommodate data-driven bin breakpoints such as splitting the support by empirical quantiles. Such a choice of breakpoints generates random basis functions and so are not nested in previously obtained results on nonparametric series estimators. Second, where results are available, they imply overly stringent conditions on smoothing parameters ruling out simple averaging in each bin (which amounts to local constant fitting). Circumventing these two issues with new theoretical results is crucial to directly study the empirical practice of binned scatter plots. Further detail is given in the Appendix and online supplement appendix (SA).

The paper proceeds as follows. We next briefly review the related literature. Then Sections 2 and 3 present our main ideas, results, and tools for the leading case of nonparametric least squares regression. Section 2 formalizes binned scatter plots as a nonparametric estimator, including clarifying the parameter of interest, the correct method for adding control variables, and the choice of J. Section 3 studies uncertainty quantification for both testing and visualization. Section 4 extends all our results to nonlinear models, covering discrete outcomes and quantiles in particular. Throughout we will use the application discussed in Figure 1 to illustrate our ideas. In addition, Section 5 we use our novel methods to revisit two recent empirical studies, Akcigit, Grigsby, Nicholas, and Stantcheva (2022) and Moretti (2021), and we show how our new tools can sharpen and improve estimation and inference in practice. Finally, Section 6 concludes. An Appendix provides a summary of the technical contributions of the paper and a Supplemental Appendix (SA) gives proofs of all our results and more detailed discussion of the technical innovations. All of our methodological results are available in fully-featured `Stata`, `R`, and `Python`. See our companion software article (Cattaneo, Crump, Farrell, and Feng, 2022) and the software repository at `https://nppackages.github.io/binsreg/`.

1.1 Related Literature

Our paper speaks directly to the applied literature using binscatter methods, which is too large to enumerate here. Starr and Goldfarb (2020) gives an overview and many references. Beyond binscatter itself, binning has a long history in both visualization and formal estimation. The most familiar case is the classical histogram. Applying binning to regression problems dates back at least to the regressogram of Tukey (1961). The core idea has been applied in such diverse areas as climate studies, for nonlinearity detection (Schlenker and Roberts, 2009); program evaluation, called subclassification (Stuart, 2010); empirical finance, called portfolio sorting (Bali, Engle, and Murray, 2016); and applied microeconomics for visualization in bunching (Kleven, 2016) and in regression discontinuity designs (Cattaneo and Titiunik, 2022). In nonparametric regression more broadly it is known as partitioning regression (Györfi, Kohler, Krzyżak, and Walk, 2002). Our work contributes to this broad line of work, and our tools can be useful in many of these areas directly.

In recent years, there has been related research looking at the importance and limitations of graphical analysis in different applied areas. For example, Korting, Lieberman, Matsudaira, Pei, and Shen (2021) conduct an experiment to investigate the role of visual inference and graphical representation in regression discontinuity designs via RD plots. They conclude that unprincipled graphical methods could lead to misleading or incorrect empirical conclusions. Similar concerns regarding graphical analysis are raised by Freyaldenhoven, Hansen, and Shapiro (2019) and Freyaldenhoven, Hansen, Pérez Pérez, and Shapiro (2021) in the context of event study designs, where they proposed principled visualization methods in that setting. Graphical and visualization methods are also being actively discussed in the machine learning community (see Wang, Chen, Wang, and Qu, 2021, and references therein, for an overview of the literature), where once again it is highlighted the importance of focusing on principled methods with well-understood properties for both in-sample and out-of-sample learning. Our paper contributes directly to this literature by offering principled approaches for visualization and inference employing binscatter methodology. Furthermore, well-executed visualization techniques can help with issues of statistical nonsignificance in empirical economics employing big data (Abadie, 2020).

Finally, our technical work is most closely related to the literature on uniform distributional approximations and their application to nonparametric linear and nonlinear series regression esti-

mation (Belloni, Chernozhukov, Chetverikov, and Kato, 2015; Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2019; Cattaneo, Farrell, and Feng, 2020, and references therein). For details see the Appendix and SA.

2 Nonparametric Framework and Estimation For Least Squares Binscatter

2.1 Model and Parameters of Interest

In this section and the next we focus on *least squares binscatter*. This refers to the loss function used in estimation, and therefore, without other covariate adjustment, a binscatter based on least squares naturally provides (a visualization of) an estimate of the conditional mean function: $\mathbb{E}[y_i|x_i]$. Interpretation of $\mathbb{E}[y_i|x_i]$ is straightforward, but in empirical work it is often important to control for additional covariates, $\mathbf{w}_i$, and this complicates interpretation. In essence, we want to visually assess how y_i and x_i relate while "controlling" for $\mathbf{w}_i$ in some precise sense. There is not a universal answer to this problem, and the empirical literature employing binscatter methods is usually imprecise. We will work with an additively separable, semi-linear model given by

$$y_i = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0 + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|x_i, \mathbf{w}_i] = 0. \tag{2.1}$$

This model naturally lends itself to least squares loss where the conditional mean is of direct interest. The structure imposed is not innocuous, but does follow practice closely, has a clear and simple interpretation, and allows our estimator to have more favorable statistical properties. Letting $\mathbf{w}_i$ enter the model fully nonparametrically would make practical implementation prohibitively complicated and would detract focus from the goal of understanding how x_i enters the model and relates to the outcome y_i. However, all our results continue to hold under general misspecification of the function $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$, in which case our results are interpreted as pertaining to the best mean-square approximation to $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$ of the form (2.1). Remark 1 discusses several extensions, including an interactive model (see also Section 3.4). Further, in Section 4 we generalize (2.1) to $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \eta(\mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0)$ for some link function $\eta(\cdot)$, and cover cases such as quantile regression and discrete outcome models, including binary (e. datag., logistic regression) and counts

(e.g., Poisson regression).

The (functional) parameter of interest that we believe is most faithful to the visualization goal is the partial mean effect. For a particular derivative of interest $v \geq 0$ and a user-selected evaluation point $\mathbf{w}$, we define

$$\Upsilon_{\mathbf{w}}^{(v)}(x) := \frac{\partial^v}{\partial x^v}\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}] = \begin{cases} \mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0 & \text{if } v = 0 \\ \mu_0^{(v)}(x) & \text{if } v > 0, \end{cases} \tag{2.2}$$

where $g^{(v)}(x) = \frac{d^v}{dx^v}g(x)$ for any function $g(\cdot)$ whenever well-defined (e.g., one-sided derivative at boundary points). The partial effect function $\Upsilon_{\mathbf{w}}^{(0)}(\cdot)$ itself captures the intuitive notion of the relationship of x_i to y_i, and is the most natural candidate for plotting. Beyond this, $\Upsilon_{\mathbf{w}}^{(v)}(\cdot)$ can be used directly to answer the substantive features generally assessed using binscatter methods. As made precise in Section 3 and shown in the applications in Section 5, we can use $\Upsilon_{\mathbf{w}}^{(v)}(\cdot)$ to formally test if parametric specifications (such as linearity in x_i) are appropriate and shape hypotheses such as monotonicity in the effect of x_i on y_i. Under general misspecification of the function $\mathbb{E}[y_i|x_i, \mathbf{w}_i]$, the first equality of (2.2) continues to define the parameter of interest, and again we interpret our findings as pertaining to the best mean-square approximation of this parameter.

The choice of the evaluation point $\mathbf{w}$ in $\Upsilon_{\mathbf{w}}^{(v)}(x)$ is important for interpretation and for numerical results, and even for the visualization itself. For example, it may be natural to select $\mathbf{w} = \mathbf{0}$, $\mathbf{w} = \mathbb{E}[\mathbf{w}_i]$, or $\mathbf{w} = \text{median}(\mathbf{w}_i)$, with $\mathbf{0}$ denoting a vector of zeros and $\mathbf{w} = \text{median}(\mathbf{w}_i)$ denoting the population median of each component in $\mathbf{w}_i$. Setting the discrete components of $\mathbf{w}$ to a base category (such as zero) is a natural choice, while the others may be more intuitive for continuous controls.

The interpretation of $\Upsilon_{\mathbf{w}}^{(v)}(x)$ changes when $\mathbf{w}$ changes, or indeed when $\mathbf{w}_i$ is coded differently, and along with this, the visualization and uncertainty quantification changes. This is explained in detail in Section 3.2, where we show that the point estimate shifts, testing is delicate, and even visually comparing to a linear fit (or other parametric specification), is affected. We note that the prevalent incorrect residualization, discussed in Section 2.2.1, masks these issues by mishandling the covariates.

The structure of equation (2.1) is the main substantive assumption we require. Aside from this,

we mostly require standard regularity conditions. As detailed in the next section, we will allow for fitting a polynomial of degree $p \geq v$ in each bin (to impose smoothness on the fit or the bands and to estimate derivatives) and we will consequently assume that the underlying functions are smooth enough relative to the desired p and v. Standard binned scatter plots correspond to $p = v = 0$, so the required smoothness is not restrictive. Our full set of assumptions on the data generating process are collected as follows. In applications without additional controls, all conditions involving $\mathbf{w}_i$ may be removed. Let $\|\cdot\|$ denote the Euclidean norm. The SA gives all results under more general, and in some cases weaker, conditions, but this is more notationally involved.

Assumption 1. *The sample $(y_i, x_i, \mathbf{w}_i')$, $i = 1, 2, \ldots, n$, is i.i.d. and satisfies Equation (2.1). The functions $\mu_0(x)$ and $\mathbb{E}[\mathbf{w}_i|x_i = x]$ are $(p+2)$-times continuously differentiable. The covariate x_i has a uniformly Lipschitz continuous density function $f_X(x)$ bounded away from zero on the compact support $\mathcal{X}$. The minimum eigenvalue of $\mathbb{V}[\mathbf{w}_i|x_i = x]$ is uniformly bounded away from zero. $\sigma^2(x) = \mathbb{E}[\epsilon_i^2|x_i = x]$ is uniformly Lipschitz 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.*

Remark 1. It is possible to extend the model (2.1) in several directions. In Section 3.4 we study group-wise comparisons, for example, to test if the relationship between y_i and x_i is the same across groups as in the case of treatment effects. This is a special case of a model with interactions between x_i and (some or all of the controls) $\mathbf{w}_i$, as in

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

with appropriate normalizations for identifiability. Our theory and methods could be extended to this setting for general $\mathbf{w}_i$. We could also consider the case where the term $\mathbf{w}_i'\boldsymbol{\gamma}_0$ of (2.1) represents an (increasing-dimension) basis function approximation to an unknown function $\gamma_0(\cdot)$ of some fixed-dimensional underlying covariates $\mathbf{x}_{2,i}$, so that the model would become $\mathbb{E}[y_i|x_i, \mathbf{x}_{2,i}] = \mu_0(x_i) + \gamma_0(\mathbf{x}_{2,i})$. ⌟

2.2 Estimation: Formalizing Least Squares Binscatter

A binscatter estimate of $\Upsilon_{\mathbf{w}}^{(v)}(x)$ has three key elements: the binning of the support of the covariate x_i, the estimation within each bin, and the way in which the controls $\mathbf{w}_i$ are handled. We discuss each of these in turn.

The partition of the support requires a choice of the number of bins, J, as well as how to divide the space. The choice of J is the tuning parameter of this estimator. In Section 2.3 we describe two methods for selecting J in applications, but for now we take $J < n$ as given. For the spacing of the J bins, we follow standard practice and use the marginal empirical quantiles of x_i. Let $x_{(i)}$ denote the i-th order statistic of the sample $(x_1, x_2, \ldots, x_n)$ and $\lfloor \cdot \rfloor$ denote the floor operator. Then the partitioning scheme is defined as $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, where

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

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. The notation $\widehat{\Delta}$ emphasizes that the partition is estimated from the data. Handling this randomness requires novel nonparametric statistical theory. Our theory can accommodate quite general partitioning schemes, both random and nonrandom, provided high-level conditions are satisfied (see Section SA-1).[1]

Given the partition $\widehat{\Delta}$, which encompasses a choice of the number of bins J, the *canonical* binscatter is the collection of J sample averages of the response variable: for each bin $\widehat{\mathcal{B}}_j$, we obtain $\bar{y}_j = \frac{1}{N_j} \sum_{i=1}^n \mathbb{1}_{\widehat{\mathcal{B}}_j}(x_i) y_i$. Typically, these sample averages are plotted as a "scatter" of points along with another estimate of the regression function $\mathbb{E}[y_i|x_i]$, frequently an ordinary least squares fit. This construction is shown in Figures 1(b) and 1(c).

This procedure is formalized as a nonparametric estimator of $\mathbb{E}[y_i|x_i]$ by recasting it as a piecewise constant fit: $\widehat{\mathbb{E}}[y_i|x_i] = \bar{y}_j$ for all points in bin $\widehat{\mathcal{B}}_j$. This is a series estimator using the Haar basis, or equivalently a zero-degree piecewise polynomial or spline. Precisely, the *canonical* binscatter is

[1] Equally spaced bins is perhaps the next most common, particularly in the nonparametric statistics literature. However, given the ubiquity of quantile binning, we focus on $\widehat{\Delta}$ as defined above.

defined as

$$\widehat{\mathbb{E}}[y_i|x_i] = \widehat{\mathbf{b}}_0(x_i)'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}_0(x_i)'\boldsymbol{\beta})^2, \tag{2.3}$$

where $\widehat{\mathbf{b}}_0(x) = [\mathbb{1}_{\widehat{\mathcal{B}}_1}(x), \mathbb{1}_{\widehat{\mathcal{B}}_2}(x), \cdots, \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)]'$ is the canonical binscatter basis given by a J-dimensional vector of orthogonal indicator variables, that is, the j-th component of $\widehat{\mathbf{b}}_0(x)$ records whether the evaluation point x belongs to the j-th bin in the partition $\widehat{\Delta}$. This piecewise constant fit is shown in Figure 1(d), and from an econometric point of view, is identical to the dots of Figures 1(b) and 1(c).

