Nonlinear Binscatter Methods*

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

July 19, 2024

Abstract

Binned scatter plots are a powerful statistical tool for empirical work in the social, behavioral, and biomedical sciences. Available methods rely on a quantile-based partitioning estimator of the conditional mean regression function to primarily construct flexible yet interpretable visualization methods, but they can also used to estimate treatment effects, assess uncertainty, and test substantive domain-specific hypotheses. This paper introduces novel binscatter methods based on nonlinear, possibly nonsmooth M-estimation methods, covering generalized linear, robust, and quantile regression models. We provide a host of theoretical results and practical tools for local constant estimation along with piecewise polynomial and spline approximations, including (i) optimal tuning parameter (number of bins) selection, (ii) confidence bands, and (iii) formal statistical tests regarding functional form or shape restrictions. Our main results rely on novel strong approximations for general partitioning-based estimators covering random, data-driven partitions, which may be of independent interest. We demonstrate our methods with an empirical application studying the relation of the percentage of individuals without health insurance to per capita income at the zip-code level. We provide general-purpose software packages implementing our methods in `Python`, `R`, and `Stata`.

Keywords: partition-based semi-linear estimators, generalized linear models, quantile regression, robust bias correction, uniform inference, binning selection, treatment effect estimation.

*We thank Isaiah Andrews, Ricardo Masini, Boris Shigida, and Rocio Titiunik for helpful comments and discussions. Ignacio Lopez Gaffney provided excellent research assistance. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805, SES-2019432, and SES-2241575. Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122, 72133002, and 72250064. 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. Software and replication files are available at https://nppackages.github.io/binsreg/.

†Department of Operations Research and Financial Engineering, Princeton University.
‡Macrofinance Studies, Federal Reserve Bank of New York.
§Department of Economics, UC Santa Barbara.
¶School of Economics and Management, Tsinghua University.

1 Introduction

Data visualization is a crucial step in any statistical analysis. In the era of big data it has become increasingly important to have simple yet informative visual tools to guide, supplement, or in some cases even supplant, numerical statistical analyses. However, it is important to maintain statistical formality and rigor to ensure the validity of any conclusions based on the data. We seek to accomplish both of these goals—effective visualization couched in a formal framework—with binned scatter plot methods.

Often known simply as a *binscatter*, the binned scatter plot has become a popular tool for visualization in large data sets, particularly in the social and behavioral sciences. The goal is to flexibly estimate, and visualize, features of the conditional distribution of a scalar outcome y_i, which may be discrete or continuous, given a covariate or treatment variable x_i, which is scalar and continuous, while possibly also controlling for a d-dimensional vector of additional factors $\mathbf{w}_i$. For estimating the conditional mean, as in the traditional regression analysis, binning has a long history: so familiar is this approach that over 60 years ago Tukey (1961), calling it a *regressogram*, went so far as to claim that "[a]ll statisticians who handle data know how to attack the simple case of this situation where y and x are both single real numbers" (p. 682). Going on to describe the construction, Tukey writes: "[t]he x-axis is to be divided into suitable intervals, the mean of all the y-values corresponding to x-values falling in each given interval is to be found, the results are then to be plotted, either as points, each located above the center of the corresponding x-interval, or better as horizontal bars, each extending over the corresponding x-interval" (p. 682).

This simple construction (perhaps disappointingly often with plotted dots rather than horizontal lines) has recently gained popularity in statistics, economics, and data science. The prevalence of binscatter plots can be partly ascribed to its intuitive construction and compelling visualization properties: given only data on x_i and y_i, the plot is a clean and interpretable depiction of the conditional mean. Moreover, several limitations of the classical scatter plot account for the rising use of binned scatter plots in modern analyses. First, in big

data sets the classical scatter plot is too dense to be informative, particularly about general "patterns" in the data which are to be modeled in subsequent analyses. Second, somewhat conversely, in cases where privacy is a concern the scatter plot is not allowed, regardless of its informational use as a visualization tool. Third, classical scatter plots do not provide a well-defined way to control for other factors, a common goal in treatment effect estimation and causal inference. Finally, particularly relevant to our setting, a scatter plot is not useful when outcomes are discrete. In contrast, a *binned* scatter plot provides a simple, yet flexible way of visualizing features of the conditional distribution of a (possibly discrete) outcome variable given a continuous covariate (or treatment) of interest, while controlling for other important factors.

Formally, a binscatter is grounded in the classical semilinear regression model. To date, however, binscatters have been available only to visualize (and estimate) conditional mean functions fitted using least squares. A common usage in this setting is comparing the nonparametric estimate to a linear fit, as a precursor to linear regression analysis. See Starr and Goldfarb (2020) for a practical review and background references. In the least squares setting, a binscatter is formally an estimator of a semilinear model for the conditional mean, nonparametric in the covariate of interest and linear in the controls, where the nonparametric component is estimated by partitioned regression. Cattaneo et al. (2024b) used that framework to derive formal statistical properties of canonical binscatter, including correcting a common mistake in empirical practice when using controls, and provide asymptotically valid confidence bands and optimal tuning parameter selection.

The restriction to least squares semilinear regression to estimate the conditional mean has limited the applicability of binscatter methods. For one, important features of the data, such as spread or variability, cannot be visualized. Further, existing methods (and theory) can be misleading in settings where the outcome is discrete or in another way restricted. For example, in the empirical illustration we use throughout, we study uninsuredness rates using a fractional outcome model, most naturally fitted using quasi-likelihood methods based on

the logistic link. Last but not least, binscatter methods for quantile regression analysis are currently lacking in the literature, despite of their usefulness for empirical work.

This paper introduces and studies a broad class of binscatter M-estimation methods, in models allowing for (i) a nonlinear and/or nonsmooth loss function and (ii) a nonlinear link function. Our results provide for the use of binned scatter plots for various visualization goals and different data types, particular leading cases being semiparametric conditional quantile regression and generalized partially linear models. We make several methodological and theoretical contributions: (i) we propose a feasible method for optimal tuning parameter selection to choose the appropriate number of bins; (ii) we provide (pointwise and) uniform inference to construct confidence bands and hypothesis tests for parametric specifications and shape restrictions, and (iii) we develop group-wise comparisons for continuous treatment effects or for treatment effect heterogeneity. Developing these methods relies on novel technical work: allowing for a large class of binning methods, including random binning, we prove new uniform (in x) Bahadur representations and strong approximations, and thus uniform distribution theory, for the broad class of nonlinear semiparametric models considered. Obtaining these results for nonlinear, nonsmooth models, with data-dependent partitions and additional covariates, represents the main technical contributions of our paper, some of which may be of independent interest.

Our proposed nonlinear binscatter methods help restore, and in cases such as discrete outcomes or additional controls, surpass, the utility of the conventional scatter plot. We offer principled ways to visually assess patterns in the data, quantify uncertainty, and develop hypothesis tests about the findings. Our results on quantile regression allow researchers to assess the spread in the conditional distribution, detect outliers or influential observations in the data, and study a larger class of treatment effects, formalizing and expanding common practices based on the classical scatter plot in small data sets, all while controlling for additional important factors. Our confidence bands properly quantify and communicate the uncertainty around the estimated function of interest, and can also be used to guide

further analyses. We also develop formal uniform hypothesis testing procedures regarding those functions, to assess shape constraints and parametric specifications. Causal inference is an important application area of our uniform inference results: studying treatment effect heterogeneity for binary treatments or the dose response function for a continuous treatment without imposing a functional form (e.g., to evaluate important hypotheses such as monotonicity in the dosage).

The paper proceeds as follows. We next discuss the connections between our work and the existing literature. Section 2 introduces binned scatter plots, defines the statistical model, and clarifies the parameters of interest. Section 3 gives details on our theoretical contributions, which are then used in Sections 4 and 5 to deliver tuning parameter selection and uniform inference. We illustrate our methods and results with a running empirical application using zip code-level data from the American Community Survey (ACS). The dependent variable, y_i, is the percentage of individuals without health insurance and the independent variable of interest, x_i, is per capita income. Section 6 concludes. The online Supplemental Appendix (SA hereafter) contains additional technical and implementation details, all mathematical proofs, and further discussion of how our technical contributions improve on the related literature. General-purpose software in `Python`, `R`, and `Stata`, as well as replication files, are available at `https://nppackages.github.io/binsreg/`. See Cattaneo et al. (2024a) for an introduction.

1.1 Related Literature

This paper contributes to several strands of the literature. First, from a practical point of view, our work builds upon and extends existing binned scatter plot methods available for applied research. See Starr and Goldfarb (2020) for a review of that literature and Cattaneo et al. (2024b) for formal results concerning least squares semilinear binscatter. Our main methodological contribution is to introduce nonlinear binscatter methods, constructed using a general, possibly nonsmooth semilinear M-estimation approach. As a result, we

propose a broad array of new binscatter methods for generalized linear models (e.g., Logit or Probit), robust semiparametric regression (e.g., Huber or trimmed least squares), and quantile regression.

Second, our theoretical results contribute to the literature on series/sieve estimation in general, and partitioning-based methods in particular (i.e., piecewise polynomials and splines approximations). See Györfi et al. (2002) for a textbook introduction, and Shen et al. (1998), Huang (2003), Belloni et al. (2015), Cattaneo and Farrell (2013), Cattaneo et al. (2020), as well as references therein, for prior convergence rates and distribution theory. These prior works studied uniform estimation and inference for linear piecewise polynomials and spline series regression without data-driven partitioning and without additional covariates, often imposing strong regularity conditions. Our primary technical contributions as compared to that recent literature are (i) allowing for general, possibly nonlinear and nonsmooth M-estimation, (ii) allowing for random partitions and hence random basis functions in the series estimator, (iii) controlling for other factors in a semilinear model, and (iv) obtaining novel strong approximations and uniform inference under weaker conditions than those previously available. A substrand of the series estimation literature studies quantile regression, the closest antecedent to our work being Belloni et al. (2019). Unlike that prior work, we consider general nonlinear, possibly nonsmooth, M-estimation problems and allow for random partitions and additional controls, and our technical results are obtained under weaker regularity conditions, which in particular permit the use of piecewise constant fitting necessary for a binned scatter plot. Finally, none of the results in Cattaneo et al. (2024b) are applicable to the large class of nonlinear binscatter estimators considered in this paper, because they only consider least squares semilinear binscatter models. Further details of how each of our individual theoretical results improves on the extant literature is given throughout the SA.

Finally, our paper also contributes to the literature on data visualization, which has become an increasingly active field of study in recent years due to the rise of big data and machine learning methods. Our results speak directly to this literature, and in particular to the need

for clear and explicit depictions of uncertainty, both in terms of variance and estimation error (Healy, 2018). These are crucial in data visualization in science and research contexts as this "builds trust and credibility" (Schwabish, 2021, p.189).

2 Setup

A binned scatter plot is designed to provide a flexible, nonparametric estimate of a regression-type function. The construction and interpretation of a binned scatter plot is simple and intuitive, which drives their appeal for applied work. But as we will see, there are some subtleties when binned scatter plots are applied to nonlinear, nonsmooth models—especially when controlling for additional covariates.

To describe the construction, it is helpful to first make precise the model and objects of interest. Our goal is to learn a regression-type function (which need not be the conditional mean) that features in the conditional distribution of a scalar outcome y_i, which may be discrete or continuous, given a covariate or treatment variable x_i, which is scalar and continuous, while possibly also controlling for a d-dimensional vector of additional factors $\mathbf{w}_i$. In applications, the goal is to flexibly study the relationship of y_i to x_i, but not necessarily to discover (or allow for) heterogeneity or nonlinearity in $\mathbf{w}_i$. Further, $\mathbf{w}_i$ is often a large-dimensional set of controls, such as fixed effects or factor variables. Consequently, we assume the regression function depends on the scalar index $\theta_0(x_i, \mathbf{w}_i) := \mu_0(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_0$, for an unknown function μ_0 and vector $\boldsymbol{\gamma}_0$, and is thus partially linear in nature. This specification is directly interpretable, and in cases where d is moderate or large, empirically convenient.

The model is defined by the following structure, which determines how the scalar index $\theta_0(x_i, \mathbf{w}_i)$ relates to the outcome y_i. Let θ be a generic value of the index. For a loss function $\rho(y; \eta(\theta))$ and inverse link function $\eta(\theta)$, let

$$(\mu_0(\cdot), \boldsymbol{\gamma}_0) = \underset{\mu \in \mathcal{M}, \boldsymbol{\gamma} \in \mathbb{R}^d}{\arg\min}\ \mathbb{E}\big[\rho(y_i; \eta(\mu(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}))\big], \tag{2.1}$$

where we assume the solution is unique, and $\rho(y;\eta(\theta))$, $\eta(\theta)$, and the function class $\mathcal{M}$ obey typical boundedness and smoothness restrictions discussed in Section 3. For different choices of $\rho(\cdot)$ and $\eta(\cdot)$ this formulation covers a large class of problems including generalized linear models, robust regression, quantile regression, and nonlinear least squares. We illustrate with some leading specific examples.

Example 1 (Least Squares Regression)**.** Setting $\eta(\theta) = \theta$ and $\rho(y;\eta) = (y-\eta)^2$ recovers semiparametric least squares regression for partially linear models. ⌟

Example 2 (Logistic Regression)**.** Assume that the binary outcome y_i, conditional on x_i and $\mathbf{w}_i$, is distributed Bernoulli with probability by $\eta(\mu(x_i)+\mathbf{w}_i'\boldsymbol{\gamma})$, where $\eta(\theta) = (1+\exp(-\theta))^{-1}$, then $\rho(y;\eta) = -y\log(\eta) - (1-y)\log(1-\eta)$. ⌟

Example 3 (Huber Regression)**.** Semiparametric robust partially linear regression sets $\eta(\theta) = \theta$ and $\rho(y;\eta) = (y-\eta)^2\mathbb{1}(|y-\eta| \leq \tau) + \tau(2|y-\eta| - \tau)\mathbb{1}(|y-\eta| > \tau)$ for some $\tau > 0$. ⌟

Example 4 (Quantile Regression)**.** Set $\rho(y;\eta) = [\tau - \mathbb{1}(y < \eta)](y-\eta)$ with $\eta(\theta) = \theta$ for some $\tau \in (0,1)$. ⌟

The key statistical challenge is (uniform in x) recovery of the function $\mu_0(x)$ for estimation and inference. Once accomplished, we can cover a wide variety of objects derived from (2.1). For concreteness we will focus on the following three objects, as they are of primary practical importance:

(i) the level of the regression function, $\vartheta_0(x,\mathsf{w}) = \eta(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0) = \eta(\theta_0(x,\mathsf{w}))$,

(ii) the nonparametric component itself (or its derivative), $\mu_0^{(v)}(x) = \frac{d^v}{dx^v}\mu_0(x)$, $v \geq 0$, and

(iii) the marginal effect $\zeta_0(x,\mathsf{w}) = \frac{\partial}{\partial x}\eta(\mu_0(x) + \mathsf{w}'\boldsymbol{\gamma}_0) = \eta^{(1)}(\theta_0(x,\mathsf{w}))\mu_0^{(1)}(x)$,

where $h^{(1)}(u) = \frac{d}{du}h(u)$ denotes the derivative of a function with respect to its scalar argument and w is a user-chosen evaluation point for the additional controls. Typically, w

is chosen as the mean or median, or for discrete variables or fixed effects, set to a baseline category. The role of w in both plotting and inference introduces some important nuances that are discussed below.

Each of these parameters corresponds to different empirical questions. The level, $\vartheta_0(x, \mathsf{w})$, is directly useful for visualization of the relationship of y_i to x_i and is commonly used in causal inference. If the variable x_i is a continuous treatment, our results yield a nonparametric estimate of the dose response function, while controlling for relevant factors $\mathbf{w}_i$. A plot of $\vartheta_0(x, \mathsf{w})$ shows this function for the subgroup defined by $\mathbf{w}_i = \mathsf{w}$. We can also obtain separate dose response functions for different subgroups of the data to be used in multi-sample comparisons. On the other hand, if x_i is a pre-treatment variable, the same multi-sample results provide an analysis of treatment effect heterogeneity for discrete (often binary) treatments, and our uniform inference allows for discovery of treatment effect heterogeneity. Finally, for visualization, obtaining $\vartheta_0(x, \mathsf{w})$ in the quantile regression case can be used to assess the spread of the conditional distribution (especially for quantiles close to zero and one) or robust measures of central tendancy, as would be done with a classical scatter plot.

The nonparametric component, $\mu_0(x)$, is most often studied to assess its functional form, generally against a parsimonious parametric specification to be considered for later analyses or for a shape restriction that is of substantive interest, such as monotonicity or convexity. Historically, a common use of binscatter was to visually (and informally) assess if $\mu_0(x)$ is well-approximated by a linear model, and if so, proceeding under that specification for the empirical results. Our results provide rigor to such practice, and expand the idea to a much richer class of models and hypotheses.

The marginal (or partial) effect $\zeta_0(x, \mathsf{w})$ is a standard object in economic analysis in nonlinear models. In binary choice models it is common to study how the probability of $y = 1$ changes as a function of x. The marginal effect at the average, obtained by setting w to the sample mean, is a standard way to summarize nonlinear models by giving the effect for the "average" individual. For example, we show that the marginal effect of income changes

sign in our application, indicating a changing response in the uninsuredness rate as a result of Medicaid. Comparing marginal effects across groups for heterogeneity analysis, in causal or noncausal settings, is a common goal in social science applications with nonlinear models.

2.1 Estimation

Given an i.i.d. sample $(y_i, x_i, \mathbf{w}_i')'$, $i = 1, \ldots, n$, the binscatter estimator is constructed by solving the empirical analogue of (2.1) using a partitioning-based approximation to the unknown function $\mu_0(x)$. This nonparametric approximation requires two choices: the partitioning of the support of x_i and the estimation within each bin.

To fix ideas, it is useful to begin with the simplest case where local constant fitting is used and $\mathbf{w}_i$ is absent. First, the support of x_i is divided into $J < n$ disjoint bins. J is the main tuning parameter for this nonparametric estimation problem, and its choice is crucial both visually and statistically. To describe the estimator, we will take $J < n$ as given at present, and return to its choice in Section 4 below.

Coupled with a choice of J is a method to divide the support. A major theoretical innovation of our work is that the bin breakpoints themselves can be data-dependent, distinct from a data-driven choice of J. The partition is denoted by $\widehat{\Delta} = \{\widehat{\mathcal{B}}_1, \widehat{\mathcal{B}}_2, \ldots, \widehat{\mathcal{B}}_J\}$, with the bins and breakpoints denoted by

$$\widehat{\mathcal{B}}_j = \begin{cases} \left[\widehat{\tau}_{j-1}, \widehat{\tau}_j\right) & \text{if } j = 1, 2, \ldots, J-1, \\ \left[\widehat{\tau}_{J-1}, \widehat{\tau}_J\right] & \text{if } j = J. \end{cases}$$

The breakpoints $\{\widehat{\tau}_j\}_{j=1}^J$ result from a user-chosen, possibly data-driven, partitioning method. Our theoretical results cover any random partition that is independent of the outcomes y_i's (given x_i's and $\mathbf{w}_i$'s), and "quasi-uniform", which intuitively requires the bins to be sufficiently similar. The formal condition is stated in the next section. The simplest approach is evenly-spaced breakpoint locations ($\widehat{\tau}_j = x_{\texttt{min}} + j(x_{\texttt{max}} - x_{\texttt{min}})/J$). The most popular

choice, however, is to use the empirical quantiles of x_i, setting $\widehat{\tau}_j = \widehat{F}^{-1}((j-1)/J)$ with $\widehat{F}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq u)$ the standard empirical cumulative distribution function, and $\widehat{F}^{-1}$ its generalized inverse. In the SA we show that both of these satisfy our generic assumptions. Other possible methods include certain adaptive regression trees and related partitioning methods, such as those with the so-called "X-property" or via sample splitting; see Devroye et al. (2013), Zhang and Singer (2010), and references therein. For concreteness, we will use quantile spacing for empirical analysis throughout the paper.

Given a partition $\widehat{\Delta}$, the binscatter estimate is formed by fitting the sample analogue of (2.1) within each bin, using only an intercept. In the simple case of least squares regression, this is identical to computing the sample average of y_i for observations in each bin, exactly as Tukey (1961) described, yielding a piecewise constant approximation to the unknown conditional expectation. The same method is followed for all other models. For example, in the case of binary data or fractional outcomes, a logistic regression of y_i on a constant is fit for each bin. For the median, or any other quantile, one simply computes the empirical quantile of y_i using observations only within the bin.