We generalize canonical binscatter in two ways: controlling for additional variables, $\mathbf{w}_i$ and allowing more flexible basis functions, $\widehat{\mathbf{b}}(x)$. Details on both are given below, but they can be incorporated straightforwardly into the squared loss, following semi-parametric partially linear regression methods matching the model (2.1). We therefore define the p-th order polynomial, $(s-1)$-times continuously differentiable, covariate-adjusted least-squares *extended* binscatter estimator as

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} (y_i - \widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta} - \mathbf{w}_i'\boldsymbol{\gamma})^2, \qquad 0 \le v, s \le p. \tag{2.4}$$

Additional controls, collected in the vector $\mathbf{w}_i$, are often important in empirical work (such as fixed effects). There has been some confusion in the binscatter literature using additional controls, and it is common in empirical work to incorrectly residualize y_i and x_i, as explained in Section 2.2.1.

Second, the additional generality of allowing for other basis functions, beyond piecewise constant, is crucial in estimating derivatives of the function of interest, and thus answering substantive questions such as monotonicity or convexity, as well as reducing the smoothing bias of the estimator. Even the common practice of comparing to a global linear fit requires derivative estimation, as discussed in Section 3.2. The user chooses both a polynomial degree p within each bin, for flexibility, and a smoothness requirement s across bins, thus encompassing piecewise polynomials and splines. It is required that $p \ge v$, so the estimate is sufficiently flexible to capture the derivative of interest (see Section 3 for examples). Further, the choice of $s \le p$, for empirical analyses, both graphical and analytical, reflects a researcher's preference for a binscatter (or associated confidence band) that exhibits some overall smoothness over the support of x_i. For instance, it is natural to construct confidence bands or conduct hypothesis tests about shape restrictions using $s > 0$.

Both $p > 0$ and $s > 0$ are accommodated by changing the basis functions in the squared loss. The extended basis will be denoted $\widehat{\mathbf{b}}(x)$ for notational simplicity, though it depends on p and s (in the SA, to make this explicit we use $\widehat{\mathbf{b}}_{p,s}(x)$). It is derived from the canonical basis $\widehat{\mathbf{b}}_0(x)$ by interacting the set of bin-specific indicators with polynomials of degree p and then imposing the smoothness restrictions. Thus $\widehat{\mathbf{b}}(x) = \widehat{\mathbf{T}}_s[\mathbb{1}_{\widehat{\mathcal{B}}_1}(x), \mathbb{1}_{\widehat{\mathcal{B}}_2}(x), \cdots, \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)]' \otimes [1, x, \cdots, x^p]'$, where $\otimes$ denotes the Kronecker product and $\widehat{\mathbf{T}}_s$ is a $[(p+1)J-(J-1)s] \times (p+1)J$ matrix of linear restrictions ensuring that the $(s-1)$-th derivative of the estimate is continuous; $s = 1$ returns a continuous but nondifferentiable function, while $s = 0$ gives a discontinuous function and $\widehat{\mathbf{T}}_s$ is the identity matrix. The form of $\widehat{\mathbf{T}}_s$ is known, and given in the SA, but it depends on the estimated quantiles and therefore must be handled with care. In this paper, we employ $\widehat{\mathbf{T}}_s$ leading to B-splines, which tend to offer good finite sample properties. For the remainder of the main paper we will focus on p as chosen by the researcher and for ease we set $s = p$. This simplifies notation, but is also natural for visualization. The SA treats the general case of any $s \leq p$.

Finally, to estimate the partial mean effect function $\Upsilon_{\mathbf{w}}^{(v)}(x)$, defined in (2.2), we use a plug-in method based on the extended binscatter (2.4):

$$\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}, \qquad \text{and} \qquad \widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x) = \widehat{\mu}^{(v)}(x) \quad \text{if } v > 1, \tag{2.5}$$

where $\widehat{\mathbf{w}}$ is a consistent estimator of the desired evaluation point $\mathbf{w}$ in (2.2). We always assume that $\widehat{\mathbf{w}}$ is either non-random (e.g., $\mathbf{w}$ is a known fixed value) or generated based on $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_n]'$. If $\mathbb{E}[y_i|x_i, \mathbf{w}_i] = \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$, as in Assumption 1, then $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ will be a consistent estimator of $\Upsilon_{\mathbf{w}}^{(v)}(x)$, and otherwise it is consistent for the pseudo-true value, as discussed above.

To summarize the discussion of this section, the least squares binscatter estimator $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x)$ not only extends canonical binscatter by allowing for v-th order derivative estimation, s-th order across-bin smoothness restrictions, and p-th order within-bin polynomial approximations, but also incorporates covariate adjustment in a principled, interpretable way (cf. Section 2.2.1).

2.2.1 Residualized Canonical Binscatter

We highlight an important methodological mistake with most applications using covariates with binscatter such as the Stata packages `binscatter` and `binscatter2`. Widespread empirical prac-

tice for covariate adjustment proceeds by first regressing out the covariates $\mathbf{w}_i$ from x_i and y_i and then applying canonical binscatter (2.3) to the residualized variables. This approach is heuristically motivated by the usual Frisch–Waugh–Lovell approach to "regressing/partialling out" other covariates in linear regression settings, and is the default implementation of covariate adjustment in commonly used binscatter software. If $\mu_0(x)$ is not linear, this method does not correctly estimate (or visualize) the functions $\mathbb{E}[y_i|x_i]$ or $\mu_0(x)$. The shape of the function and even its support may be incorrect, and therefore can lead to incorrect empirical findings. Further, this incorrect residualizing also mechanically masks the importance of the covariates for quantifying uncertainty and comparing to parametric specifications, as discussed in Section 3.2.

Under mild assumptions, this covariate-residualized binscatter approximates the conditional expectation $\mathbb{E}[y_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}|x_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}]$, where $\widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}$ and $\widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}$ can be interpreted as the best (in mean square) linear approximation to $\mathbb{E}[y_i|\mathbf{w}_i]$ and $\mathbb{E}[x_i|\mathbf{w}_i]$, respectively, with $\widetilde{\mathbf{w}}_i = (1, \mathbf{w}_i')'$. The conditional expectation $\mathbb{E}[y_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{y.\widetilde{w}}|x_i - \widetilde{\mathbf{w}}_i'\boldsymbol{\delta}_{x.\widetilde{w}}]$ is difficult to interpret in general and does not align with standard economic reasoning.

Figure 2 illustrates how different the results can be when using the correct and incorrect residualization. We present both approaches in three empirical settings, with the incorrect approach on the left and our method on the right. The top row uses the insurance rate data of Figure 1. We can immediately see two counterfactual implications of the incorrect residualization: first, we do not observe a decline in uninsurance rates for low incomes which stands in contrast to the presence of Medicaid programs; second, we observe uninsurance rates *rising* for zip codes with higher per capita income (compare this shape to the rate data of Figure 1(a) with no controls). Even if we increase the number of bins, these features of the shape remain. In contrast, using the correct residualization we see both a decline in uninsurance rates at lower incomes and a steady decline at higher incomes as we would expect.

The middle and bottom rows use the applications of Akcigit, Grigsby, Nicholas, and Stantcheva (2022) and Moretti (2021) that we revisit in Section 5. Figures 2(c) and 2(e) replicate the original figures from the two papers but on the correct scale of the data.[2] We will discuss these applications

[2]For the presentation of the results we use the choice of $J = 50$ rather than $J = 100$ used in Akcigit, Grigsby, Nicholas, and Stantcheva (2022) because when the correct residualization is used there is insufficient variation in the variable of interest to feasibly accommodate the larger choice of bins. In Section 5 we present results using the optimal choice of $J = 12$.

in more detail in Section 5 but we would like to make two observations. First, notice the extreme compression of the support of the estimate using the incorrect residualization. This generally comes about because the variability of both the dependent and variable of interest have been overly suppressed. Second, compare the shape of the points to the right column when the correct residualization is used. We can observe a much clearer shape of the estimate of the conditional expectation. Said differently, the bottom two left-hand plots look more like conventional scatter plots, which they are not, rather than an estimate of a function, which they are (though not the conditional mean; see Section 2.2.1). In Section 5 we will demonstrate further how our new tools can improve and sharpen inference in empirical applications using these examples.

In all cases, we overlay the line from a linear regression fit, as is typical. Even this is not straightforward when using covariates. The lines on the left and right in each row are the same. We plot the line on the range of the data itself, not the compressed support of the incorrect residualization. Moreover, the intercept of the line depends on the chosen evaluation point, as discussed in Section 3.2. See Figure 3 as a demonstration of how this can impact the visualization once the covariates are correctly handled.

2.3 Choice of J

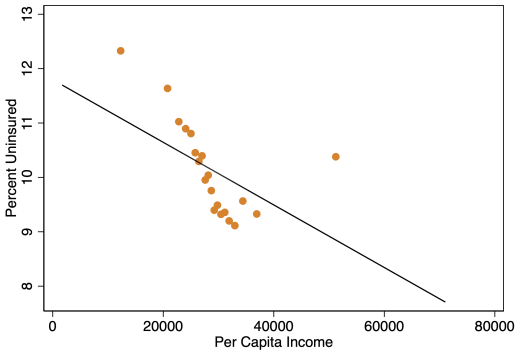
The main tuning parameter for binned scatter plots is the number of bins, J. Consistent nonparametric estimation requires J to diverge with the sample size but not too rapidly, to control both the variance and, along with p, the bias of the estimator. Theorem 1 below formalizes this (a uniform consistency result is given in Corollary SA-2.2, and shares the rate of Theorem 1 up to a factor of $\log(J)$). A wide range of choices for J will, in large sample theory, yield a consistent estimator, but such rate restrictions are not informative enough to guide practice. We therefore consider two practical implementations, reflecting the two goals of binscatter analyses: nonparametric estimation and data visualization.

The starting point of both is studying the density-weighted integrated mean squared error (IMSE) of the plug-in estimator $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x)$, given in (2.5). The proof of this result, and all others, is given in the supplement.

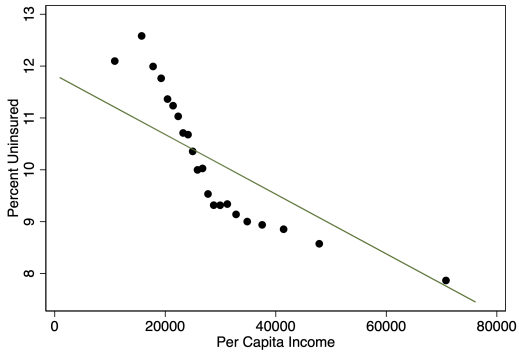
Theorem 1. *Let* $\mathbf{X} = [x_1, \cdots, x_n]'$ *and* $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_n]'$. *Let Assumption 1 hold,* $0 \leq v \leq p$,

Figure 2: **Comparison of Covariate Adjustment Approaches.** This figure demonstrates the implications of covariate adjustment approaches. In the top row, the dependent variable and the independent variable of interest are the same as in Figure 1. The control variables are: percentage of residents with a high school degree, percentage of residents with a bachelor's degree, median age of residents, and the local unemployment rate. The middle row is based on data from Akcigit, Grigsby, Nicholas, and Stantcheva (2022) and the bottom row is based on date from Moretti (2021) (see Section 5 for details.)

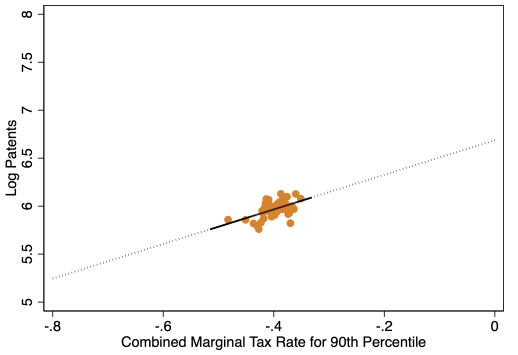
(a) *Residualized Canonical Binscatter*



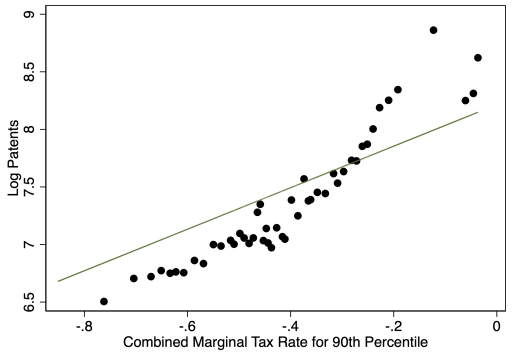
(b) *Semi-linear Canonical Binscatter*



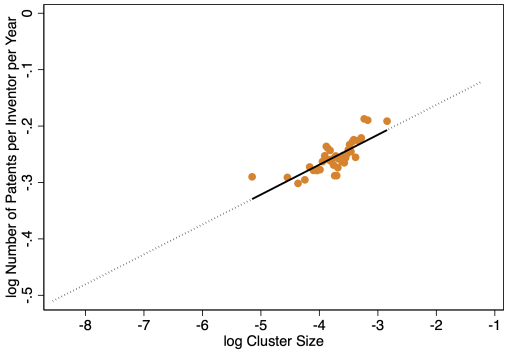
(c) *Residualized Canonical Binscatter*



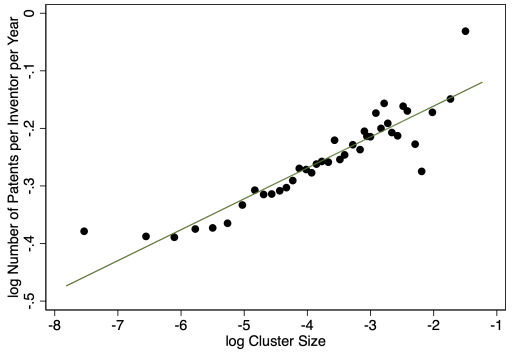
(d) *Semi-linear Canonical Binscatter*



(e) *Residualized Canonical Binscatter*



(f) *Semi-linear Canonical Binscatter*



and $J\log(J)/n \to 0$ *and* $nJ^{-4p-5} \to 0$. *In addition,* $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$. *Then*

$$\int_{\mathcal{X}} \mathbb{E}\Big[\Big(\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)\Big)^2 \Big| \mathbf{X}, \mathbf{W}\Big] f_X(x)dx$$
$$= \frac{J^{1+2v}}{n}\mathscr{V}_n(p,v) + J^{-2(p+1-v)}\mathscr{B}_n(p,v) + o_{\mathbb{P}}\left(\frac{J^{1+2v}}{n} + J^{-2(p+1-v)}\right),$$

where $\mathscr{V}_n(p,v)$ and $\mathscr{B}_n(p,v)$ are non-random, n-varying bounded sequences (see Section SA-2.5).