Formally, we define the basis functions $\widehat{\mathbf{b}}_0(x) = [\mathbb{1}_{\widehat{\mathcal{B}}_1}(x), \mathbb{1}_{\widehat{\mathcal{B}}_2}(x), \cdots, \mathbb{1}_{\widehat{\mathcal{B}}_J}(x)]'$, consisting of indicators for each bin. We then obtain

$$\widehat{\mu}(x) = \widehat{\mathbf{b}}_0(x)'\widehat{\boldsymbol{\beta}}, \qquad \widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_0(x_i)'\boldsymbol{\beta}\big)\Big). \tag{2.2}$$

A graphical illustration of this procedure is shown in Figure 1. The data are obtained from the 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, y_i, is the percentage of individuals without health insurance, and the independent variable of interest, x_i, is per capita income. The fractional nature of the outcome motivates the use of logistic quasi-maximum likelihood for estimation and inference (Papke and Wooldridge,

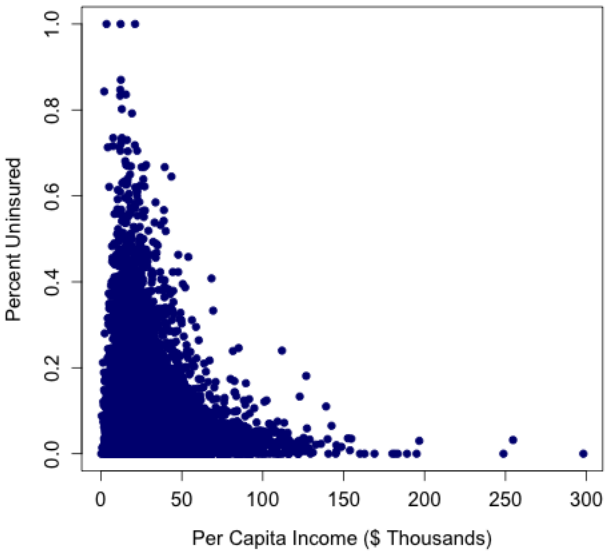
1996). Figure 1(a) shows the classical scatter plot of the raw data. This data set has about 32,000 observations, far from the millions commonly encountered, and already this plot fails to be useful for assessing the functional form: the visualization is dominated by a dense cloud of data with a few outlying observations. Figure 1(b) shows a binned scatter plot being constructed, with the raw data in the background. The dots are the fitted values of applying (2.2) following Example 2, i.e., we show $\widehat{\vartheta}(x) = \eta(\widehat{\mu}(x))$. Figure 1(c) isolates the binscatter and overlays a linear fit (i.e., a global logistic quasi-likelihood with $\mu_0(x)$ assumed linear in x). The linear approximation to $\mu_0(x)$ appears satisfactory at first, but this is because the nonparametric estimate is undersmoothed. Figure 1(d) presents the estimate using the optimal number of bins (Section 4), and shows that the informal analysis, relying on an ad hoc choice of J, would miss an important feature of the data: the presence of the Medicaid program which provides subsidized health insurance for limited-income individuals. As a preview, Table 1 below shows that formal tests reject polynomial parametric specifications and reject the hypothesis that the uninsurance rate is monotonically decreasing with per capita income.

Graphs like Figures 1(c) and (d) have a long tradition in statistics and data science, and have recently become ubiquitous in applied microeconomics. Visually assessing functional forms is the typical use. Importantly, in this case the visualization shows an estimate of $\vartheta_0(x, \mathsf{w})$, not $\mu_0(x)$ directly. Further, although the binned scatter plot invites the viewer to "connect the dots" smoothly, the actual estimator is piecewise constant, which generally gives a less appealing visualization but underpins any formal analysis.

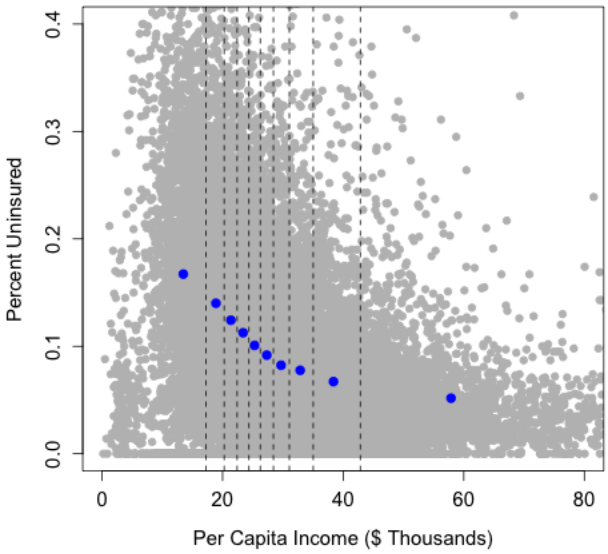
We expand on (2.2) in two ways: adding the covariates $\mathbf{w}_i$ and enriching the set of allowable basis functions. The covariates can be directly incorporated into the loss, exactly as they are in (2.1). Moreover, the additively separable and linear nature of the controls makes this generalization straightforward empirically. Importantly, the presence of controls invalidates bin-by-bin estimation, as the coefficients $\boldsymbol{\gamma}_0$ are global parameters. The SA discusses different approaches to estimating μ_0 and $\boldsymbol{\gamma}_0$.

Figure 1: **Illustration of Nonlinear Binned Scatter Plots.** This figure illustrates the construction of a nonlinear binned scatter plot using the ACS data. 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.

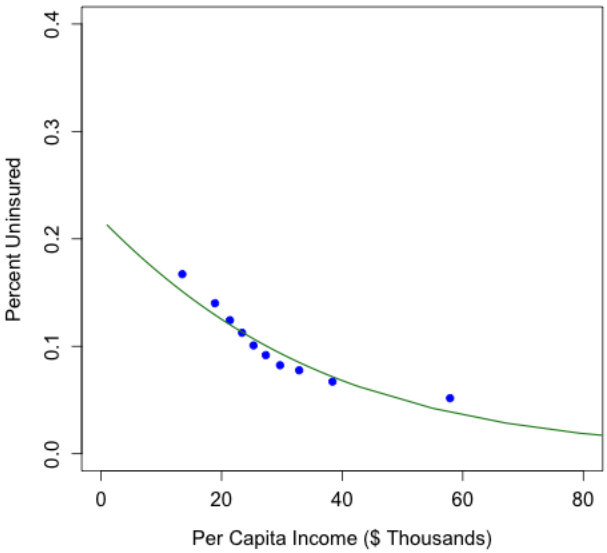
(a) *Raw Scatter Plot*



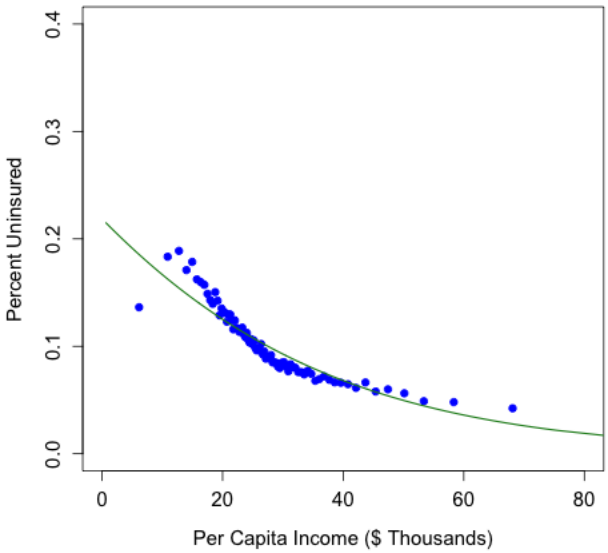
(b) *Bin Cutoffs and Binscatter*



(c) *Ad-Hoc Binscatter Plot*



(d) *IMSE-Optimal Binscatter*



Next, we replace the piecewise constant approximation based on $\widehat{\mathbf{b}}_0(x)$ with an order-p polynomial approximation in each bin that is $(s-1)$-times continuously differentiable at the breakpoints of the bins, with the convention that $s = 0$ corresponds to discontinuous fits and

$s = 1$ is continuous but nondifferentiable. This change to the basis gives additional flexibility that is crucial for bias reduction and derivative estimation, the latter being instrumental for studying shape restrictions and specification testing. By construction, $p \geq s \geq 0$, and derivative estimation requires that the derivative of interest, $v \geq 0$, is no larger than p. The general basis is defined as $\widehat{\mathbf{b}}_{p,s}(x) := \widehat{\mathbf{T}}_s[\widehat{\mathbf{b}}_0(x)' \otimes (1, x, \ldots, x^p)']$, where $\otimes$ denotes the Kronecker product and $\widehat{\mathbf{T}}_s$ is a $[(p+1)J-(J-1)s] \times (p+1)J$ transformation matrix ensuring that the $(s-1)$-th derivative of the estimate is continuous. The exact form of $\widehat{\mathbf{T}}_s$ is available in Section SA-5.2 of Cattaneo et al. (2024b), and we note that $\widehat{\mathbf{T}}_s$ also depends on the random partitions. When $s = 0$, $\widehat{\mathbf{T}}_s$ is the identity matrix, and the fit is a piecewise polynomial of degree p. The piecewise constant fit (as bars or as dots) corresponds to $s = p = 0$. Another popular choice are cubic B-splines, obtained by setting $s = p = 3$. On account of its popularity and to simplify notation, we will assume throughout the paper that $s = p$ and use the notation $\widehat{\mathbf{b}}_p(x) := \widehat{\mathbf{b}}_{p,p}(x)$. The SA treats the general case of $s \leq p$.

The generalized, covariate-adjusted binscatter can now be defined. We first solve

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{bmatrix} = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \sum_{i=1}^{n} \rho\Big(y_i;\ \eta\big(\widehat{\mathbf{b}}_p(x_i)'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma}\big)\Big). \tag{2.3}$$

Using (2.3) the estimators of the three functions of interest are:

$$\widehat{\vartheta}_p(x, \widehat{\mathsf{w}}) = \eta(\widehat{\theta}_p(x, \widehat{\mathsf{w}})), \tag{2.4}$$

$$\widehat{\mu}_p^{(v)}(x) = \widehat{\mathbf{b}}_p^{(v)}(x)'\widehat{\boldsymbol{\beta}}, \tag{2.5}$$

$$\widehat{\zeta}_p(x, \widehat{\mathsf{w}}) = \eta^{(1)}(\widehat{\theta}_p(x, \widehat{\mathsf{w}}))\widehat{\mu}_p^{(1)}(x), \tag{2.6}$$

respectively, where $\widehat{\theta}_p(x, \widehat{\mathsf{w}}) = \widehat{\mu}_p(x) + \widehat{\mathsf{w}}'\widehat{\boldsymbol{\gamma}}$, is the plug-in estimator of the true index $\theta_0(x, \mathsf{w})$, and $\widehat{\mathsf{w}}$ (non-random or generated based on $\{\mathbf{w}_i\}_{i=1}^n$) is a consistent estimator for the desired evaluation point w. We will often make the polynomial order p explicit, as this is needed for clarity when constructing confidence bands and hypothesis tests in Section 5; dependence

on $\widehat{\Delta}$ and choice of J is suppressed, but also important.