The terms $\mathscr{V}_n(p,v)$ and $\mathscr{B}_n(p,v)$ capture the asymptotic variance and (squared) bias of binscatter, respectively, as a function of the derivative of interest (v) and the user-selected polynomial order (p). Full details are in the SA, including dependence on the smoothness imposed when $s < p$. All that matters at present is that the variance $\mathscr{V}_n(p,v)$, depending on $\sigma^2(x)$ and $f_X(x)$, is bounded and bounded away from zero under minimal assumptions, while the (squared) bias $\mathscr{B}_n(p,v)$, depending on $\mu_0^{(p+1)}(x)$ and $f_X(x)$, is generally bounded and bounded away from zero except for a very special case (Remark SA-2.6).

By balancing the variance and (squared) bias, we can give an IMSE-optimal choice of J, for any given (p,v) with $p \geq v$. Let $\lceil \cdot \rceil$ denote the ceiling operator. Then

$$J_{\texttt{IMSE}}(p,v) = \left\lceil \left(\frac{2(p-v+1)\mathscr{B}_n(p,v)}{(1+2v)\mathscr{V}_n(p,v)} \right)^{\frac{1}{2p+3}} n^{\frac{1}{2p+3}} \right\rceil. \tag{2.6}$$

Notice that $J_{\texttt{IMSE}}(p,v)$ explicitly accounts for the derivative targeted for estimation and the polynomial order. When it causes no confusion we will write $J_{\texttt{IMSE}}$, but the dependence is important. In general, one expects to use fewer bins and a higher p for derivatives than for levels. Further, although it is not explicit, the best choice of J will depend on the presence and properties of the controls $\mathbf{w}_i$ and the evaluation point $\mathbf{w}$ chosen.

Setting $J = J_{\texttt{IMSE}}(p,v)$ (or its feasible analogue) yields optimal nonparametric estimation and can be used for valid uncertainty quantification (e.g., confidence bands and testing, as in Section 3). However, for visualization this choice may be unappealing. In practice it is common to employ piecewise constant binscatter with a fixed, data-independent number of bins, such as 20 or 50. Such choices may yield attractive and even informative plots, but do not come with any theoretical guarantees.

To balance this visualization-driven choice with the desire for valid estimation and inference, we propose a novel method for selecting the polynomial order p (and along with it, the smoothness s) while keeping J fixed. Let $J = \texttt{J}$ denote the fixed value selected by the user, such as $\texttt{J} = 20$. We then look for p such that $J_{\texttt{IMSE}}(p,v)$ approximates the selected $\texttt{J}$:

$$p_{\texttt{IMSE}}(\texttt{J},v) = \arg\min_{p \in \mathcal{P}} \left| J_{\texttt{IMSE}}(p,v) - \texttt{J} \right|, \tag{2.7}$$

where in principle $\mathcal{P} = \mathbb{N}_0$, but in practice $\mathcal{P} = \{p_{\mathtt{min}}, p_{\mathtt{min}}+1, \ldots, p_{\mathtt{max}}-1, p_{\mathtt{max}}\} \subseteq \mathbb{N}_0$, for some integers $p_{\mathtt{min}} \leq p_{\mathtt{max}}$. The motivation behind this idea is to remove bias: fixing $J = \mathtt{J}$ reduces the flexibility of the nonparametric estimator, and this must be offset by changing p. Analogously to $J_{\mathtt{IMSE}}(p, v)$, the number of bins $\mathtt{J}$ and the derivative of interest are explicit in $p_{\mathtt{IMSE}}(\mathtt{J}, v)$, though we will often omit them. The choice of $\mathtt{J}$, even for visualization, may depend on v. The controls $\mathbf{w}_i$ also matter here.

Feasible implementation of $J_{\mathtt{IMSE}}$ and $p_{\mathtt{IMSE}}$ is straightforward, and we defer details to Section SA-5. The high-level message is that under mild regularity conditions a feasible $\widehat{J}_{\mathtt{IMSE}}$ satisfies $\widehat{J}_{\mathtt{IMSE}}/J_{\mathtt{IMSE}} \to_{\mathbb{P}} 1$, where $\to_{\mathbb{P}}$ denotes convergence in probability. Then $\widehat{p}_{\mathtt{IMSE}}$ may be found analogously to $p_{\mathtt{IMSE}}$, using $\widehat{J}_{\mathtt{IMSE}}$ in (2.7).

3 Characterizing Uncertainty and Testing Substantive Hypothesis

In this section we provide both analytical and visualization tools to capture the uncertainty around the mean estimate $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ (or $\widehat{\mu}^{(v)}(x)$ if $\mathbf{w} = 0$ or no additional controls are used). These tools depend on the underlying variance in the data and the heteroskedasticity pattern, and while the visualizations do reflect these quantities, they are not directly shown. This is exactly analogous to how a simple confidence interval for the mean reflects only estimation uncertainty about the parameter, even though the interval depends on the variance of the data. For visualizing the "spread" and detecting outliers conditional quantiles may be more useful; see Section 4.

Importantly, all our inference is uniform over $x \in \mathcal{X}$, as opposed to pointwise for a given x. This uniformity is required both to answer the substantive questions of interest in empirical work and to provide an honest visualization of the uncertainty. The uniform inference theory is a major technical contribution of this paper. Section 3.1 sketches our key theoretical contributions and points to further details in the online supplement.

Section 3.1 is technical; the tools for applied work are discussed in Sections 3.2, 3.3, and 3.4. §3.2 tests for substantive hypotheses about $\mu_0(x)$, focusing on testing parametric specifications (e.g., linearity) and shape restrictions (e.g., monotonicity). Therein we also discuss the important and subtle role of the controls $\mathbf{w}_i$ in uncertainty quantification. Our main visualization tools are confidence bands for $\Upsilon^{(0)}_{\mathbf{w}}(x)$ or $\mu^{(v)}_0(x)$, described in §3.3. §3.4 gives an extension to multi-sample

comparisons. Finally, §3.5 illustrates the ideas using the uninsurance data behind Figure 1.

3.1 Theoretical Foundation

The starting point of our theoretical analysis is the Studentized t-statistic that centers and scales the estimator $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ (recall the definitions in (2.4) and (2.5)). We index important objects with p (recall we let $s = p$ in the paper, but the SA treats the general case). We study

$$T_p(x) = \frac{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) - \Upsilon^{(v)}_{\mathbf{w}}(x)}{\sqrt{\widehat{\Omega}(x)/n}}, \tag{3.1}$$

where $\widehat{\Omega}(x) = \widehat{\mathbf{b}}^{(v)}(x)'\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{Q}}^{-1}\widehat{\mathbf{b}}^{(v)}(x)$, $\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{b}}(x_i)\widehat{\mathbf{b}}(x_i)'$, and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{b}}(x_i)\widehat{\mathbf{b}}(x_i)'(y_i - \widehat{\mathbf{b}}(x_i)'\widehat{\boldsymbol{\beta}} - \mathbf{w}_i'\widehat{\boldsymbol{\gamma}})^2$. We seek a distributional approximation for the entire stochastic process $\{T_p(x) : x \in \mathcal{X}\}$ because this allows us to study the visualization and econometric properties of the entire binscatter fit $\{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) : x \in \mathcal{X}\}$ simultaneously (but the SA also gives pointwise results). Using this strong approximation we can compute the critical value $\mathfrak{c}$ for valid hypothesis testing and constructing confidence bands. We will see how our strong approximation gives a simple, tractable method for computing this critical value based on draws from the Normal distribution.

Our novel strong approximations make two important contributions compared to the prior theoretical literature, both of which are needed to study binned scatter plots. First, we have optimal rate conditions (up to $\log(n)$ terms), and therefore our results are sharp enough to allow for (possibly covariate-adjusted) canonical binscatter ($p = 0$), a result excluded by the prior literature. Second, we allow for random partitions (from the binning based on empirical quantiles). These two are the most salient improvements on the prior literature, but not the only ways in which our theory is innovative. Appendix A gives a complete summary of how our technical work improves on the non-/semi-parametric least squares series estimation literature. Further details are given for each result in Section SA-2 for least squares binscatter, and Sections 4 and SA-3 move beyond least squares loss to cover nonlinear and nonsmooth losses.

The randomness of the partition $\widehat{\Delta}$ (and thus in the basis functions themselves) is not just ruled out by the assumptions of prior work, but rather it is not even possible to obtain a valid strong approximation for the entire stochastic process $\{T_p(x) : x \in \mathcal{X}\}$ exactly because the randomness

causes uniformity to fail. As an alternative, we establish a conditional Gaussian strong approximation as the key building block for uniform inference. Heuristically, our strong approximation begins by establishing the following two approximations uniformly over $x \in \mathcal{X}$:

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

where $\mathbf{N}^{\star}_{p+J}$ denotes a $(p+J)$-dimensional standard Gaussian random vector, independent of the data. The first approximation is a stochastic linearization (Theorem SA-2.1) and directly implies the variance formula $\widehat{\Omega}(x)$. This step is reminiscent of standard least squares algebra. The second approximation corresponds to a conditional coupling (Theorems SA-2.4 and SA-2.5). It is not difficult to show that $\widehat{\mathbf{Q}}$ and $\widehat{\mathbf{\Sigma}}$ are sufficiently close in probability to well-defined non-random matrices in the necessary norm (Lemma SA-2.1 and Theorem SA-2.2). However, $\widehat{\mathbf{b}}^{(v)}(x)$ fails to be close in probability to its non-random counterpart *uniformly* in $x \in \mathcal{X}$ due to the sharp discontinuity introduced by the indicator functions entering the binning procedure. Nevertheless, inspired by the work in Chernozhukov, Chetverikov, and Kato (2014a,b), our approach circumvents that technical hurdle by developing a strong approximation that is conditionally Gaussian first, retaining some of the randomness introduced by $\widehat{\Delta}$, and then using such coupling to deduce a distributional approximation for specific functionals of interest (e.g., suprema); see Section SA-4.1 for details.

We state the formal results in two steps: the first derives an infeasible strong approximation and the second shows that, given the data, a feasible version can be constructed.

Theorem 2 (Feasible Strong Approximation)**.** *Let Assumption 1 hold and let* $\{a_n : n \geq 1\}$ *be a sequence of non-vanishing constants such that* $n^{-1/2}J(\log J)^2 + J^{-1} + nJ^{-2p-3} = o(a_n^{-2})$*. Assume that* $\|\widehat{\mathbf{w}} - \mathbf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$*. Then, on a properly enriched probability space, there exists a standard Gaussian random vector* $\mathbf{N}_{p+J}$*, of length* $p+J$*, such that for any* $\xi > 0$*,*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_p(x) - Z_p(x)| > \xi a_n^{-1}\Big) = o(1), \qquad Z_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}(x)'\mathbf{Q}_0^{-1}\mathbf{\Sigma}_0^{1/2}}{\sqrt{\Omega(x)}}\mathbf{N}_{p+J}.$$

Also, there exists a standard Gaussian random vector $\mathbf{N}^{\star}_{p+J}$*, of length* $p+J$*, independent of the*

data $\mathbf{D} = \{(y_i, x_i, \mathbf{w}_i') : i = 1, 2, \ldots, n\}$, *such that for any* $\xi > 0$,

$$\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\widehat{Z}_p(x) - Z_p(x)| > \xi a_n^{-1} \Big| \mathbf{D} \Big) = o_{\mathbb{P}}(1), \qquad \widehat{Z}_p(x) = \frac{\widehat{\mathbf{b}}^{(v)}(x)' \widehat{\mathbf{Q}}^{-1} \widehat{\mathbf{\Sigma}}^{1/2}}{\sqrt{\widehat{\Omega}(x)}} \mathbf{N}^{\star}_{p+J}.$$

This result forms the basis of the inference tools in the following subsections. In principle, we can now approximate the distribution of any functional of the t-statistic process $T_p(x)$ using a plug-in approach based on $\widehat{Z}_p(x)$. This prescription is easy to put into practice, because it depends only on Gaussian draws and the already-computed elements $\widehat{\mathbf{b}}(x)$, $\widehat{\mathbf{Q}}$, $\widehat{\mathbf{\Sigma}}$, and $\widehat{\Omega}(x)$, and therefore the process $\widehat{Z}_p(x)$ is simple to simulate. For example, the distribution of $\sup_{x \in \mathcal{X}} \left|T_p(x)\right|$ is well approximated by that of $\sup_{x \in \mathcal{X}} \left|\widehat{Z}_p(x)\right|$, conditional on the data, and we can use this to obtain critical values for testing or forming confidence bands, as shown in the next subsections.

However, and crucially for applied practice, one must choose J such that the approximation is valid. In addition, ideally, the choice of J would be optimal in some way and the resulting inference would be robust to small fluctuations in J. The IMSE-optimal choice $J_{\texttt{IMSE}}(p, v)$ cannot be directly used, as it is too "small" to remove enough bias for the t-statistic $T_p(x)$ to be correctly centered. To address this problem, we rely on robust bias correction (Calonico, Cattaneo, and Farrell, 2018) to form valid uniform inference based on an IMSE-optimal binscatter, that is, without altering the partitioning scheme $\widehat{\Delta}$ used. Further, robust bias-corrected inference is well-documented to be more robust to the choice of J. For a choice of p, we construct the binscatter (point) estimate $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ using either method of Section 2.3, and the implied $\widehat{\Delta}$, and then for inference we employ $T_{p+1}(x)$ (recall that we take $s = p$, so $T_{p+1}(x)$ uses $s + 1$).

Finally, we note that from a theoretical point of view, the rate conditions of Theorem 2 are seemingly minimal and improve on prior results. In fact, it can be shown that when $a_n = \sqrt{\log n}$ and a subexponential moment restriction holds for the error term, it suffices that $J/n = o(1)$, up to $\log n$ terms. In contrast, a strong approximation of the t-statistic process for general series estimators was obtained based on Yurinskii coupling in Belloni, Chernozhukov, Chetverikov, and Kato (2015), which requires $J^5/n = o(1)$, up to $\log n$ terms. Alternatively, a strong approximation of the *supremum* of the t-statistic process can be obtained under weaker rate restrictions, such as the requirement of $J/n^{1-2/\nu} = o(1)$ (up to $\log n$) used by Chernozhukov, Chetverikov, and Kato (2014a), but this result applies exclusively to the suprema of the stochastic process. Our

theoretical improvements have direct practical consequences as the rate conditions are weak enough to accommodate the canonical binascatter (i.e., the piecewise constant ($p = 0$) estimator), which would otherwise not be possible. See Appendix A for more information.

3.2 Hypothesis Testing: Parametric Specifications and Shape Restrictions

We first utilize our strong approximation theory to give a formal treatment of what is perhaps the most common use of binned scatter plots: assessing the functional form of $\mu_0(x)$. Typically a binscatter will precede an ordinary least squares regression of y_i on x_i and $\mathbf{w}_i$. The idea is that if the binscatter "dots" are roughly on a line, then $\mu_0(x)$ is approximately linear and regression analysis is justified. Beyond evaluating the evidence for linearity, binned scatter plots are also utilized to assess other shape restrictions (see, for example, Shapiro and Wilson (2021) or Feigenberg and Miller (2021)). In this section we provide a rigorous formulation of this idea.

We must be careful with the controls $\mathbf{w}_i$ in two ways. First, and most obviously, we must include them correctly, as discussed in Section 2.2.1. But second, and more subtly, the user-selected point of evaluation $\mathbf{w}$ will matter, as will the estimates of the coefficients $\boldsymbol{\gamma}_0$. It is important to emphasize that these issues matter not just for formal testing, but even for the standard informal procedure of putting a parametric fit over a binned scatter plot (see Figure 3).

One would think that $\mathbf{w}$ and $\boldsymbol{\gamma}_0$ should not matter, because we are interested only in testing how x enters $\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$. But this intuition overlooks the fact that the function $\mu_0(x)$ is only defined relative to how $\mathbf{w}_i$ is coded. We will first show why $\mathbf{w}$ and $\boldsymbol{\gamma}_0$ matter when the test is designed to exactly match the common visualization analysis. We then use this to argue for an alternative testing procedure, based on derivatives of $\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$, which satisfies the original intuition. This shows why having inference results for derivatives is crucial for practice: 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.

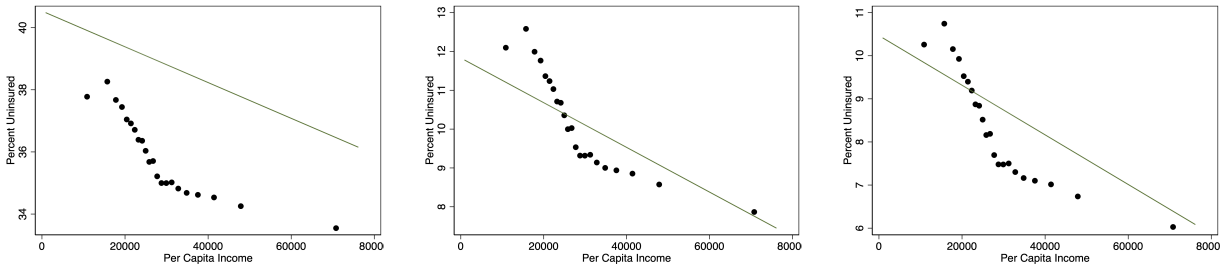
To see the problem, let us first formalize the hypothesis testing procedure behind the informal practice of checking if the "dots" are roughly linear, and then running ordinary least squares regression of y_i on x_i and $\mathbf{w}_i$. This idea motivates the standard practice of plotting the fitted regression line along with the binned scatter plot, as in Figures 1 and 2. In this case, the null

hypothesis is *not* merely that $\mu_0(x) = \theta_0 + \theta_1 x$, i.e. a linear function, but rather that the full model is linear, so that $\Upsilon_{\mathbf{w}}(x) = \theta_0 + x\theta_1 + \mathbf{w}'\boldsymbol{\gamma}_0$ (cf (2.2)). Under the partially linear assumption of the model (2.1), these would seem identical, because in either case $\mathbf{w}$ enters linearly. But this is not so in practice for two reasons: the estimates of the coefficients $\boldsymbol{\gamma}_0$ will differ in general, as will the implied intercepts, and the chosen $\mathbf{w}$ will impact the uncertainty about θ_0. These issues will be important for visualizing uncertainty as well, as discussed below in Section 3.3.

The "dots" show the semiparametric estimate $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}$, defined in (2.5), while the plotted line is obtained from the parametric fit $\widetilde{\theta}_0 + x\widetilde{\theta}_1 + \widehat{\mathbf{w}}'\widetilde{\boldsymbol{\gamma}}$, obtained from least squares regression. Thus, while we are only interested in assessing the linearity of $\mu_0(x)$, we are *actually* testing these two functional forms for $\Upsilon_{\mathbf{w}}(x)$, and the fact that $\widehat{\boldsymbol{\gamma}} \neq \widetilde{\boldsymbol{\gamma}}$ becomes important. Moreover, because $\widetilde{\theta}_0 + x\widetilde{\theta}_1 + \widehat{\mathbf{w}}'\widetilde{\boldsymbol{\gamma}}$ is a global parametric fit while $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x) = \widehat{\mu}(x) + \widehat{\mathbf{w}}'\widehat{\boldsymbol{\gamma}}$ is local and nonparametric, the implied intercept when plotted depends the chosen $\widehat{\mathbf{w}}$, and this can shift the line away from the dots. Figure 3 demonstrates this by example: everything is identical between the three plots except for choice of $\widehat{\mathbf{w}}$. Notice the shift in absolute position (note the y axis) and the change in the relative position of the line and the binscatter. This phenomenon is unavoidable in this setting, and the user must select $\widehat{\mathbf{w}}$ appropriately. (Again, this does not occur when using the incorrect residualization because the covariates are mishandled.)

Figure 3: **Role of the Evaluation Point.** This figure demonstrates that the choice of $\widehat{\mathbf{w}}$ shifts both the absolute position (note the y axis) of the visualization and estimator, but also affects the comparison to parametric fits. The data is the same as in Figure 1.

(a) $\widehat{\mathbf{w}} = \mathbf{w}_{min}$ (b) $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$ (c) $\widehat{\mathbf{w}} = \mathbf{w}_{max}$



Beyond the visual inspection of a plot like Figure 3, we wish to formally assess if $\mu_0(x) = m(x;\boldsymbol{\theta})$ where $m(x;\boldsymbol{\theta})$ is a function known up to a finite dimensional parameter $\boldsymbol{\theta}$. (In the case of linearity, $\boldsymbol{\theta} = (\theta_0, \theta_1)'$ and $m(x;\boldsymbol{\theta}) = \theta_0 + x\theta_1$.) Under the null $\Upsilon_{\mathbf{w}}(x) = M_{\mathbf{w}}(x;\boldsymbol{\theta},\boldsymbol{\gamma}_0) = m(x;\boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}_0$.

The testing problem that corresponds exactly to the typical visual inspection is

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

Then, assuming that there exists an estimator $(\widetilde{\boldsymbol{\theta}}', \widetilde{\boldsymbol{\gamma}}')'$ that consistently estimates $(\boldsymbol{\theta}', \boldsymbol{\gamma}_0')'$ under the null hypothesis and is well behaved under the alternative, we can construct the appropriate test statistic,

$$\dot{T}_p(x) = \frac{\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x) - M_{\widehat{\mathbf{w}}}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}}, \tag{3.2}$$

and conduct the specification test as follows:

$$\text{Reject } \dot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} |\dot{T}_p(x)| \geq \mathfrak{c}, \tag{3.3}$$

where for a chosen level α the critical value is $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+1}(x)| \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$. (All this is made precise in Theorem 3 below.)

The testing procedure defined by (3.2) and (3.3) formalizes the idea of visually examining a binned scatter plot compared to a parametric specification; a common step before regression analysis. But it also formalizes the problematic dependency on the evaluation point $\mathbf{w}$ and the difference between $\widehat{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$. First, despite the fact that $\mathbf{w}'\boldsymbol{\gamma}_0$ cancels out in both the null and alternative statements, the numerator of the t-statistic depends on $\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}})$, because in finite samples $\boldsymbol{\gamma}_0$ is unknown. Therefore our uncertainty about how x enters the model depends on the controls $\mathbf{w}_i$. This is in fact unavoidable because $\mu_0(x)$ is only defined relative to $\mathbf{w}_i$.

Consider the case where $\mathbf{w}_i$ is an indicator (or fixed effect). Then setting $\widehat{\mathbf{w}} = \mathbf{0}$ would seem to remove the problem, because the numerator of $\dot{T}_p(x)$ depends only on $\widehat{\mu}(x)$ and $m(x;\widetilde{\boldsymbol{\theta}})$, while setting $\widehat{\mathbf{w}} = \mathbf{1}$ maximizes it. This is correct, but is then sensitive to how the researcher has coded $\mathbf{w}_i$, i.e., which category is considered the baseline. Thus we can get a different answer to the test depending on which category of $\mathbf{w}$ we consider, even though the hypothesis applies to both. This is intuitively the same as the fact that in a linear model with dummy variables the standard error of the intercept changes depending on how $\mathbf{w}$ is coded. The case of a continuous $\mathbf{w}_i$ (especially

with large support, such as annual income) is perhaps worse: if $\widehat{\boldsymbol{\gamma}} \neq \widetilde{\boldsymbol{\gamma}}$, then there is *always* some value $\widehat{\mathbf{w}}$ for which we reject the null. Thus, using (3.2) and (3.3) to test parametric specifications is potentially confusing at best, and at worst is vulnerable to p-hacking. It is worth noting in most papers studying the partially linear model, the parameter of interest is $\boldsymbol{\gamma}_0$, and so these concerns have gone largely unnoticed. (And are masked by construction when using the incorrect residualization approach.)

To avoid these issues, and motivated by the fact that the central point of binscatter is to study how y_i relates to x_i, controlling for $\mathbf{w}_i$, we advocate reformulating the hypothesis as pertaining to the *derivative* of $\mu_0(x)$, instead of the level. Under the partially linear model maintained throughout, any derivative of $\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$ is exactly $\mu_0^{(v)}(x)$, and is by definition $\Upsilon_{\mathbf{w}}^{(v)}(x)$; see (2.2). Therefore, instead of testing the null $\Upsilon_{\mathbf{w}}(x) = M_{\mathbf{w}}(x;\boldsymbol{\theta},\boldsymbol{\gamma}_0) = m(x;\boldsymbol{\theta}) + \mathbf{w}'\boldsymbol{\gamma}_0$, we test the equivalent hypothesis that $\Upsilon_{\mathbf{w}}^{(v)}(x) = M_{\mathbf{w}}^{(v)}(x;\boldsymbol{\theta},\boldsymbol{\gamma}_0) = m^{(v)}(x;\boldsymbol{\theta})$ for some $v \geq 1$. For example, instead of testing that $\mu_0(x)$ is linear, we test that it has constant first derivative. To test if $\mu_0(x)$ itself is constant, the null would be that $\mu_0^{(1)}(x) = m^{(1)}(x;\boldsymbol{\theta}) = 0$.

In general, we obtain this (more robust) testing problem:

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

In this case, the appropriate test statistic is, with $v \geq 1$,

$$\ddot{T}_p(x) = \frac{\widehat{\Upsilon}_{\widehat{\mathbf{w}}}^{(v)}(x) - M_{\widehat{\mathbf{w}}}^{(v)}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\Omega}(x)/n}} = \frac{\widehat{\mu}^{(v)}(x) - m^{(v)}(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}(x)/n}}, \tag{3.4}$$

where the dependence on $\widehat{\mathbf{w}}$, $\widehat{\boldsymbol{\gamma}}$, and $\widetilde{\boldsymbol{\gamma}}$ is gone. We conduct the test as follows:

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

for an appropriate critical value. This procedure is formalized in the following result, which leans on Theorem 2 to use $\widehat{Z}_p(x)$ to approximate the distribution of $\sup_{x\in\mathcal{X}} |\ddot{T}_p(x)|$. Note that (3.1) and Theorem 2 are fully general in the derivative v. Recall that for inference we use robust bias

correction.

Theorem 3 (Hypothesis Testing: Parametric Specification)**.** *Let Assumption 1 hold and* $J = J_{\texttt{IMSE}}(p,v)$*. Also, assume* $\widehat{\mathbf{w}}$ *is a* $\sqrt{n}$*-consistent estimator of the evaluation point* $\mathbf{w}$ *in (2.2). Let* $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} |\widehat{Z}_{p+1}(x)| \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$*. Then under* $\ddot{\mathsf{H}}_0$*, if* $\sup_{x\in\mathcal{X}} |M^{(v)}_{\widehat{\mathbf{w}}}(x;\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\gamma}}) - \Upsilon^{(v)}_{\mathbf{w}}(x)| = O_{\mathbb{P}}(n^{-1/2})$*, then*

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

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

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

The conditions on $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ are mild and satisfied in most standard parametric models used in applied economics (including linear and nonlinear regression, discrete choice, censored and truncation models, and more). We essentially require $\boldsymbol{\theta}$ to be $\sqrt{n}$-estimable, provided some mild regularity holds for the known regression function $m(x;\boldsymbol{\theta})$. For example, a simple sufficient condition is $\sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = O_{\mathbb{P}}(1)$ and $m^{(v)}(x,\boldsymbol{\theta})$ is continuous in x and continuously differentiable in $\boldsymbol{\theta}$.