It is worth mentioning that nonlinear binscatter methods can be constructed for both fixed $J < \infty$ and large $J \to \infty$ as $n \to \infty$, naturally leading to different interpretations. The functions of interest defined at the beginning of this section cannot be recovered when J is fixed, but coarsened versions thereof will be, and these objects can have an interesting interpretation: if the parition "settles" as $n \to \infty$ to some fixed Δ_0 with associated fixed basis $\mathbf{b}_p(x)$ (see Assumption 4 below for a precise definition), then the probability limit of the fixed-J binscatter is the solution to (2.1) where the function class $\mathcal{M}$ is restricted to be $\mathcal{M} = \{\mu(x) = \mathbf{b}_p(x)'\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^K, K = \dim(\mathbf{b}_p(x))\}$. This is most natural with $p = 0$ and quantile-spacing, because the binscatter shows average outcomes across quantiles of a continuous covariate. For $p = 0$ and $J = 100$, for example, the results allow for comparison of y_i across percentiles of x_i (possibly controlling for $\mathbf{w}_i$), which is standard for x_i variables such as test scores or measures of wealth. All our estimation and inference results remain valid when J is fixed, provided the target parameter is adjusted accordingly; see Cattaneo et al. (2024a,b) for further discussion of fixed-J binscatter methods. For the remaining of the paper, we will consider only the case $J \to \infty$ as $n \to \infty$ to streamline the presentation.

3 Theory

This section presents two main novel technical results: a uniform Bahadur representation and a feasible strong approximation. The methodological results in subsequent sections—tuning parameter selection, confidence bands, and hypothesis testing—are built from these results. To conserve space and notation, in this section we only show the results for $\widehat{\vartheta}_p(x, \widehat{\mathsf{w}})$. The analogous results for $\widehat{\mu}_p^{(v)}(x)$ and $\widehat{\zeta}_p(x, \widehat{\mathsf{w}})$ are deferred to the SA (specific references below) and are conceptually similar. The SA also gives other important technical results and additional discussion of how our theory improves on the existing literature.

First we state the assumptions required. The class of data generating processes, encom-

passing the model and loss, are restricted by the following.

Assumption 1 (Data Generating Process)**.**

(i) $\{(y_i, x_i, \mathbf{w}_i') : 1 \leq i \leq n\}$ *are i.i.d. random vectors satisfying* (2.1) *and supported on* $\mathcal{Y} \times \mathcal{X} \times \mathcal{W}$*, where* $\mathcal{X}$ *is a compact interval and* $\mathcal{W}$ *is a compact set.*

(ii) The marginal distribution of x_i *has a Lipschitz continuous (Lebesgue) density bounded away from zero on* $\mathcal{X}$*.*

(iii) The conditional distribution of y_i *given* $(x_i, \mathbf{w}_i')$ *has a (conditional) density with respect to some sigma-finite measure uniformly bounded over its support and* $\mathcal{X} \times \mathcal{W}$*.*

This assumption is fairly standard. It restricts attention to cross-sectional data and bounded covariates with minimal regularity imposed on the underlying joint distribution. Requiring x_i to be continuously distributed is natural given the visualization and estimation goals, but our results can also be applied to discrete x_i by taking each mass point as its own bin to conduct simultaneous estimation and inference over those support points.

The following assumption restricts the class of statistical models.

Assumption 2 (Statistical Model)**.**

(i) $\rho(y;\eta)$ *is absolutely continuous with respect to* $\eta \in \mathbb{R}$ *and admits a derivative* $\psi(y,\eta) := \psi^\dagger(y-\eta)\psi^\ddagger(\eta)$ *almost everywhere.* $\psi^\ddagger(\cdot)$ *is continuously differentiable and strictly positive or negative. If the conditional distribution of* y_i *given* $(x_i, \mathbf{w}_i')$ *does not have a Lebesgue density, then* $\psi^\dagger(\cdot)$ *is Lipschitz continuous, otherwise it is piecewise Lipschitz with finitely many discontinuity points.*

(ii) $\rho(y;\eta(\theta))$ *is convex with respect to* θ *and* $\eta(\cdot)$ *is strictly monotonic and three-times continuously differentiable.*

(iii) $\mathbb{E}[\psi(y_i, \eta(\theta_0(x_i, \mathbf{w}_i)))|x_i, \mathbf{w}_i] = 0$*.* $\sigma^2(x, \mathbf{w}) := \mathbb{E}[\psi(y_i, \eta(\theta_0(x_i, \mathbf{w}_i)))^2|x_i = x, \mathbf{w}_i = \mathbf{w}]$ *is bounded away from zero uniformly over* $\mathcal{X} \times \mathcal{W}$*.* $\mathbb{E}[\eta^{(1)}(\theta_0(x_i, \mathbf{w}_i)^2\sigma^2(x_i, \mathbf{w}_i)|x_i = x]$ *is*

Lipschitz continuous on $\mathcal{X}$*, and* $\mathbb{E}[|\psi(y_i, \eta(\theta_0(x_i, \mathbf{w}_i)))|^\nu | x_i = x, \mathbf{w}_i = \mathbf{w}]$ *is uniformly bounded over* $\mathcal{X} \times \mathcal{W}$ *for some* $\nu > 2$*.* $\mathbb{E}[\psi(y_i, \eta)|x_i = x, \mathbf{w}_i = \mathbf{w}]$ *is twice continuously differentiable with respect to* η*.*

(iv) For $\Upsilon(x, \mathbf{w}) := \frac{\partial}{\partial \eta}\mathbb{E}[\psi(y_i, \eta)|x_i = x, \mathbf{w}_i = \mathbf{w}]|_{\eta=\eta(\theta_0(x,\mathbf{w}))}$*,* $\Upsilon(x, \mathbf{w})\eta^{(1)}(\theta_0(x, \mathbf{w}))^2$ *is bounded away from zero uniformly over* $\mathcal{X} \times \mathcal{W}$ *and* $\mathbb{E}[\Upsilon(x_i, \mathbf{w}_i)\eta^{(1)}(\theta_0(x_i, \mathbf{w}_i))^2 | x_i = x]$ *is Lipschitz continuous on* $\mathcal{X}$*.*

(v) $\mu_0(\cdot)$ *is* ς*-times continuously differentiable for some* $\varsigma \geq p + 1$*.*

This assumption imposes regularity on the statistical model in (2.1), particularly on the loss function and resulting parameters of interest. The complexity of part (i) reflects the breadth of the class of models and parameters we cover. When y_i is continuous the loss function can have points of nondifferentiability, but for discrete outcomes the loss must be smoother. To illustrate, consider first Example 4 in Section 2: the loss function for quantile regression is continuous but not differentiable everywhere, which is covered by our assumptions with $\psi(y, \eta) = \text{sign}(y-\eta)\eta^{(1)}$, where $\psi^\dagger(y-\eta) = \text{sign}(y-\eta) = 1-2\mathbb{1}(y-\eta \leq 0)$ exhibits a mass point, and $\psi^\ddagger(\eta) = \eta^{(1)}$ is smooth. Alternatively, for logistic regression (Example 2) $y_i \in \{0, 1\}$ and we have $\psi(y, \eta) = (y - \eta)[\eta(1 - \eta)]^{-1}$, which exactly matches the required structure of $\psi^\dagger(y - \eta)\psi^\ddagger(\eta)$. Both functions are clearly as smooth as required and the definition of $\eta(\theta)$ ensures that $\psi^\ddagger > 0$. If nonlinear least squares is used instead, the conditions are trivially satisfied. The rest of the assumption gives standard moment and boundedness conditions to ensure that the parameters and their estimators are well-defined; those regularity conditions are also satisfied in all examples of interest. Finally, the nonparametric object $\mu_0(\cdot)$ is assumed to be smooth, as is standard in the nonparametric inference literature.

We next give several high-level conditions on the estimation procedure. These conditions ensure that the partitioning scheme is sufficiently regular and that the evaluation point for the control variables $\mathbf{w}_i$ and the Gram matrix can be estimated sufficiently well.

Assumption 3 (High-Level Estimation Conditions)**.**

(i) *The partition* $\widehat{\Delta}$ *is independent of* $\{y_i\}_{i=1}^n$ *given* $\{x_i, \mathbf{w}_i\}_{i=1}^n$ *and, w.p.a.* 1, $\max_{1\leq j\leq J} |\widehat{\tau}_j - \widehat{\tau}_{j-1}| \leq C \min_{1\leq j\leq J} |\widehat{\tau}_j - \widehat{\tau}_{j-1}|$*, for an absolute constant* $C > 0$*.*

(ii) $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(1)$ *and* $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$*, where* $\|\cdot\|$ *is the Euclidean norm.*

(iii) *For the infeasible Gram matrix* $\bar{\mathbf{Q}}_p := n^{-1}\sum_{i=1}^n \widehat{\mathbf{b}}_p(x_i)\widehat{\mathbf{b}}_p(x_i)'\Upsilon(x_i, \mathbf{w}_i)\eta^{(1)}(\theta_0(x_i, \mathbf{w}_i))^2$*, there is an estimator* $\widehat{\Upsilon}(x_i, \mathbf{w}_i)$ *such that* $\|\bar{\mathbf{Q}}_p - \widehat{\mathbf{Q}}_p\| = O_{\mathbb{P}}\big(J^{-p-1} + \big(\frac{J\log n}{n^{1-2/\nu}}\big)^{1/2}\big)$*, where* $\widehat{\mathbf{Q}}_p := n^{-1}\sum_{i=1}^n \widehat{\mathbf{b}}_p(x_i)\widehat{\mathbf{b}}_p(x_i)'\widehat{\Upsilon}(x_i, \mathbf{w}_i)\eta^{(1)}(\widehat{\theta}_p(x_i, \mathbf{w}_i))^2$*, and* $\|\cdot\|$ *is the operator norm.*

The requirement that the partition intervals are not too dissimilar in length is satisfied for evenly spaced partitioning, trivially, and is shown to hold for quantile spacing in the SA (Lemma SA-5.2). For other data-driven methods this condition must be checked. This assumed property of the random bining structure is often called quasi-uniformity (Cattaneo et al., 2020; Huang, 2003), and is important for controlling the approximation bias and, when combined with the assumptions on the density of x_i, for ensuring that each bin contains sufficient data to control the variance. Part (ii) requires that the desired evaluation point of $\mathbf{w}_i$ (such as the mean) can be estimated consistently and that the coefficient vector $\boldsymbol{\gamma}_0$ can be estimated sufficiently accurately. Generally neither is restrictive, as the nonparametric estimation of $\mu_0(x)$ is the most statistically difficult estimation in this setting. Finally, part (iii) ensures that we have a feasible estimator of the Gram matrix that converges rapidly enough. The infeasible Gram matrix $\bar{\mathbf{Q}}_p$ defined above is not a population object, but rather retains the randomness of the estimated basis. This will be key in our results and is discussed following Theorem 1. See Section SA-4.1 for examples of $\widehat{\Upsilon}(x_i, \mathbf{w}_i)$ for different models.

Our first theoretical result is a uniform (in x) Bahadur representation for $\widehat{\vartheta}_p(x, \widehat{\mathsf{w}})$ as defined in (2.4).