So far we have discussed only parametric specifications, but binned scatter plots can also be used to assess other shape restrictions about $\Upsilon^{(v)}_{\mathbf{w}}(x)$ or $\mu_0(x)$ that cannot be implemented using Theorem 3. That result dealt exclusively with two-sided null hypotheses, but features such as negativity, monotonicity, and concavity of $\Upsilon_{\mathbf{w}}(x)$ all correspond to one-sided statements, respectively given by $\Upsilon_{\mathbf{w}}(x) \leq 0$, $\Upsilon^{(1)}_{\mathbf{w}}(x) \leq 0$, and $\Upsilon^{(2)}_{\mathbf{w}}(x) \leq 0$. To enable testing of this type of shape restrictions, we will also give results for the following null and alternative hypotheses:

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

For this testing problem, the test statistic is

$$\dddot{T}_p(x) = \frac{\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)}{\sqrt{\widehat{\Omega}(x)/n}},$$

and we then conduct the test as:

$$\text{Reject } \dddot{\mathsf{H}}_0 \qquad \text{if and only if} \qquad \sup_{x\in\mathcal{X}} \dddot{T}_p(x) \geq \mathfrak{c}, \tag{3.6}$$

again for an appropriate choice of critical value $\mathfrak{c}$ to control false rejections (Type I error). Because of its one-sided nature, this test is conservative in general. Of course, the other one-sided hypothesis tests are constructed in the obvious symmetric way. This procedure is formalized and validated in the following result. The key idea again is to use Theorem 2 to approximate the distribution of the appropriate functional of $\dddot{T}_p(x)$.

Theorem 4 (Hypothesis Testing: Shape Restrictions)**.** *Let Assumption 1 hold and* $J = J_{\texttt{IMSE}}(p, v)$*. Also, assume* $\widehat{\mathbf{w}}$ *is a* $\sqrt{n}$*-consistent estimator of the evaluation point* $\mathbf{w}$ *in* (2.2)*. Let* $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} \widehat{Z}_{p+1}(x) \leq c \;\big|\; \mathbf{D}\big] \geq 1-\alpha\big\}$*. Then under* $\dddot{\mathsf{H}}_0$*,*

$$\lim_{n\to\infty} \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \dddot{T}_{p+1}(x) > \mathfrak{c}\Big] \leq \alpha,$$

and under $\dddot{\mathsf{H}}_A$*,*

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

Theorem 4 corresponds to the one-sided "left" hypothesis test, but of course the analogous theorem "to the right" also holds. Our software allows for all three possibilities: one-sided (left or right) and two-sided hypothesis testing. Finally, in SA-4.3 we include a more general version of Theorem 4, where we introduce one-sided tests against a parametric fit. For example, setting $v = 0$ and abstracting from covariate adjustment for simplicity, this more involved testing procedure might be useful to assess whether $\mu_0(x)$ always resides "below" the line defined by $m(x, \bar{\theta}; a) = a + x\bar{\theta}$, when $\widetilde{\theta}$ is the OLS estimator based on y_i and x_i, $\bar{\theta}$ is its probability limit (i.e., $x\bar{\theta}$ denotes the best linear predictor of y_i and x_i), and a is a user-chosen positive constant.

Remark 2 (Other Metrics)**.** The tests above are based on the maximum (uniform) discrepancy between one binscatter fit and possibly some parametric fit. Some practitioners, however, may prefer to assess the discrepancy by means of an alternative metric. For instance, one can construct a testing procedure using the mean squared difference between the parametric and nonparametric

fits:

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

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

⌟

3.3 Confidence Bands

The testing procedures in Section 3.2 provide analytical tools to accompany binned scatter plots, but do not enhance the visualization. For that, we now turn to confidence bands. Loosely speaking, a confidence band is simply a confidence "interval" for a function, and like a traditional confidence interval, it is given by the area between two "endpoints", which are now functions (of x), say $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x;\mathtt{U})$ and $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x;\mathtt{L})$. In repeated samples the area between them covers the true parameter function $\Upsilon^{(v)}_{\mathbf{w}}(x)$ with a prespecified probability. Heuristically, a plotted confidence band shows the plausible functions compatible with the data at hand. Visually, the size of the band reflects the uncertainty in the data, both in terms of overall sampling uncertainty and any heteroskedasticity patterns, as it pertains to estimation of $\Upsilon^{(v)}_{\mathbf{w}}(x)$. Thus it is an important step in restoring to binscatter the ability to visually assess uncertainty, just as in a classical scatter plot. One may also wish to assess the spread in the outcomes directly, and for this, the conditional quantiles of Section 4 are useful.

We construct confidence bands that are dual to the t statistic (3.1). Also, because confidence bands are for quantifying uncertainty, we employ robust bias correction. We therefore construct the band $\{\widehat{I}_{p+1}(x) : x \in \mathcal{X}\}$ as

$$\widehat{I}_{p+1}(x) = \Big[\; \widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x) \pm \mathfrak{c} \cdot \sqrt{\widehat{\Omega}(x)/n} \;\Big], \quad \text{with} \quad \mathfrak{c} = \inf\Big\{c \in \mathbb{R}_+ : \mathbb{P}\Big[\sup_{x\in\mathcal{X}} \big|T_{p+1}(x)\big| \leq c\Big] \geq 1-\alpha\Big\}. \tag{3.7}$$

The quantile $\mathfrak{c}$ as shown is infeasible because the distribution of $\sup_{x\in\mathcal{X}} \big|T_{p+1}(x)\big|$ is unknown, but again we can use Theorem 2 to approximate it, just as we did with testing. These ideas are formalized as follows.

Theorem 5 (Confidence Bands)**.** *Suppose Assumption 1 holds,* $\widehat{\mathbf{w}}$ *is a* $\sqrt{n}$*-consistent estimator of*

the evaluation point $\mathbf{w}$ *in* (2.2)*, and* $J = J_{\texttt{IMSE}}(p, v)$ *for a fixed* (p, v)*. Then*

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

provided that $\mathfrak{c} = \inf\big\{c \in \mathbb{R}_+ : \mathbb{P}\big[\sup_{x\in\mathcal{X}} \big|\widehat{Z}_{p+1}(x)\big| \leq c \,\big|\, \mathbf{D}\big] \geq 1-\alpha\big\}$,

Theorem 5 shows how to add valid confidence bands to any binned scatter plot, whether for the level of the function $\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$ or its derivatives (with respect to x). This visual assessment of uncertainty is an important step in any analysis, just like reporting uncertainty numerically in a regression analysis. It is common to follow a binned scatter plot with a regression analysis that quantifies the relationship between y_i and x_i. All regression parameters are reported along with their standard errors or other measures of uncertainty; this is good empirical practice. In the same way, and for the same reasons, adding confidence bands to binned scatter plots is good practice: the reader can see not only the estimate of the relationship (the "dots" of the binscatter) but the uncertainty around this estimate.

The visual appearance of the band will be impacted by the evaluation point $\mathbf{w}$ (or its feasible version $\widehat{\mathbf{w}}$) as discussed above. This is important to keep in mind when evaluating these plots. By definition each plot shows only one choice of $\mathbf{w}$, and therefore while the shape of $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x)$ is unchanged, a level shift will occur and the size of the band can change. For an intuitive example, consider the case where $\mathbf{w}$ is categorical, and some categories have much larger or smaller sample sizes. These different sample sizes will naturally be reflected in the uncertainty for $\mathbb{E}[y_i|x_i = x, \mathbf{w}_i = \mathbf{w}]$.

For this reason, we must be careful when using confidence bands as visual aids in parametric specification testing. If we plot $\widehat{\Upsilon}_{\widehat{\mathbf{w}}}(x)$ and its associated confidence band, it is tempting to say that if this band does not contain a line (or quadratic function), then we say that at level α we reject the null hypothesis that $\mu_0(x)$ is linear (or quadratic). Although this is formally justified, we must interpret such analyses with caution because of the evaluation point as discussed above.

3.4 Extension to Multi-Sample Estimation and Testing

Beyond studying the relationship between y_i and x_i in general, researchers often want to compare mean, quantile, and other effects across different groups, after controlling for $\mathbf{w}_i$. This is a common goal in program evaluation and causal inference settings, where the groups are defined by treatment

arms, and the differences define heterogeneous (in x_i) effects. Our results extend naturally to this setting.

For a grouping, or treatment arm, variable $t_i = \{0, \ldots, L\}$, we extend the partially linear model of (2.1) to allow for general interactions:

$$y_i = \sum_{t=0}^{L} \mathbb{1}(t_i = t) \left(\mu_{0,t}(x_i) + \mathbf{w}_i' \boldsymbol{\gamma}_{0,t} \right) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i | x_i, \mathbf{w}_i, t_i] = 0.$$

The parameter of interest is then $\Upsilon^{(v)}_{\mathbf{w},t}(x)$, defined as in (2.2) but for the subsample $t = 0, 1, 2, \ldots, L$. For example, in a randomized experiment with a binary treatment, $\Upsilon^{(v)}_{\mathbf{w},0}(x)$ and $\Upsilon^{(v)}_{\mathbf{w},1}(x)$ may denote the partial mean effect for control and treatment units, respectively, and thus $\Upsilon^{(v)}_{\mathbf{w},1}(x) - \Upsilon^{(v)}_{\mathbf{w},0}(x)$ can be interpreted as the average treatment effect conditional on $x_i = x$ and $\mathbf{w}_i = \mathbf{w}$. The latter parameter naturally captures treatment effect heterogeneity along the x_i dimension.

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

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

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

All of these testing problems are subject to the same concerns regarding the chosen evaluation point $\mathbf{w}$ discussed above. For example, by setting $\mathbf{w} = \mathbf{0}$, the null $\mathsf{H}_0 : \Upsilon^{(v)}_{\mathbf{w},0}(x) = \Upsilon^{(v)}_{\mathbf{w},1}(x)$ focuses only on the nonparametric component, $\mathsf{H}_0 : \mu^{(v)}_{0,1}(x) = \mu^{(v)}_{0,0}(x)$. However, this is now sensitive to the coding of $\mathbf{w}$: the test will change depending on which category is labeled as zero. Further, the

value $\mathbf{w} = \mathbf{0}$ may not be appropriate for all controls (such as age). These issues are unavoidable; researchers must be careful when implementing the tests and interpreting the results.

3.5 Empirical Illustration

We now illustrate our new inference procedures using data on uninsurance rates versus per capita income across U.S. zip codes. Heuristically, our formal parametric specification testing approach is based on comparing the maximal empirical deviation between the binscatter and the desired parametric specification for $\mu_0(x)$. If the parametric specification is correct, then there should be no deviation beyond what is explained by random sampling for all evaluation points x. Hence, when there are no covariates included in the model we may compare directly the parametric fit and the confidence band for $\mu_0(x)$ as in the left plot of Figure 4. We clearly see that the linear fit and the cubic fit are not fully enveloped by the confidence band. This is confirmed by the inference results shown in Table 1 when no covariates are included in the model. We strongly reject both the linear regression model and the cubic regression model; moreover, we also strongly reject when we use the L_2 metric as discussed in Remark 2.[3] When we add covariates to the model, we must be mindful of the dependence on the evaluation point of $\mathbf{w}$ as we have discussed earlier. The right plot of Figure 4 presents the visualization of the test based on comparing the confidence band and the parametric fit when $\widehat{\mathbf{w}}$ is chosen as $\bar{\mathbf{w}}$. Clearly, both fitted parametric functions lie outside the confidence band for some income levels. This corresponds directly to the middle row of each of the top two panels in Table 1 which report the test statistic in equation (3.2) for these data. We again reject the parametric forms for both choices of distance metric. However, to demonstrate how inference results may be sensitive to the choice of $\mathbf{w}$, the table also includes test statistics and associated p-values when $\widehat{\mathbf{w}}$ is evaluated at the marginal minimum $\mathbf{w}_{min}$ and maximum $\mathbf{w}_{max}$ for each covariate. Although the supremum test statistic continues to strongly reject the null hypothesis at conventional significance levels, we fail to reject the null hypothesis of a linear fit or a cubic fit with the L_2 norm when $\widehat{\mathbf{w}} = \mathbf{w}_{max}$ (cf. Figure 3). We can circumvent these ostensible inconsistencies by using the test statistic of equation (3.4) based on the first derivative. This allows us to avoid choosing where to evaluate $\mathbf{w}$. Here, we strongly reject the null hypotheses

[3] We could also consider hypothesis tests where the null hypothesis is specified on restricted portions of the support of x_i. Our results apply directly to that case as well.

of linear or cubic regression models and can conclude that neither models are appropriate for these data.

We can also test shape restrictions on the bivariate relation between uninsurance rates and per capita income. A natural hypothesis is a null hypothesis of "decreasingness" corresponding to uninsurance rates monotonically declining as per capita income rises. This corresponds to a test of $\sup_{x\in\mathcal{X}} \Upsilon_{\mathbf{w}}^{(1)}(x) \leq 0$. The third panel of Figure 1 presents the test statistics and associated p-values for this null hypothesis in the model with and without covariates. We emphatically reject the null hypothesis for both specifications which is likely driven by the presence of medicaid coverage for lower income zip codes. In the bottom panel of Table 1 we present a test of convexity of the conditional mean function—a test of $\inf_{x\in\mathcal{X}} \Upsilon_{\mathbf{w}}^{(2)}(x) \geq 0$—which necessitates estimation of the second derivative. We also strongly reject this shape restriction for both specifications.

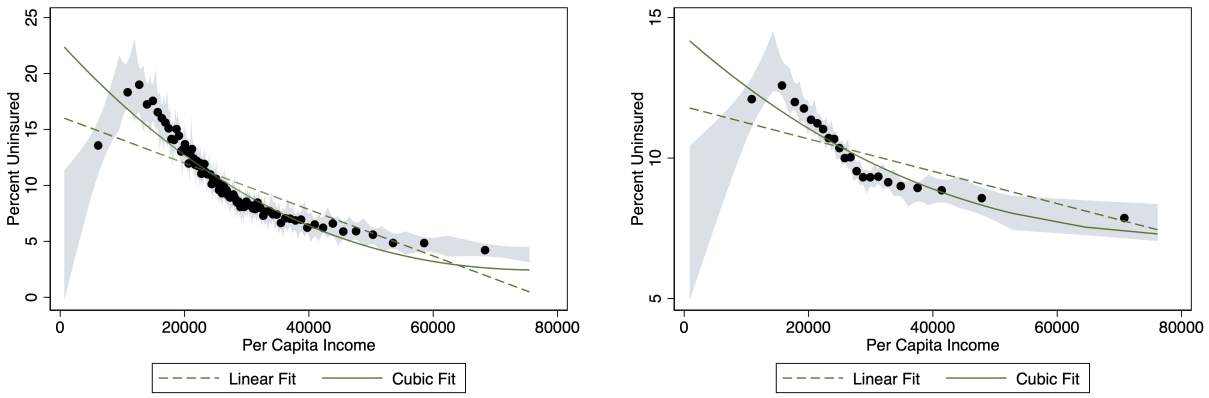
Table 1: Formal Tests for Parametric Specifications and Shape Restrictions.

	Sup norm		**L_2 norm**		
	Test Statistic	P-value	Test Statistic	P-value	$\hat{J}_{\texttt{IMSE}}$
Test of Linear Fit					
No Covariates	59.90	0.000	7.48	0.000	81
Covariates, $\widehat{\mathbf{w}} = \mathbf{w}_{min}$	5.14	0.000	3.80	0.000	22
Covariates, $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$	8.37	0.000	4.42	0.000	22
Covariates, $\widehat{\mathbf{w}} = \mathbf{w}_{max}$	5.53	0.000	0.93	0.357	22
Covariates, First Derivative	10.53	0.000	2.83	0.000	13
Test of Cubic Fit					
No Covariates	25.96	0.000	3.60	0.000	81
Covariates, $\widehat{\mathbf{w}} = \mathbf{w}_{min}$	5.50	0.000	1.98	0.042	22
Covariates, $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$	6.28	0.000	2.78	0.005	22
Covariates, $\widehat{\mathbf{w}} = \mathbf{w}_{max}$	3.41	0.005	0.66	0.533	22
Covariates, First Derivative	13.41	0.000	2.90	0.000	13
Test of Monotonic Decline					
No Covariates	8.75	0.000			16
Covariates	4.64	0.000			13
Test of Convexity					
No Covariates	-12.75	0.000			7
Covariates	-6.25	0.000			6

Notes. This table reports the test statistics and associated p-values from hypothesis tests of parametric specifications and shape restrictions using the Census data as in Figures 1 and 2. The top panel reports test results under the null of a linear regression model whereas the second panel reports test results under the null of a cubic regression model. The bottom two panels report test results under the null of a monotonic decline (third panel) or convexity (fourth panel) of the conditional expectation. p-values based on 50,000 simulations.

Figure 4: **Graphical Representation of Parametric Specification Testing.** This figure compares the binned scatter plot and associated confidence bands to parametric fits of the conditional expectation function. The dependent variable, the independent variable of interest, and covariates are the same as in Figures 1 and 2. In the specification with covariates, $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$. Shaded regions denote confidence bands for a nominal level of 95%.

(a) *Without Covariates* (b) *With Covariates*



Finally, we demonstrate the multisample testing results. We divide states into two groups based on their population density as measured by the average population per square mile.[4] Specifically, we label low and high density states as those with population densities below or above 100, respectively. Figure 5 plots the binscatter fits for two groups and their corresponding confidence bands. The left chart omits control variables whereas the right plot includes these covariates with a conditional mean estimate evaluated at their sample mean. We observe that the point estimates correspond to higher uninsured rates in zip codes in low population density states as compared to high density states. Without controls, there is generally overlap in the confidence bands throughout the range of per capita income shown. In contrast, when covariates are added, there is a much clearer delineation between the two groups at all but the lowest of income levels.

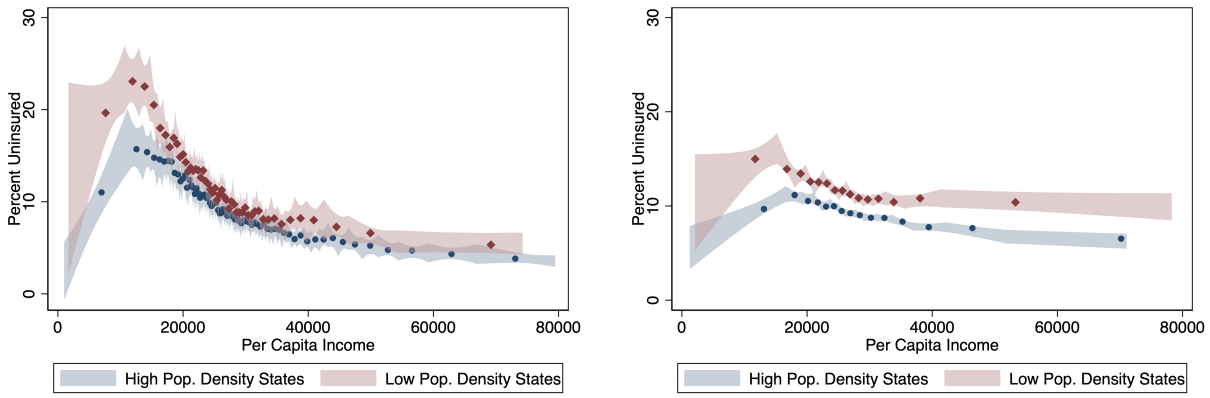
4 Generalized Nonlinear Binscatter

The discussion so far has focused on least squares covariate-adjusted binscatter and estimation and inference for (features of) the conditional mean function. This section studies a more general class of binscatter methods, including examples such as nonlinear least squares regression, quantile

[4]Data available from the Census Bureau at https://www.census.gov/data/tables/time-series/dec/density-data-text.html.

Figure 5: **Two-Sample Comparison.** This figure utilizes the same data as in Figures 1 and 2. Low density states are defined as states with average population per square mile of less than a 100 whereas high density states are those with above 100. Each chart shows the binned scatter plot for each group along with associated confidence bands. The right chart includes covariates as described in the caption to Figure 2 with $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$. Shaded regions denote confidence bands for a nominal level of 95%.

(a) *Without Covariates* (b) *With Covariates*



regression, and MLE methods including generalized linear models such as logistic and Poisson regression. We will keep the discussion brief here, focusing on what is new relative to Sections 2 and 3.2, while relegating most of the detail to the SA.

There are two main messages in this section. First, we are able to obtain the same theoretical results for this more general case as we obtained for least squares. The technical work is more involved for this class of estimators and we give a number of novel results, discussed below and in Section SA-3, that may be of independent interest. Second, thanks to these theoretical results established, we are able to deliver the same visual and analytical tools. This means we can extend binscatter methods to different data types, such as discrete outcomes where classical scatter plots are useless for visualization (e.g., Figure 6), and to other features of the data, such as conditional quantiles for assessing spread (e.g., Figure 1(f)).

Instead of the model (2.1), we now assume that the underlying parameters $\mu_0(\cdot)$ and $\boldsymbol{\gamma}_0$ are defined by

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \underset{\mu \in \mathcal{M}, \boldsymbol{\gamma} \in \mathbb{R}^d}{\arg\min} \; \mathbb{E}[\rho(y_i; \eta(\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}))], \tag{4.1}$$

where the loss function $\eta \mapsto \rho(\cdot; \eta)$ is assumed to be absolutely continuous with a piecewise Lipschitz continuous weak derivative $\psi(y; \eta) = \psi(y - \eta)$ exhibiting at most a finite number of discontinuity

points, the (inverse) link function $\eta(\cdot)$ is assumed to be strictly monotonic and thrice continuously differentiable, and $\mathcal{M}$ is an appropriate space of functions satisfying certain (smoothness) conditions made precise in the supplemental appendix. We assume $u \mapsto \rho(\cdot;\eta(u))$ is convex to simplify our results, but this restriction can be dropped with additional technical work. The least squares setting studied in previous sections corresponds to the choice $\rho(y;\eta) = (y-\eta)^2$ and $\eta(u) = u$.

The parameter of interest in this context that is most natural, and most closely corresponds to (2.2) is, for a user-chosen evaluation point $\mathbf{w}$,

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

For example, if $v = 0$, then $\vartheta_{\mathbf{w}}(x) = \eta(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)$ captures the relationship between x and y evaluated at level $\mathbf{w}$. If $v = 1$, then $\vartheta_{\mathbf{w}}^{(1)}(x) = \eta^{(1)}(\mu_0(x) + \mathbf{w}'\boldsymbol{\gamma}_0)\mu_0^{(1)}(x)$ is the marginal partial effect of x on y at level $\mathbf{w}$.

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

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

For estimation, we define the *generalized nonlinear* binscatter estimator as

$$\widehat{\mu}^{(v)}(x) = \widehat{\mathbf{b}}^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \qquad \begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma}\big)\Big). \tag{4.3}$$

Under regularity conditions, (4.1) is the probability limit of (4.3), and therefore a natural plug-in

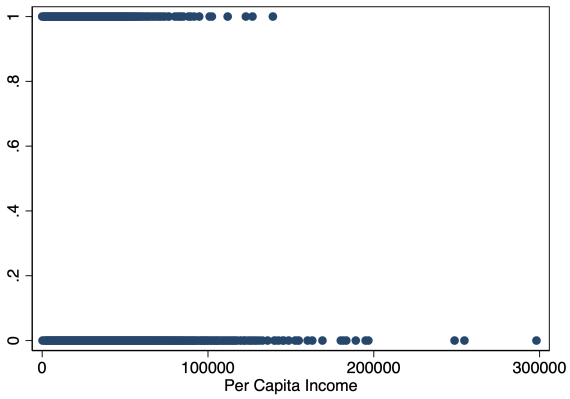
binscatter estimator for $\vartheta_{\mathbf{w}}^{(v)}(x)$ is

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

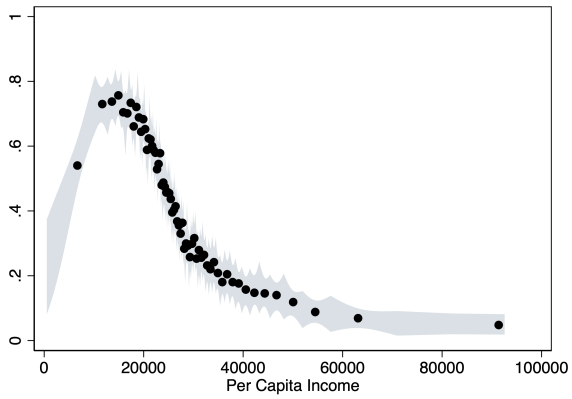
where $\widehat{\mathbf{w}}$ is a consistent estimator of the desired evaluation point $\mathbf{w}$ in (4.2).

Figure 6: **Nonlinear Binned Scatter Plots.** This figure utilizes the same data as in Figure 1. The dependent variable is defined as a binary variable which takes on the value of one when a zip code has an uninsured rate above 10% and zero otherwise. In the specification with covariates, $\widehat{\mathbf{w}} = \bar{\mathbf{w}}$. Shaded regions denote confidence bands for a nominal level of 95%.

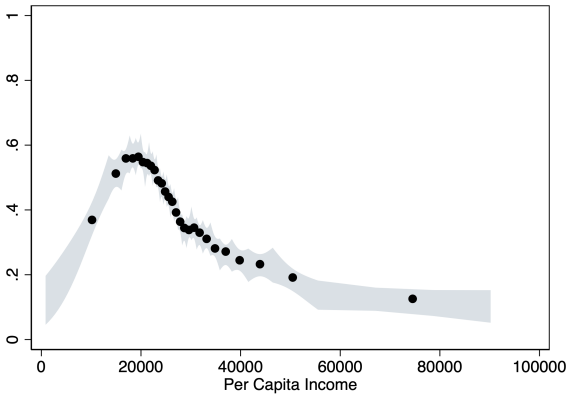
(a) *Classical Scatter Plot*



(b) *Binned Scatter Plot*



(c) *Binned Scatter Plot w/Covariates*



4.1 Technical Results

Here we give a summary of our technical results for generalized nonlinear binscatter estimators, highlighting a few central issues. The Appendix and SA provide more details.

The technical work, and the results, are more substantially more involved than those for least

squares regression, but nonetheless we are able to give analogous theorems. In a nutshell, we are able to extend all the results presented in previous sections for $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ in (2.5) to now accommodate $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$ in (4.4). A central difficulty that we overcome is that, in general, a closed-form expression is not available in the present case. This means that in some cases stronger conditions are required in the general case compared to least squares where the closed-form expression is leveraged substantially. The supplemental appendix (Sections SA-3 and SA-4) provide all the details.

The crux of our theoretical work is a novel uniform Bahadur representation (Theorem SA-3.1) for the generalized nonlinear binscatter estimator $\widehat{\mu}^{(v)}(x)$ in (4.3). To give a summary of that result, under standard but cumbersome assumptions (Assumptions SA-DGP and SA-GL, taking ν defined there to be four to match Assumption 1), we show that if $J^{\frac{8}{3}}\log^{\frac{4}{3}}(n)/n \to 0$ and $\log(n)J^{-1} \to 0$, then

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

uniformly in $x \in \mathcal{X}$, where $\bar{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}(x_i)\widehat{\mathbf{b}}(x_i)'\Psi_{i,1}\eta_{i,1}^2$ with $\Psi_{i,1} = \frac{\partial}{\partial\eta}\mathbb{E}[\psi(y_i;\eta)|x_i,\mathbf{w}_i]\big|_{\eta=\eta_{i,0}}$, and $\eta_{i,v} = \eta^{(v)}(\mu_0(x_i)+\mathbf{w}_i'\boldsymbol{\gamma}_0)$. This result substantially generalizes the least squares version, under essentially the same rate restrictions previously imposed, with an error of approximation that is optimal up to $\log(n)$ terms. Our results are on par with, or improve upon, prior theory for kernel estimators (Kong, Linton, and Xia, 2010) and series estimation for quantile regression (Belloni, Chernozhukov, Chetverikov, and Fernandez-Val, 2019); see Remark SA-3.3. As before, (4.5) motivates the asymptotic variance, but as discussed in the supplemental appendix, estimation of the asymptotic variance is more complicated, and our results there provide general high-level conditions justifying several alternatives commonly used in practice.