Theorem 1 (Bahadur Representation)**.** *Suppose that Assumptions 1, 2, and 3 hold, and*

that $J^{\frac{\nu}{\nu-2}}\log n + J(\log n)^{7/3} + (J^2(\log n))^{\frac{\nu}{\nu-1}} = o(n)$ *and* $\log n = o(J)$. *Then,*

$$\sup_{x\in\mathcal{X}} \left|\widehat{\vartheta}_p(x,\widehat{\mathsf{w}}) - \vartheta_0(x,\mathsf{w}) - \widehat{\mathsf{L}}_p(x,\mathsf{w})\right| = O_{\mathbb{P}}(r_n),$$

where

$$\widehat{\mathsf{L}}_p(x,\mathsf{w}) := \eta^{(1)}(\theta_0(x,\mathsf{w}))\widehat{\mathbf{b}}_p(x)'\bar{\mathbf{Q}}_p^{-1}\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{b}}_p(x_i)\eta^{(1)}(\theta_0(x_i,\mathbf{w}_i))\psi(y_i,\eta(\theta_0(x_i,\mathbf{w}_i))),$$

and $r_n := \left(\frac{J\log n}{n}\right)^{3/4}\log n + J^{-\frac{p}{2}}\left(\frac{\log^2 n}{n}\right)^{1/2} + J^{-p-1} + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| + \|\widehat{\mathsf{w}} - \mathsf{w}\|$.

This result is essentially a stochastic linearization of the estimator, and yields important consequences including the mean squared error expansion used to choose J and the asymptotic variance formula for inference. The form of the variance is reminiscent of its parametric counterpart (e.g., for generalized linear models), but estimation is more complicated. Herein we maintain general high-level conditions justifying several alternatives commonly used in practice. These are discussed in SA-4.1. The analogous Bahadur representations for $\widehat{\mu}_p^{(v)}(x)$ and $\widehat{\zeta}_p(x,\widehat{\mathsf{w}})$, under the same assumptions, are given in Theorem SA-3.1. The "linear" term is slightly different to account for the different structure of the three estimands and the remainder rate for derivative estimation is slower.

With the Bahadur representation in place, we can develop tools for inference. Our main result is a strong approximation for the (Studentized) t-statistic process for each of the three estimators, allowing us to obtain a feasible asymptotic distributional approximation. Again we give the details only for $\widehat{\vartheta}_p(x,\widehat{\mathsf{w}})$ and defer the others to the SA. The variance is an immediate consequence of the expansion in Theorem 1, and is made feasible by replacing unknown objects by their estimators. For a given p, define the statistic

$$T_{\vartheta,p}(x) = \frac{\widehat{\vartheta}_p(x,\widehat{\mathsf{w}}) - \vartheta_0(x,\mathsf{w})}{\sqrt{\widehat{\Omega}_{\vartheta,p}(x)/n}}, \qquad \widehat{\Omega}_{\vartheta,p}(x) := \eta^{(1)}(\widehat{\theta}_p(x,\widehat{\mathsf{w}}))^2\widehat{\mathbf{b}}_p(x)'\widehat{\mathbf{Q}}_p^{-1}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{Q}}_p^{-1}\widehat{\mathbf{b}}_p(x), \quad (3.1)$$

where $\widehat{\mathbf{\Sigma}}_p := n^{-1}\sum_{i=1}^n \widehat{\mathbf{b}}_p(x_i)\widehat{\mathbf{b}}_p(x_i)'\psi(y_i,\eta(\widehat{\theta}_p(x_i,\mathbf{w}_i)))^2\eta^{(1)}(\widehat{\theta}_p(x_i,\mathbf{w}_i))^2$ and $\widehat{\mathbf{Q}}_p$ is defined in Assumption 3. The t-statistics $T_{\mu^{(v)},p}(x)$ and $T_{\zeta,p}(x)$, for $\mu_0^{(v)}(x)$ and $\zeta_0(x,\mathsf{w})$ respectively, are entirely analogous, and are defined in Section SA-3.3.

Our inference results will follow from the next key theorem.

Theorem 2 (Strong Approximation)**.** *Suppose that Assumptions 1, 2, and 3 hold, and let* $(a_n : n \geq 1)$ *be a sequence of non-vanishing constants such that* J *and* $\widehat{\mathsf{w}}$ *obey*

$$\frac{J(\log n)^2}{n^{1-\frac{2}{\nu}}} + \Big(\frac{J(\log n)^7}{n}\Big)^{1/2} + nJ^{-2p-3} + \frac{(\log n)^2}{J^{p+1}} + nJ^{-1}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2 = o(a_n^{-2}),$$

$\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(a_n^{-1}\sqrt{J/n})$*, and* $(J^2\log(n))^{\nu/(\nu-1)} = o(n)$*. Then, on a properly enriched probability space, there exists a* $(J+p)$*-dimensional standard Normal random vector* $\mathbf{N}$ *such that for any* $\xi > 0$*,*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|T_{\vartheta,p}(x) - \bar{Z}_{\vartheta,p}(x)| > \xi a_n^{-1}\Big) = o(1), \qquad \bar{Z}_{\vartheta,p}(x) = \frac{\widehat{\mathbf{b}}_p(x)'\eta^{(1)}(\theta_0(x,\mathsf{w}))\bar{\mathbf{Q}}_p^{-1}\bar{\mathbf{\Sigma}}_p^{1/2}}{\sqrt{\bar{\Omega}_{\vartheta,p}(x)}}\mathbf{N},$$

where $\bar{\mathbf{\Sigma}}_p$ *and* $\bar{\Omega}_{\vartheta,p}(x)$ *are shown in Section SA-3. On a further enriched space, there exists a conformable standard Normal vector* $\mathbf{N}^\star$*, independent of* $\{(y_i,x_i,\mathbf{w}_i')'\}_{i=1}^n$*, such that for any* $\xi > 0$*,*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{X}}|\bar{Z}_{\vartheta,p}(x) - \widehat{Z}_{\vartheta,p}(x)| > \xi a_n^{-1}\Big|\{(y_i,x_i,\mathbf{w}_i')'\}_{i=1}^n\Big) = o_{\mathbb{P}}(1),$$
$$\widehat{Z}_{\vartheta,p}(x) = \frac{\widehat{\mathbf{b}}_p(x)'\eta^{(1)}(\widehat{\theta}_p(x,\widehat{\mathsf{w}}))\widehat{\mathbf{Q}}_p^{-1}\widehat{\mathbf{\Sigma}}_p^{1/2}}{\sqrt{\widehat{\Omega}_{\vartheta,p}(x)}}\mathbf{N}^\star.$$

The approximating process, $\bar{Z}_{\vartheta,p}(\cdot)$, is a Gaussian process conditional on $\{x_i,\mathbf{w}_i\}_{i=1}^n$ by construction, and the elements of $\bar{\mathbf{\Sigma}}_p$ and $\bar{\Omega}_{\vartheta,p}(x)$ reflect this conditioning. This process is infeasible but the second result shows that all the unknown quantities in in $\bar{Z}_{\vartheta,p}(\cdot)$ can be replaced by their sample analogues to obtain a feasible approximation. Theorems SA-3.5 and SA-3.6 give the corresponding results for $T_{\mu^{(v)},p}(x)$ and $T_{\zeta,p}(x)$, under the same assumptions.

Pointwise inference results are also given in the SA for completeness.

Theorems 1 and 2 substantially generalize the least squares results in Cattaneo et al. (2024b), under essentially the same rate restrictions, 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 of nonlinear models (Kong et al., 2010) and series estimation for quantile regression (Belloni et al., 2019). There are several key improvements. First, having sharp rate conditions allows us to accommodate $p = 0$, which is generally excluded by the prior literature but necessary for binned scatter plots. Note that these theorems give approximations for the entire t-statistic process, and not just functionals thereof, under such weak conditions. Prior work has obtained such sharp results only for the supremum of the process. Further, we allow for random partitioning (i.e. series estimation with data-dependent basis functions), which represents a major technical hurdle, and also allow for additional control variables.

In fact, beyond being ruled out by prior work, the randomness in the basis functions requires a novel theoretical approach. The key motivation behind this approach is that the basis functions $\widehat{\mathbf{b}}_p^{(v)}(x)$ do not converge uniformly to a nonrandom counterpart, due to the sharp discontinuity of the (random) indicator functions. It is not possible to obtain the uniform results of Theorem 1 (or the strong approximations below) by expanding around a nonrandom limit. Thus $\widehat{\mathbf{b}}_p^{(v)}(x)$ is left as random in the Bahadur representation, including in the matrix $\bar{\mathbf{Q}}_p$. This further separates our results from prior literature. The more general theorems in the SA are followed by remarks detailing how our work improves on the relevant literature in each case.

4 Tuning Parameter Selection

We can use our theoretical results from the previous section to directly inform implementation in empirical applications. Our first task is selecting the number of bins. The choice of J determines both the visual and statistical properties of the estimator. Consistent estimation

and valid inference is possible for a range of diverging sequences of J, but this does not provide sufficiently precise guidance for implementation. Thus, our first methodological result is a Nagar-type integrated mean squared error expansion that enables an optimal choice of J in empirical applications.

To obtain the result, we need one further assumption to characterize the leading terms of the expansion. Intuitively, we require that the random partition $\widehat{\Delta}$ "converges" to a fixed one which obeys the same restrictions as in Assumption 3. This assumption is not needed for any other results.

Assumption 4. *There exists a non-random partition* $\Delta_0 = \{\mathcal{B}_1, \cdots, \mathcal{B}_J\}$ *with* $\mathcal{B}_j = [\tau_{j-1}, \tau_j)$ *for* $j \leq J-1$ *and* $\mathcal{B}_J = [\tau_{J-1}, \tau_J]$ *such that* $\max_{1\leq j\leq J} |\tau_j - \tau_{j-1}| \leq C \min_{1\leq j\leq J} |\tau_j - \tau_{j-1}|$, *for an absolute constant* $C > 0$ *and* $\max_{1\leq j\leq J} |\hat{\tau}_j - \tau_j| = o_{\mathbb{P}}(J^{-1})$.

This condition trivially holds for well-behaved nonrandom partitions, but also holds for the leading case of quantile-spacing, since sample quantiles converge to their population counterparts. In more general cases with data-driven partitions this condition could fail. However, all our other results remain valid, and furthermore, even if this condition fails a rule-of-thumb choice of J is available, which has the optimal rate but suboptimal constants. See Section SA-4.2 for discussion.

Our IMSE result for $\widehat{\vartheta}_p(x, \widehat{\mathsf{w}})$ is given by the following. The corresponding results for $\widehat{\mu}_p^{(v)}(x)$ and $\widehat{\zeta}_p(x, \widehat{\mathsf{w}})$ are stated in Theorem SA-3.4.