Taking the uniform Bahadur representation as the starting point, we establish the following results (along with other technicalities):

1. Theorem SA-3.6 establishes a Nagar-type approximate IMSE expansion for $\widehat{\mu}^{(v)}(x)$ in (4.3), and shows that the resulting optimal choice of J is equivalent to $J_{\texttt{IMSE}}$ in (2.6), but with different bias and variance constants $\mathscr{B}_n(p,v)$ and $\mathscr{V}_n(p,v)$, which are substantially more complicated (as expected). We can also select the polynomial order analogously to $p_{\texttt{IMSE}}$ of (2.7). To avoid dealing with the complicated constants, $J_{\texttt{IMSE}}$ of (2.6) can always be used as a valid rule-of-thumb IMSE-optimal estimator of the number of bins for generalized nonlinear

binscatter, which is valid by Theorem SA-3.6. This theorem is new to the literature, even in the case of non-random partitioning and without covariate adjustment, for both general nonlinear series estimators and binscatter (piecewise polynomials and splines) nonlinear series estimators in particular.

2. Theorems SA-3.4 and SA-3.5 establish strong approximations and valid uniform inference for $\widehat{\mu}^{(v)}(x)$ in (4.3). Similar results were only available for special cases and under stronger conditions. In particular, see Remark SA-3.5 for a comparison to Belloni, Chernozhukov, Chetverikov, and Fernandez-Val (2019), who study series-based quantile regression. Section SA-4.4 then establishes strong approximation results for $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$. Those results provide all the necessary tools to establish hypothesis testing and confidence bands, matching Sections 3.2, 3.3, and 3.4. Formal statements are give in Sections SA-4.1, SA-4.2, and SA-4.3.

4.2 Empirical Illustration

We illustrate the generalized nonlinear binscatter using a nonparametric logit specification. Again using our data on the uninsured rate in each zip code, we define our dependent variable as an indicator function of whether this rate is above 10 percent. The top left chart in Figure 6 shows a classical scatter plot for this binary variable. Clearly, there is only limited information we can gather from such a plot. It is clear that there are no observations with high uninsured rates and high incomes; however, at lower income levels, nothing can be gleaned. In contrast, the top right chart shows the nonlinear binned scatter plot along with an associated 95% confidence band. We can observe a clear hump shape in the conditional expectation which persists even with additional covariates (bottom plot).

5 Applications

In this section we further demonstrate some of the capabilities of our new toolkit by revisiting two recent papers which utilize binned scatter plots: Akcigit, Grigsby, Nicholas, and Stantcheva (2022) and Moretti (2021). We defer to those papers for a complete description of their data and analysis. Our goal here is to show that the empirical results in each of these papers are enhanced and strengthened when re-analyzed using the methodology introduced in the earlier sections.

Akcigit, Grigsby, Nicholas, and Stantcheva (2022) study the effect of corporate and personal taxes on innovation in the United States over the twentieth century. In the benchmark estimation the authors use both linear regressions and binned scatter plots to study the relation between log patents and marginal tax rates utilizing a rich set of control variables including fixed effects (See Table II and Figure I of Akcigit, Grigsby, Nicholas, and Stantcheva (2022)). In their macro-level approach, the authors show that higher taxes negatively affect the quantity of innovation.

To start, the top left plot of Figure 7 presents a raw scatter plot of log patents and the variable of interest, transformed marginal tax rates.[5] Despite a sample size of about 3,000 observations it is difficult to draw any inferences about the data from the scatter plot. In the top right plot of Figure 7 we replicate Figure I of Akcigit, Grigsby, Nicholas, and Stantcheva (2022) using $J = 50$.[6] This is the *same* binscatter as Figure 2(c) but on a tighter (incorrect) scale.

It is intuitive to view and interpret this figure as one would a conventional scatter plot: as a cloud of points with a regression line fit to the "data" and we would conclude that there may be a positive but noisy relationship between these two variables. This interpretation is tempting, and indeed the very name "binscatter" invites this, but it is incorrect: the dots here are not data points but estimates of the conditional mean function. This is emphasized in Figure 7(c), which is *formally identical* to the figure in the original paper (Figure 7(b)), but visually very different, and now assuming the wiggly step function is well-approximated by a line seems clearly inappropriate. However, there are two issues here: the incorrect residualization is done and the number of bins is too large, leading to massive undersmoothing. Figure 7(d) addresses the former, applying our corrected approach to covariates overlaying the residualized version now at the correct scale, making the difference even starker. Correctly adjusting for other covariates presents a very different picture of the empirical conclusions to be drawn from the data than do Figures 7(b) and (c): now the estimated conditional expectation is approximately linear with a slope above 2. In this sense, the correct approach actually harmonizes the precisely estimated regression coefficient reported in Akcigit, Grigsby, Nicholas, and Stantcheva (2022) and the binned scatter plot, strengthening the

[5] The authors use the logarithm of one minus the marginal tax rate so this transformed variable implies that a positive relation between y and x implies that higher marginal tax rates are associated with lower quantity of innovation.

[6] When presenting the replication of this figure we have added back the mean of y and x (as is done in the `binscatter` command). Akcigit, Grigsby, Nicholas, and Stantcheva (2022) do not make this addition; however, we emphasize this is for visual simplicity only and has no effect on our conclusions.

original conclusions. This visual pattern is even more apparent in the bottom left plot where we address undersmoothing by using the IMSE-optimal J. With fewer bins ($J_{\texttt{IMSE}} = 12$) the point estimate of the conditional expectation function appears even closer to linearity.

Akcigit, Grigsby, Nicholas, and Stantcheva (2022) use a number of control variables in their regressions, including year and state fixed effects. As we discussed in Section 3.2 we can conduct inference on the derivative of the conditional expectation to draw conclusions which are invariant to how categorical variables are defined. The bottom right plot shows the estimated derivative of the conditional expectation function along with an associated 95% confidence band. We clearly reject the null of no relationship (i.e., the conditional expectation is equal to a constant) as the confidence band does not envelop the horizontal line at $y = 0$. However, we fail to reject the null of a linear relationship as the confidence band clearly accommodates horizontal lines at *positive* values on the y-axis consistent with a positive linear relationship. Taken in concert, our visual and formal results strengthen the evidence in favor of the original baseline regressions.

We now turn to Moretti (2021), examining the relation between the productivity of top inventors and high-tech clusters, where clusters are defined as activity in a city of a specific research field (e.g., computer scientists in Silicon Valley). The paper estimates an elasticity of number of patents in a year with respect to cluster size of 0.0676. The statistically significant positive relationship aligns with the observation that increasingly large subsidies are being offered by states and localities for high-tech firms to relocate within their regions.

We again begin our analysis with a raw scatter plot of the data (top left of Figure 8). With close to one million observations, the scatter plot is both dense and uninformative. In the top right plot we replicate Figure 4 in Moretti (2021) which is a binned scatter plot controlling for year, research field and city effects, and in the middle left we show the implied estimate of the mean function. In the middle right plot we present the results of the correct and incorrect residualization on the same chart. When placed on the same scale, we again see that the support of the binned scatter plot with incorrect residualization is substantially curtailed. As in the previous example, the correct residualization pains a clearer picture of the joint relationship between y and x. The estimate of the conditional expectation appears approximately flat for smaller clusters before rising steadily as the cluster size grows. This pattern is largely replicated when the optimal choice of J is used in the middle right plot.

The only control variables used in Moretti (2021) are fixed effects of different kinds. The main specification employs 11 different fixed effects and so we again focus on the derivative of the conditional expectation function rather than the function itself. The bottom left plot presents the estimated derivative of the conditional expectation function along with an associated 95% confidence band. We clearly reject the null of no relationship between the variables as the confidence band is outside the horizontal line at zero for large clusters. However, we also reject the null of linearity as there is no horizontal line that can be enveloped by the whole confidence band. Instead we find strong evidence of a nonlinear relation for larger-sized clusters and little evidence of an effect for smaller clusters. This added nuance to the results of Moretti (2021) obtained through our new tools is not inconsequential. Taken at face value, it would imply that states and localities which have only small clusters of inventors might have to offer very large incentives in order to grow their cluster size sufficiently large to generate the positive agglomeration effects presented in Moretti (2021).

6 Conclusion

Data visualization is a powerful tool for effectively conveying empirical results in a simple and intuitive form. Binned scatter plots have become a popular tool to present a flexible, yet cleanly interpretable, estimate of the relationship between an outcome and covariate of interest. However, despite their visual simplicity and conceptual appeal, there has been no work to establish that they provide a high quality, or even accurate, visualization of the data. This hampers their reliability and usability in applications.

We introduce a suite of formal and visual tools based on binned scatter plots to improve, and in some cases correct, empirical practice. Our methods offer novel visualization tools, principled covariate adjustment, estimation of conditional mean, quantile, and other nonlinear functions, visualization of variance and precise uncertainty quantification, and formal tests of substantive hypotheses such as linearity or monotonicity. We illustrate our methods with three substantive empirical applications, two of them revisiting recently published papers (Akcigit, Grigsby, Nicholas, and Stantcheva, 2022; Moretti, 2021) in economics, and show in particular the pitfalls of employing binned scatter methods incorrectly in practice. Further, our empirical reanalysis showcase how

applying binned scatter correctly can strengthen the empirical findings in those papers. All of our results are fully implemented in publicly available software (Cattaneo, Crump, Farrell, and Feng, 2022).

Appendix A Summary of Technical Contributions

The online supplement is a comprehensive collection and discussion of all our new theoretical results for generalized nonlinear partitioning-based estimators with semi-linear covariate-adjustment and random partitioning based on empirical quantiles. Canonical binscatter and all other binscatter methods discussed in the main paper are special cases of the generic setup considered therein. Many of our results contribute to the broader literature on series estimation, and are thus of independent interest outside of binscatter contexts. Here we give a brief summary of the new theoretical results, pointing to specific places in the supplemental appendix. Further, at the end of each technical subsection of the supplement, we include a remark labelled "Improvements over literature" that discusses more details of the technical improvements presented in that subsection and gives related references.

Section SA-1.2 presents new technical lemmas for random partitions based on empirical quantiles. Those results include general characterizations of the "regularity" of the random partitioning scheme (Lemmas SA-1.1 and SA-1.2) and of the associated random basis functions (Lemmas SA-1.3 and SA-1.4). These results give sharp control on the underlying random binning scheme of binscatter methods.

Section SA-2 studies large sample point estimation and distributional properties of the least squares estimator of $\mu_0(x)$ in (2.1). We study this case separately because its closed form solution allows for sharper results under weaker regularity conditions, and because least squares binscatter is arguably the most popular approach in empirical work. New results include:

1. technical lemmas for Gram matrix (Lemma SA-2.1), asymptotic variance (Lemmas SA-2.2 and SA-2.3), approximation error (Lemma SA-2.4) and covariate adjustments (Lemma SA-2.5);

2. stochastic linearization and uniform convergence rate (Theorem SA-2.1 and Corollary SA-2.2) and variance estimation (Theorem SA-2.2);

3. pointwise distributional approximation (Theorem SA-2.3);

4. conditional strong approximation (Theorem SA-2.4) and feasible implementation thereof (Theorem SA-2.5);

5. integrated mean squared error (IMSE) expansions (Theorem SA-2.6) and IMSE-optimal tuning parameter selection.

All these results explicitly account for the random binning scheme. The most noteworthy novel result in this section is the conditional strong approximation, which circumvents a fundamental lack of uniformity of the random binning basis $\widehat{\mathbf{b}}_{p,s}^{(v)}(x)$, while still delivering a sufficiently fast uniform coupling requiring only $J^2/n \to 0$ (up to $\log(n)$ terms). In fact, if a subexponential moment restriction holds for the error term, it suffices that $J/n \to 0$ (up to $\log(n)$ terms). Such rate conditions not only improve on previous results in the literature, but also allow for canonical binscatter (i.e., our results show that there exists a sequence $J \to \infty$ such that bias and variance are simultaneously controlled even when $p = s = 0$).

Section SA-3 studies large sample point estimation and distributional properties of the generalized nonlinear estimator of $\mu_0(x)$ in (4.1). New results include:

1. technical lemmas for Gram matrix (Lemma SA-3.1), asymptotic variance (Lemmas SA-3.2 and SA-3.3), approximation error (Lemma SA-3.4) and uniform consistency (Lemma SA-3.5);

2. stochastic linearization and uniform convergence rate (Theorem SA-3.1 and Corollary SA-3.2) and variance estimation (Theorem SA-3.2);

3. pointwise distributional approximation (Theorem SA-3.3);

4. conditional strong approximation (Theorem SA-3.4) and feasible implementation thereof (Theorem SA-3.5);

5. IMSE expansions (Theorem SA-3.6) and IMSE-optimal tuning parameter selection.

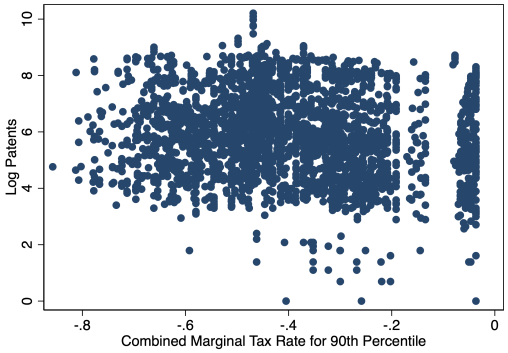
All these results explicitly account for the random binning scheme. This section includes two most noteworthy novel results. First, a sharp Bahadur representation for general nonlinear semiparametric partitioning-based estimators with much faster rate of convergence than previously available in

the literature is established: the result requires only $J^{\frac{8}{3}}/n \to 0$ (up to $\log(n)$ terms), while previous results required $J^4/n \to 0$ (up to $\log(n)$ terms) or worse. In fact, if a subexponential moment condition holds for the error term and a piecewise polynomial ($s = 0$) is employed, we only need the minimal assumption $J/n \to 0$ (up to $\log(n)$ terms). Therefore, our results allow for canonical binscatter and generalizations thereof, which would have been excluded by prior results (i.e., for previous technical results there was no sequence $J \to \infty$ such that bias and variance are simultaneously controlled). Furthermore, our new Bahadur representation allows us to employ the same novel conditional strong approximation approach mentioned above, albeit with some important technical differences, to establish uniform inference results for generalized nonlinear binscatter methods under essentially the same tuning parameter conditions imposed for least square binscatter methods. Second, a new Nagar-type IMSE expansion for generalized nonlinear partitioning-based estimators with semi-linear covariate-adjustment and random partitioning based on empirical quantiles is established, which has no antecedent in the literature to the best of our knowledge. Lastly, our methods also allow for a large class of loss functions (e.g., L_p or Huber regression) and for semi-linear covariate adjustment in nonlinear series estimation settings, leading to new results that were previously unavailable in the literature.