Theorem 3. *Set* $\omega(x)$ *to be a continuous weighting function over* $\mathcal{X}$ *bounded away from zero. Suppose that Assumptions 1, 2, 3, and 4 hold, and let* $J^{\frac{\nu}{\nu-2}} \log n + J^{\frac{2\nu}{\nu-1}} (\log n)^{\frac{\nu}{\nu-1}} + J(\log n)^7 = o(n)$ *and* $\log n = o(\sqrt{J})$, *and* $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(\sqrt{J/n} + J^{-p-1})$. *Then,*

$$\int_{\mathcal{X}} \Big(\widehat{\vartheta}_p(x, \widehat{\mathsf{w}}) - \vartheta_0(x, \mathsf{w})\Big)^2 \omega(x) dx = \mathtt{AISE}_\vartheta + o_{\mathbb{P}}\Big(\frac{J}{n} + J^{-2(p+1)}\Big)$$

where $\texttt{AISE}_\vartheta$ *obeys*

$$\mathbb{E}[\texttt{AISE}_\vartheta | \{(x_i, \mathbf{w}_i')'\}_{i=1}^n, \widehat{\Delta}] = \frac{J}{n}\mathscr{V}_n(p) + J^{-2(p+1)}\mathscr{B}_n(p) + o_{\mathbb{P}}\Big(\frac{J}{n} + J^{-2(p+1)}\Big),$$

for nonrandom terms $\mathscr{V}_n(p)$ *and* $\mathscr{B}_n(p)$ *shown in Theorem SA-3.4 that are bounded and nonzero in general.*

This result is stated in terms of J, as it is the tuning parameter, but the rates and constants depend on the fixed polynomial order p (recall that we set $p = s$, see Section 2.1). An L_2 convergence rate immediately follows from this result. An L_∞ convergence rate is also available in the SA (Corollary SA-3.1). This theorem, and the L_2 and L_∞ rates, are new to the literature, even in the case of non-random partitioning and without covariate adjustment, for nonlinear series estimators and binscatter methods.

The practical consequence of Theorem 3 is that we can balance the (squared) bias and variance to obtain an IMSE-optimal choice of J, which is given by

$$J_{\texttt{IMSE}}(p) := \left(\frac{2(p+1)\mathscr{B}_n(p)}{\mathscr{V}_n(p)}\right)^{\frac{1}{2p+3}} n^{\frac{1}{2p+3}}. \tag{4.1}$$

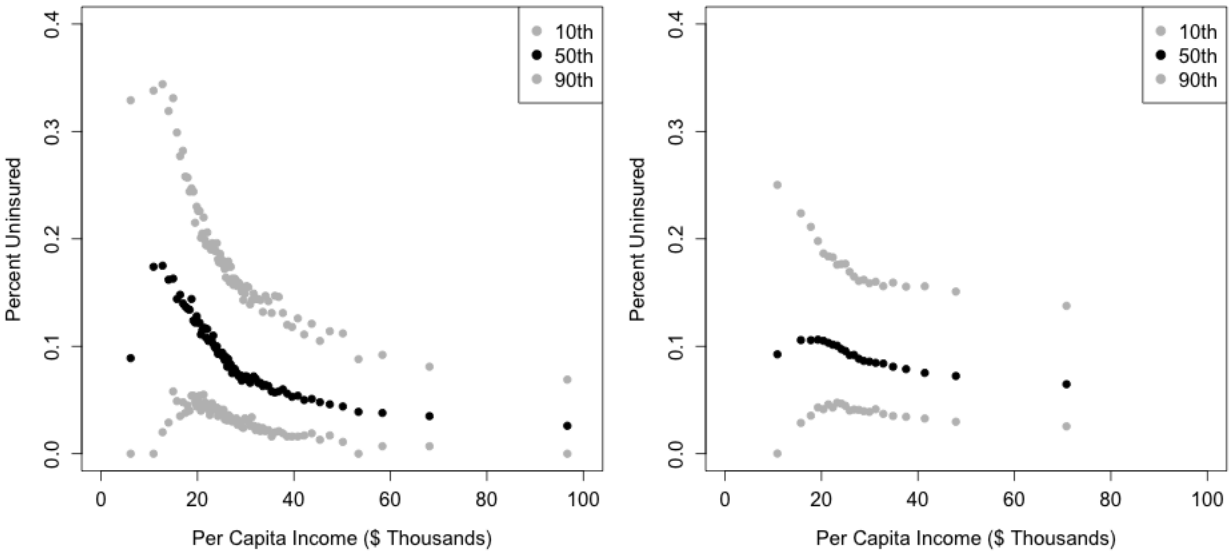
Implementing binscatter with this J is optimal in the sense of providing the IMSE-optimal estimate of the unknown function $\vartheta(\cdot, \mathsf{w})$. In the next section we discuss the use of $J_{\texttt{IMSE}}(p)$ for inference, where, as is typical, a bias correction must be applied. A feasible version of $J_{\texttt{IMSE}}(p)$ is described in Section SA-4 and implemented in the `binsreg` package (Cattaneo et al., 2024a). Section SA-4.3 discusses binned scatter plots with a fixed choice of J, which can be visually appealing and interpretable in some applications.

Figure 2 demonstrates the use of the optimal J for quantile estimation (see Figure 1(d) for the mean). Quantile regression can be used to visualize the spread of the conditional distribution. Observe that Figure 2 restores the visualization of the variability in the data that is present in Figure 1(a) but hidden by the averaging in Figure 1(c). Figure 2(a) shows that there is much larger variance in the fraction insured in lower income areas, but Figure

2(b) shows that this gap narrows when controlling for demographics. In this case we control for (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, and (iv) the local unemployment rate.

Figure 2: **Conditional Quantiles.** This figure illustrates the use of quantile regression to visualize the spread in the ACS data (see Example 4). As the link function is the identity, panel (a) shows estimates of $\vartheta_0(x, \mathsf{w})$ for $\tau = 0.1, 0.5, 0.9$, while panel (b) shows the same quantiles including additional covariates controlling for demographics: (i) percentage of residents with a high school degree, (ii) percentage of residents with a bachelor's degree, (iii) median age of residents, and (iv) the local unemployment rate.

(a) *Quantile Regression w/o Controls* (b) *Quantile Regression w/ Controls*



5 Uniform Inference

We now turn to uniform inference for the three functions defined in Section 2. Uniform inference is required to make statistical statements about the functions $\vartheta_0(x, \mathsf{w})$, $\mu_0^{(v)}(x)$, and $\zeta_0(x, \mathsf{w})$, rather than about their values at a specific point x. Pointwise inference methods (e.g. confidence intervals) will not suffice for our main applications of interest, including treatment effect heterogeneity and continuous treatment effects, as well as shape restrictions and functional forms. For completeness, pointwise inference results are given in the SA and implemented in the `binsreg` package, but omitted here to save space.

A key element of our uniform inference results—from a practical point of view—is the pairing of a feasible tuning parameter choice with valid inference. To give the most accurate estimate, and therefore also visualization, of $\vartheta_0(x,\mathsf{w})$ we prefer to use $J_{\texttt{IMSE}}(p)$ of (4.1). However, as is typical for nonparametric problems, the (I)MSE-optimal tuning parameter choice delivers invalid inference, as it fails to eliminate a first-order bias. We therefore use robust bias correction to ensure that $J_{\texttt{IMSE}}(p)$ remains a valid choice across all uses, delivering optimal estimation and valid inference. With this eye toward practicality, we state all results below specifying $J = J_{\texttt{IMSE}}(p)$, but the SA gives the more general results under mild rate restrictions on J.

Bias correction involves estimating, and removing, the leading smoothing bias term, and is made "robust" by correcting the standard errors to account for the additional sampling variability that has been introduced. Robust bias correction has theoretically superior higher-order inference properties (Calonico et al., 2018, 2022), performs well in simulations, and has been empirically validated in specific contexts (Hyytinen et al., 2018). Robust bias correction is operationalized in the present context by (i) selecting a degree p and creating a partition $\widehat{\Delta}$ based on $J_{\texttt{IMSE}}(p)$ to form the optimal point estimate of $\widehat{\vartheta}_p(x,\widehat{\mathsf{w}})$ and then (ii) conducting inference using $T_{\vartheta,p+1}(x)$ (or its feasible analogue $\widehat{Z}_{\vartheta,p+1}(x)$), i.e. the statistic formed using a higher degree polynomial but the partitioning scheme based on p in (i): $\widehat{\Delta} = \widehat{\Delta}(J_{\texttt{IMSE}}(p))$. Any higher order polynomial may be used, but $p+1$ is simple and robust. Cattaneo et al. (2020) give further discussion of robust bias correction in the context of partition regression, including alternative strategies.

5.1 Confidence Bands

Our first uniform inference result delivers confidence bands for the functions $\vartheta_0(x,\mathsf{w})$, $\mu_0^{(v)}(x)$, and $\zeta_0(x,\mathsf{w})$. Confidence bands are similar in spirit as the more familiar concept of confidence intervals, but instead cover the entire function (uniformly over $x \in \mathcal{X}$) with a prespecified probability. Confidence bands are the appropriate tool for visualizing the uncertainty

around the estimated function. The size of the band also changes to reflect the presence of heteroskedasticity in the data. These bands can be used directly to identify interesting or important features of the function, for example, regions where it is statistically indistinguishable from zero or from a constant function. Bands are also useful for assessing the functional form or shape, such as regions of linearity or monotonicity, and therefore visually complement the formal hypothesis tests we introduce below.

The confidence bands are built from Theorem 2, coupled with robust bias correction, as discussed above. Confidence bands are defined as the area between an upper and lower bounding function. Recall that we employ robust bias correction, so that $\widehat{\vartheta}_{p+1}(x,\widehat{\mathsf{w}})$ is the bias-corrected version of $\widehat{\vartheta}_p(x,\widehat{\mathsf{w}})$, and is thus the "center" of the confidence band, and using $\widehat{\Omega}_{\vartheta,p+1}(x)$ in the standard error accounts for the additional variability. The robust bias-corrected confidence band for $\vartheta_0(x,\mathsf{w})$ is given by

$$\widehat{I}_{\vartheta,p+1}(x) = \left[\widehat{\vartheta}_{p+1}(x,\widehat{\mathsf{w}}) \pm \mathfrak{c}_\vartheta \sqrt{\widehat{\Omega}_{\vartheta,p+1}(x)/n}\right], \tag{5.1}$$

where the critical value is determined by

$$\mathfrak{c}_\vartheta = \inf\left\{ c \in \mathbb{R}_+ : \mathbb{P}\left[\sup_{x\in\mathcal{X}} |\widehat{Z}_{\vartheta,p+1}(x)| \leq c \,\middle|\, \{(y_i,x_i,\mathbf{w}_i')'\}_{i=1}^n\right] \geq 1-\alpha \right\}. \tag{5.2}$$

The asymptotic validity of this confidence band follows from Theorem 2, which allows us to approximate the distribution of the supremum of $T_{\vartheta,p+1}(x)$ by applying the same functional to $\widehat{Z}_{\vartheta,p+1}(x)$. This yields the following result. Here we assume directly that the optimal J is used. Theorem SA-3.8 states a more general result, valid for a range of J, as well as inference results for $\mu_0^{(v)}(x)$ and $\zeta_0(x,\mathsf{w})$.

Theorem 4. *Set $J = J_{\mathtt{IMSE}}(p)$. Suppose that Assumptions 1, 2, and 3 hold, with $p+1$ in place of p and $\nu > 3$, and let $\|\widehat{\mathsf{w}} - \mathsf{w}\| = o_{\mathbb{P}}(\sqrt{J/(n\log J)})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = o_{\mathbb{P}}(\sqrt{J/(n\log J)})$.*

Then, for $\widehat{I}_{\vartheta,p+1}(x)$ *defined in* (5.1),

$$\mathbb{P}\Big[\vartheta_0(x,\mathsf{w}) \in \widehat{I}_{\vartheta,p+1}(x),\ \textit{for all } x \in \mathcal{X}\Big] = 1-\alpha+o(1).$$

This result establishes valid confidence bands for generalized, covariate-adjusted binscatters. We can use this result to visually assess uncertainty about the form and shape of the regression function. One can visually "test" hypotheses of interest, though formal testing (Section 5.2) is recommended. Plotting both the estimate $\widehat{\vartheta}_p(x,\widehat{\mathsf{w}})$ and the band $\widehat{I}_{\vartheta,p+1}(x)$ is advisable in applications because doing so presents both the IMSE-optimal point estimate and a valid measure of uncertainty (and one that uses the same bins).

In nonlinear models, particularly in social sciences, partial effects are often the preferred way of summarizing the relationship (causal or not) of x_i to y_i, controlling for $\mathbf{w}_i$. In our setting this corresponds to the estimate of $\zeta_0(x,\mathsf{w})$ and its associated confidence band. When x_i is a treatment variable, $\zeta_0(x,\mathsf{w})$ captures the effect of increasing the treatment dosage, and the band can help identify regions of $\mathcal{X}$ with the largest effects, or any other noteworthy shape.

Figure 3 shows examples of confidence bands using our running empirical application. The confidence band in Figure 3(a) displays the uncertainty surrounding the estimate first shown in Figure 1(d). The presence of the Medicaid program is clearly delineated by the shape of the band at lower income levels. From the band we can immediately conclude that the relationship is nonmonotonic. This is further emphasized in Figure 3(b), showing the marginal effect. Using the bands, we can reject the null hypothesis of monotonicity as the band lies completely on either side of zero at low and high income levels.

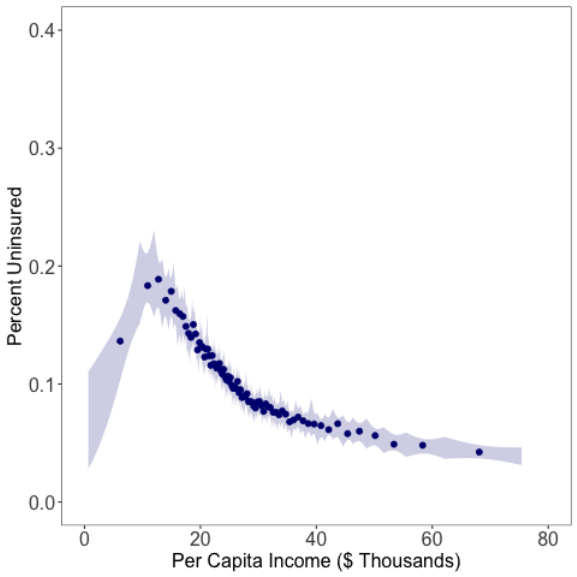
There are two features of our confidence bands that warrant mention. First, the user-selected point of evaluation w can impact the shape, placement, and size, of the confidence band. One might expect that since the additional controls are modeled as additively linear, the evaluation point w (and the coefficient $\boldsymbol{\gamma}_0$) should not impact conclusions about the

nonparametric relationship between y and x. But this intuition overlooks the fact that the function $\mu_0(x)$ is only defined relative to how $\mathbf{w}_i$ is coded. For example, if $\mathbf{w}_i$ contains a binary variable indicating groups of substantially different sizes, then the estimation uncertainty will be different between the two groups. This can cause a level shift in $\widehat{\mu}_p^{(v)}(x)$ and alter the uncertainty around the estimate. For $\vartheta_0(x, \mathsf{w})$ and $\zeta_0(x, \mathsf{w})$, the shape may also change. This impacts all aspects of inference, both visual and the formal tests below. This is not particular to our method; it is always present in analysis of models like (2.1).

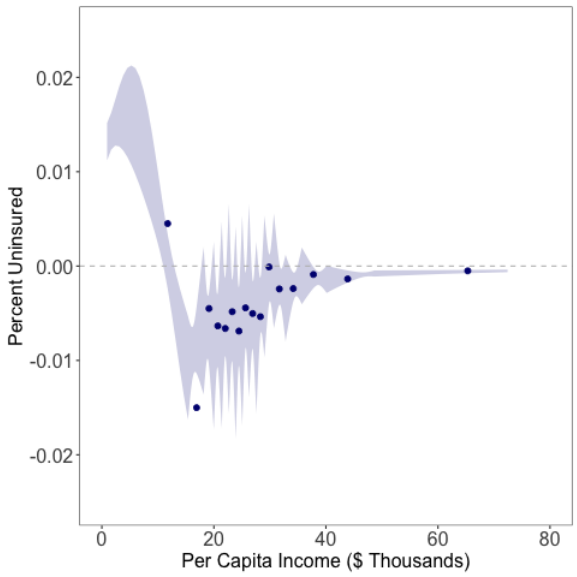
Second, the bias correction may result in the point estimate lying outside the confidence band. This occurs in regions of high bias. This is formally correct but can be visually unappealing. Figure 3(b) shows an example of this phenomenon. This can arise in any application of bias correction methods, and is not necessarily a failing: the point estimate remains IMSE-optimal and inference remains valid.

Figure 3: **Confidence Bands.** This figure illustrates confidence bands in a nonlinear binned scatter plot using the ACS data. Panel (a) shows the point estimate (dots) and robust bias corrected confidence band (shaded region) for the conditional mean function with no controls, i.e., $\vartheta_0(x) = \eta(\mu_0(x))$, while panel (b) shows the corresponding point estimate and confidence band for the marginal effect, $\zeta_0(x)$. Shaded regions denote 95% confidence bands and are based on 50,000 random draws.

(a) *Conditional Mean*



(b) *Marginal Effect*



5.2 Hypothesis Testing

We also provide formal hypothesis tests for substantive questions including functional form or shape restrictions for $\vartheta_0(x,\mathsf{w})$, $\mu_0(x)$, and $\zeta_0(x,\mathsf{w})$. Our discussion here is brief. Full details are given in the SA.

A leading case is testing a parametric functional form for $\mu_0(x)$. This is a two-sided testing problem where under the null there exists some finite dimensional parameter $\boldsymbol{\theta}$ such that $\mu_0(x) = m(x;\boldsymbol{\theta})$, uniformly in $x \in \mathcal{X}$ (we can also test any derivative of $\mu_0(x)$). The testing problem is

$$\begin{aligned} \dot{\mathsf{H}}_0^{\mu}: &\quad \sup_{x\in\mathcal{X}}\left|\mu_0(x) - m(x;\boldsymbol{\theta})\right| = 0, \qquad \text{for some } \boldsymbol{\theta}, \qquad vs. \\ \dot{\mathsf{H}}_{\mathrm{A}}^{\mu}: &\quad \sup_{x\in\mathcal{X}}\left|\mu_0(x) - m(x;\boldsymbol{\theta})\right| > 0, \qquad \text{for all } \boldsymbol{\theta}. \end{aligned}$$

This test formalizes the notion of visual inspection in plots like Figure 1(c) and (d), beyond what is already done by adding a confidence band. We test this hypothesis using the statistic

$$\dot{T}_{\mu,p+1}(x) := \frac{\widehat{\mu}_{p+1}(x) - m(x;\widetilde{\boldsymbol{\theta}})}{\sqrt{\widehat{\Omega}_{\mu,p+1}(x)/n}},$$

where $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\gamma}}$ are estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_0$ that are consistent under $\dot{\mathsf{H}}_0^{\mu}$. Theorem 2 again provides the tools to obtain the appropriate critical value. Theorem SA-3.9 gives the formal result showing size control and consistency of the test, as well as the corresponding tests for $\vartheta_0(x,\mathsf{w})$ and $\zeta_0(x,\mathsf{w})$. The tests can be performed using any L_q norm for $q \geq 1$, instead of L_∞ as shown above. Last, we also provide for testing shape restrictions, which are conceptually similar but are generally one-sided testing problems. A leading example would be testing monotonicity of $\vartheta_0(x,\mathsf{w})$ or $\zeta_0(x,\mathsf{w})$.

Table 1 shows several testing examples using the L_∞ norm. Consider first the left column of results. We test against the linear specification model, formalizing the visual comparison in Figure 1(d). We also test against a cubic (in x) logistic quasi-likelihood model for

added flexibility. Both parametric specifications are rejected, and moreover, also rejected when including other controls. This highlights the need for nonparametric modeling in this application. Finally we test the substantive null hypothesis that the uninsuredness rate is monotonically decreasing with income. This null is also strongly rejected due to the existence of Medicaid. This motivates the right column of results, where we repeat the analysis after restricting the sample to zip codes with per capita income above 138% of the 2013–2017 average federal poverty line for a single-person household ($16,248). This is the cutoff for expanded Medicaid eligibility based only on income. When we restrict to this sample, which diminishes the influence of the Medicaid program, we fail to reject the null hypothesis of a monotonic decline, but still reject the parametric specifications. This aligns with the need for flexible estimation and matches the conclusions we draw from the shape of the confidence bands shown in Figure 3.

Table 1: **Specification and Shape Testing**

	Full Sample			**Above Income Cutoff**		
	Test Stat.	p-value	$\hat{J}_{\texttt{IMSE}}$	Test Stat.	p-value	$\hat{J}_{\texttt{IMSE}}$
Test of Linear Fit						
No Covariates	3113.083	0.000	80	4315.983	0.000	40
Covariates, $\widehat{\mathsf{w}} = \bar{\mathbf{w}}$	1979.468	0.000	22	2908.763	0.000	12
Test of Cubic Fit						
No Covariates	2245.499	0.000	80	14814.862	0.000	40
Covariates, $\widehat{\mathsf{w}} = \bar{\mathbf{w}}$	1981.105	0.000	22	3587.529	0.000	12
Test of Monotonic Decline						
No Covariates	23.991	0.000	16	0.644	0.998	10
Covariates, $\widehat{\mathsf{w}} = \bar{\mathbf{w}}$	6.997	0.000	13	-0.016	1.000	8

Notes. This table reports the test statistics and associated p-values along with the IMSE-optimal choice of J from hypothesis tests of parametric specifications and shape restrictions using the ACS data. The first and second panels report test results for the null hypotheses of linear and cubic (in x) logistic quasi-likelihood models, respectively, while the third panel reports results for the null hypothesis of a monotonic decline in the level (i.e., negative derivative). All tests are performed with and without control variables (control variables are same as in Figure 2). The left panel ("Full Sample") reports results for the full sample whereas the right panel ("Above Income Cutoff") restricts to the sample of zip codes with per capita income above $16,248. All p-values are based on 50,000 random draws.