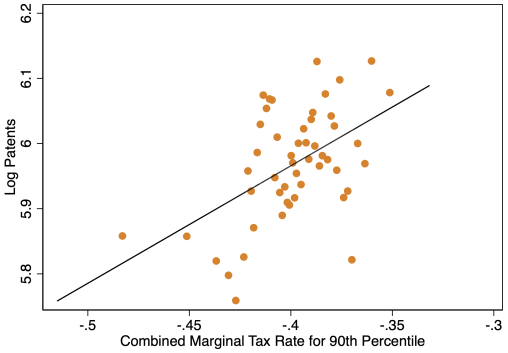
Section SA-4 employs the technical results in Sections SA-2 and SA-3 to study estimation and inference for $\widehat{\Upsilon}^{(v)}_{\widehat{\mathbf{w}}}(x)$ and $\widehat{\vartheta}^{(v)}_{\widehat{\mathbf{w}}}(x)$, respectively. New results include valid confidence band estimators, consistent hypothesis tests about parametric specification and shape restrictions, and a detailed discussion on other parameters of interest, among other results. All these results explicitly account for the random binning scheme and semi-linear covariate-adjustment with random evaluation point. The most noteworthy novel result in this section is the proof technique to transform our strong approximation results (Theorems SA-2.4 and SA-3.4), and their feasible versions (Theorems SA-2.5 and SA-3.5), into statements about the Kolmogorov distance for the suprema and related functionals of the t-statistic processes of interest. Our technical approach again circumvents a fundamental lack of uniformity of the random binning basis $\widehat{\mathbf{b}}^{(v)}_{p,s}(x)$, while still delivering a sufficiently fast uniform coupling, requiring only $J^2/n \to 0$ (up to $\log(n)$ terms) in the least squares case and $J^{\frac{8}{3}}/n \to 0$ (up to $\log(n)$ terms) in the general nonlinear case. This proof technique can also be used to analyze other functionals such as the L_p distance, Kullback–Leibler divergence, and arg max statistic.

Figure 7: **Effect of taxes on innovation.** This figure uses the data from Akcigit, Grigsby, Nicholas, and Stantcheva (2022). The top left plot shows a raw scatter plot of the log number of patents per state per year versus the marginal tax rate for the 90th percentile earners. The top right plot shows Figure I(A) in Akcigit, Grigsby, Nicholas, and Stantcheva (2022) whereas the middle left plot shows the implied estimated conditional mean function. The incorrect residualization versus the semi-linear specification introduced in Section 2 (both for 50 bins) is shown in the middle right chart. The bottom left chart uses the optimal choice of J introduced in Section 2.3. The bottom right plot shows the estimated derivative of the conditional mean function along with the associated 95% confidence bands using a cluster-robust variance estimator two-way clustered by state × five-year period and year based on the specification in Akcigit, Grigsby, Nicholas, and Stantcheva (2022, Table II, Column (1))

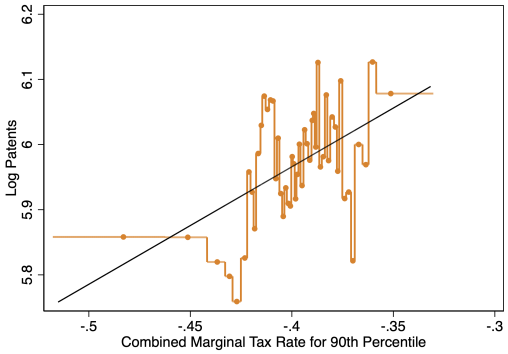
(a) *Raw Scatter Plot*



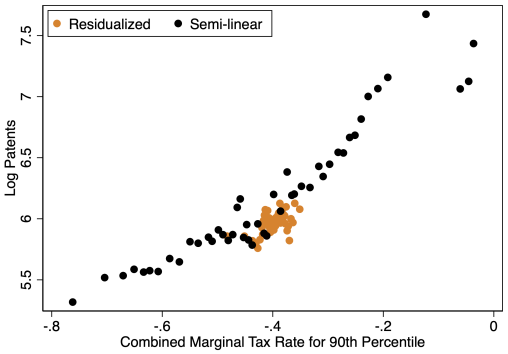
(b) *Fig. I of Akcigit et al. (2022)*



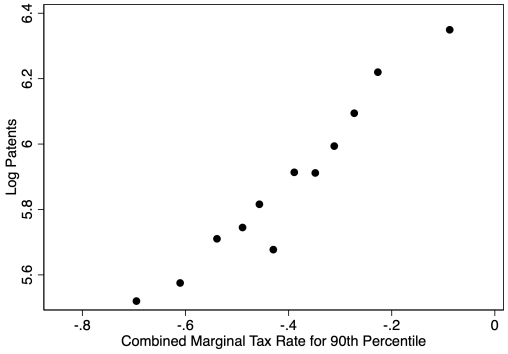
(c) *Incorrect Residualization*



(d) *Covariate Adjustment*



(e) *Optimal J*



(f) *Deriv. of Cond. Exp.*

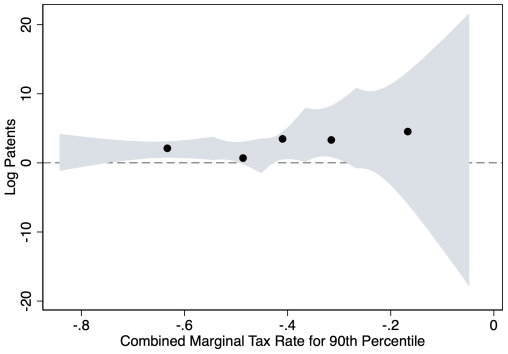
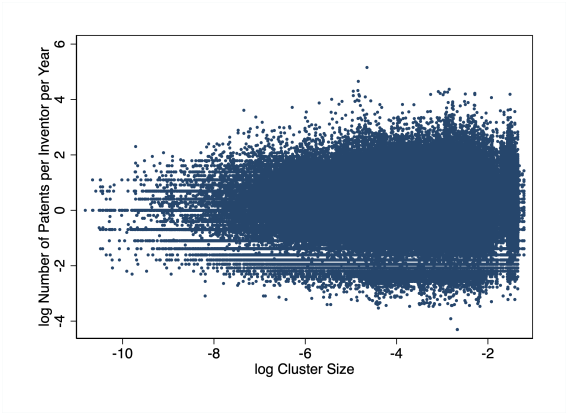
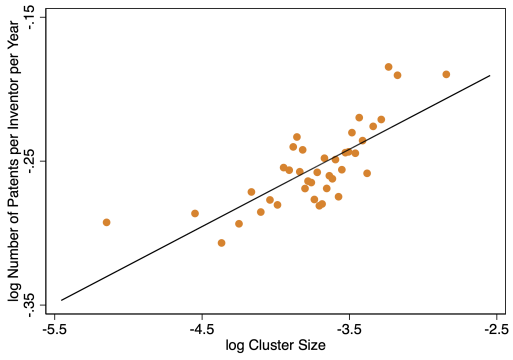


Figure 8: **Relation between productivity of top inventors and high-tech clusters.** This figure uses the data from Moretti (2021). The top left plot shows a raw scatter plot of the log number of patents per inventor per year versus the log cluster size. The top right plot shows Figure 4 in Moretti (2021) whereas the middle left plot shows the implied estimated conditional mean function. The incorrect residualization versus the semi-linear specification introduced in Section 2 (both for 40 bins) is shown in the middle right chart. The bottom left chart uses the optimal choice of J introduced in Section 2.3. The bottom right plot shows the estimated derivative of the conditional mean function along with the associated 95% confidence bands using a cluster-robust variance estimator clustered by city $\times$ field based on the specification in Moretti (2021, Table 3, Columnn (8)).

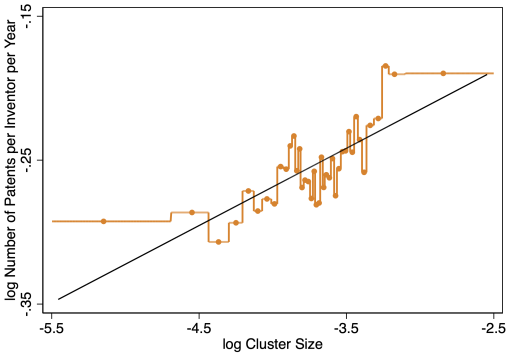
(a) *Raw Scatter plot*



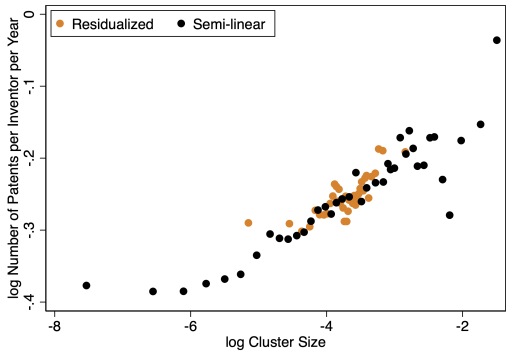
(b) *Fig. 4 of Moretti (2021)*



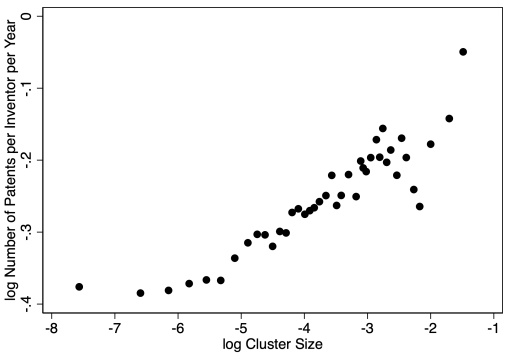
(c) *Incorrect Residualization*



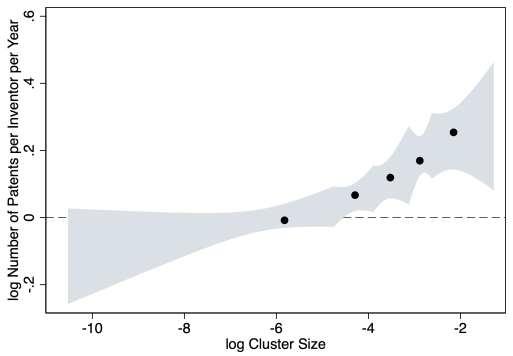
(d) *Covariate Adjustment*



(e) *Optimal J*



(f) *Deriv. of Cond. Exp.*



References

Abadie, A. (2020): "Statistical nonsignificance in empirical economics," *American Economic Review: Insights*, 2(2), 193–208.

Akcigit, U., J. Grigsby, T. Nicholas, and S. Stantcheva (2022): "Taxation and Innovation in the Twentieth Century," *Quarterly Journal of Economics*, 137(1), 329–385.

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

Bali, T. G., R. F. Engle, and S. Murray (2016): *Empirical Asset Pricing: The Cross Section of Stock Returns*. John Wiley & Sons.

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

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

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.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2022): "Binscatter Regressions," in preparation for the *Stata Journal*.

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

Cattaneo, M. D., and R. Titiunik (2022): "Regression Discontinuity Designs," *Annual Review of Economics*, 14, 821–851.

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.

Feigenberg, B., and C. Miller (2021): "Racial Divisions and Criminal Justice: Evidence from Southern State Courts," *American Economic Journal: Economic Policy*, 13(2), 207–240.

Freyaldenhoven, S., C. Hansen, J. Pérez Pérez, and J. M. Shapiro (2021): "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design," in *Advances in Economics and Econometrics - Twelfth World Congress*. forthcoming.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019): "Pre-event trends in the panel event-study design," *American Economic Review*, 109(9), 3307–38.

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

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

Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and Its Application to the Additive Model," *Econometric Theory*, 26(5), 1529–1564.

Korting, C., C. Lieberman, J. Matsudaira, Z. Pei, and Y. Shen (2021): "Visual Inference and Graphical Representation in Regression Discontinuity Designs," *arXiv preprint arXiv:2112.03096*.

Moretti, E. (2021): "The Effect of High-Tech Clusters on the Productivity of Top Inventors," *American Economic Review*, 111(10), 3328–3375.

Schlenker, W., and M. J. Roberts (2009): "Nonlinear temperature effects indicate severe damages to US crop yields under climate change," *Proceedings of the National Academy of sciences*, 106(37), 15594–15598.

Shapiro, A. H., and D. J. Wilson (2021): "Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis," *The Review of Economic Studies*, 89(5), 2768–2805.

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

Stuart, E. A. (2010): "Matching methods for causal inference: A review and a look forward," *Statistical Science*, 25(1), 1–21.

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.

Wang, Q., Z. Chen, Y. Wang, and H. Qu (2021): "A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization," *IEEE Transactions on Visualization and Computer Graphics*.