5.3 Multi-Sample Comparisons

Our results extend to comparisons between different samples, or groups, within the data. This is a common goal in program evaluation and causal inference settings. With discrete (e.g., binary) treatments, the groups are defined by treatment arms and the differences define heterogeneous (in x_i) effects. In the continuous case, the grouping is the dimension of heterogeneity and x_i is the treatment. Our results extend naturally to this setting. For a grouping indicator $t_i = 0, \ldots, L$, we replace the scalar index in the model (2.1) with $\theta_0(x_i, \mathbf{w}_i, t_i) := \sum_{t=0}^{L} \mathbb{1}\{t_i = t\}\theta_{0,t}(x_i, \mathbf{w}_i)$, where each $\theta_{0,t}(x_i, \mathbf{w}_i) = \mu_{0,t}(x_i) + \mathbf{w}_i'\boldsymbol{\gamma}_{0,t}$. The level and marginal effect can then be defined groupwise, as $\vartheta_{0,t}(x, \mathsf{w}) = \eta(\theta_{0,t}(x, \mathsf{w}))$ and $\zeta_{0,t}(x, \mathsf{w}) = \eta^{(1)}(\theta_{0,t}(x, \mathsf{w}))\mu_{0,t}^{(1)}(x)$ for some evaluation point w of control variables.

For example, in a randomized experiment $\vartheta_{0,1}(x, \mathsf{w}) - \vartheta_{0,0}(x, \mathsf{w})$ is the conditional average treatment effect (CATE) function, and the binscatter naturally captures treatment effect heterogeneity along the x_i dimension holding fied $\mathbf{w}_i = \mathsf{w}$. The rate of change in this heterogeneity is $\zeta_{0,1}(x, \mathsf{w}) - \zeta_{0,0}(x, \mathsf{w})$. Our methods can be used to formally test the null hypothesis that $\vartheta_{0,1}(x, \mathsf{w}) = \vartheta_{0,0}(x, \mathsf{w})$ 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:

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

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.

Figure 4 shows an example of this type of analysis. We divide states into two groups based on their population density, with low and high density states as those with population

densities below or above 100 people per square mile, respectively. Density is defined as the average population per square mile, and the data is available from the Census Bureau. Panels (a) shows $\widehat{\vartheta}_t(x)$ for each group, i.e. without controls, while panel (b) adds controls and shows $\widehat{\vartheta}_t(x, \widehat{\mathsf{w}})$, with $\widehat{\mathsf{w}}$ set to the sample mean. The point estimates show 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 except for very low incomes. In contrast, when covariates are added, there is a much clearer delineation between the two groups at all but the lowest of income levels. This is made clear in panels (c) and (d), which plot the point estimate of the difference (the CATE) and the associated confidence band. The null hypothesis that $\vartheta_{0,1}(x, \mathsf{w}) = \vartheta_{0,0}(x, \mathsf{w})$ for all $x \in \mathcal{X}$ is rejected in both cases, with test statistics of 7.719 and 8.308, respectively, and negligible p-values. Multi-sample comparisons share the same sensitivity to the chosen evaluation point as discussed above. These issues are unavoidable; researchers must be mindful when implementing the tests and interpreting the results.

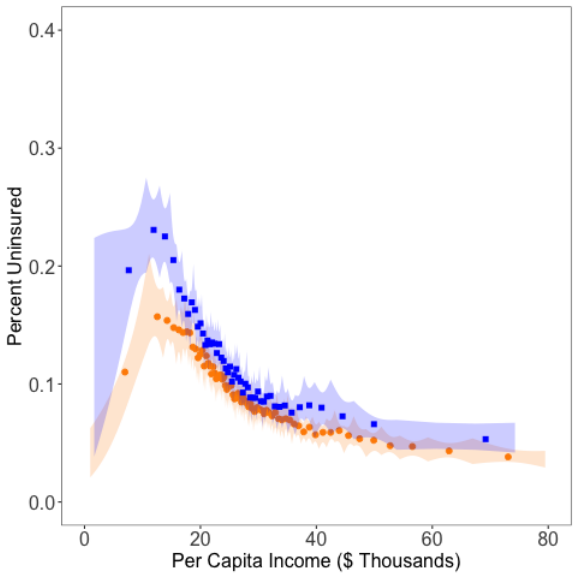
6 Conclusion

With the rise of large data sets, new visualization tools, such as binned scatter plots, have emerged and gained in popularity. This paper has thoroughly studied binned scatter plots in nonlinear, nonsmooth regression models. Our main contributions are to propose novel nonlinear binscatter methods, together with IMSE-optimal tuning parameter selection and uniform inference methods, including valid confidence bands and functional testing. Our companion `binsreg` software package makes these tools available for applications.

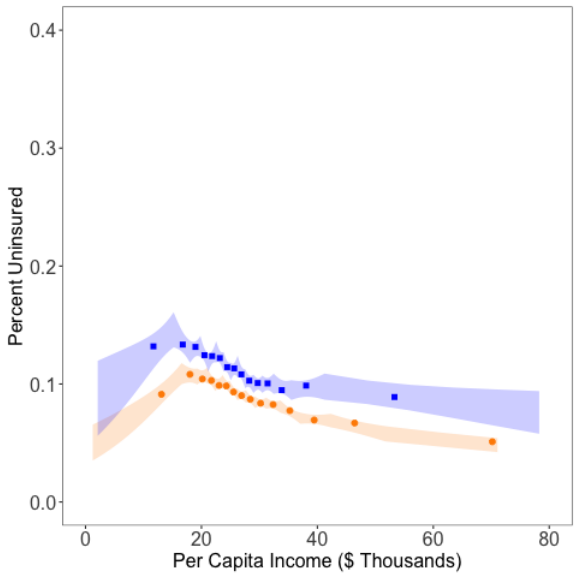
One avenue for future work would be to generalize the analysis beyond a scalar covariate of interest. For example, in two dimensions such an approach would produce "heat maps" which are the bivariate extension of binned scatter plots. Extending our results to that case would be a valuable addition to the practitioner's toolkit.

Figure 4: **Two Sample Comparison.** This figure uses the same ACS data to compare areas in low density states (blue) and high density states (orange). Low density states are defined as those with average population per square mile below 100. Panels (a) and (b) show the point estimate (squares or dots) and robust bias corrected confidence band (shaded region) for each group, first without control variables and then with controls added (the same controls as in Figure 2). Panels (c) and (d) show the estimated difference (evaluated using the binning of the low density states) and the associated confidence bands. Shaded regions denote 95% confidence bands and are based on 50,000 random draws.

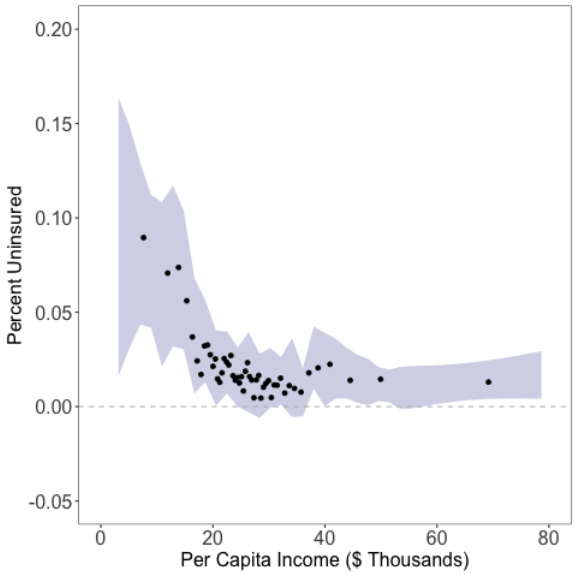
(a) *Conditional means w/o controls*



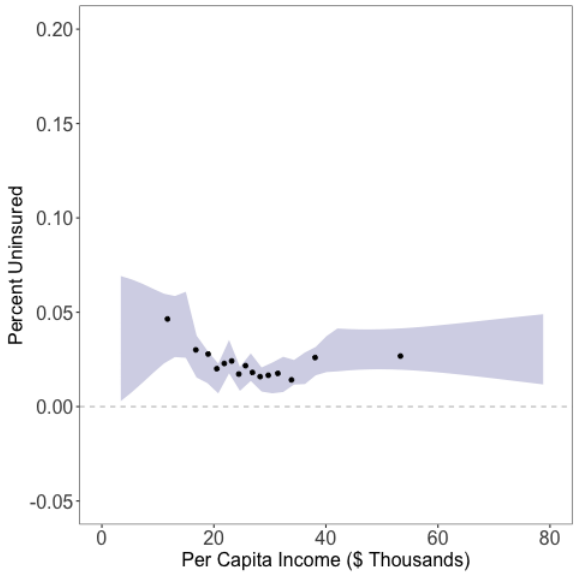
(b) *Conditional means w/ controls*



(c) *Difference of means w/o controls*



(d) *Difference of means w/ controls*



References

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

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

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

——— (2022), "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," *Bernoulli*, 28, 2998–3022.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024a), "Binscatter Regressions," *Stata Journal, revised and resubmit*.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024b), "On Binscatter," *American Economic Review*, 114, 1488–1514.

Cattaneo, M. D., and Farrell, M. H. (2013), "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127–143.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020), "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, 48, 1718–1741.

Devroye, L., Györfi, L., and Lugosi, G. (2013), *A Probabilistic Theory of Pattern Recognition*, Vol. 31, Springer Science & Business Media.

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

Healy, K. (2018), *Data Visualization: A Practical Introduction*, Princeton University Press.

Huang, J. Z. (2003), "Local Asymptotics for Polynomial Spline Regression," *Annals of Statistics*, 31, 1600–1635.

Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O., and Tukiainen, J. (2018), "When Does Regression Discontinuity Design Work? Evidence from Random Election Outcomes," *Quantitative Economics*, 9, 1019–1051.

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

Papke, L. E., and Wooldridge, J. M. (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 619–632.

Schwabish, J. (2021), *Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks*, Columbia University Press.

Shen, X., Wolfe, D., and Zhou, S. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *Annals of Statistics*, 26, 1760–1782.

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

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

Zhang, H., and Singer, B. H. (2010), *Recursive Partitioning and Applications*, Springer Science & Business Media.