EMPIRICAL LIKELIHOOD FOR
REGRESSION DISCONTINUITY DESIGN

By

Taisuke Otsu and Ke-Li Xu

May 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1799

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Empirical Likelihood for Regression Discontinuity Design[*]

Taisuke Otsu[†]
Yale University

Ke-Li Xu[‡]
Texas A&M University and University of Alberta

1st May 2010

Abstract

This paper proposes empirical likelihood based inference methods for causal effects identified from regression discontinuity designs. We consider both the sharp and fuzzy regression discontinuity designs and treat the regression functions as nonparametric. The proposed inference procedures do not require asymptotic variance estimation and the confidence sets have natural shapes, unlike the conventional Wald-type method. These features are illustrated by simulations and an empirical example which evaluates the effect of class size on pupils' scholastic achievements. Bandwidth selection methods, higher-order properties, and extensions to incorporate additional covariates and parametric functional forms are also discussed.

Keywords: Empirical likelihood; Nonparametric methods; Regression discontinuity design; Treatment effect

JEL Classifications: C12; C14; C21

1 Introduction

Since the seminal work of Thistlethwaite and Campbell (1960), regression discontinuity design (RDD) analysis has been a fundamental tool to investigate causal effects of treatment assignments on outcomes of interest. There are numerous methodological developments and empirical applications of RDD

[*] We would like to thank Li Gan, Timothy Gronberg, Hidehiko Ichimura, Susumu Imai, Steven Lehrer, Qi Li, James MacKinnon, Vadim Marmer, Thanasis Stengos, an associate editor, an anonymous referee, and seminar participants at Queen's University, Texas A&M University, University of Guelph, University of Tokyo, SETA 2009 in Kyoto, and FEMES 2009 in Tokyo for helpful comments. Our research is supported by the National Science Foundation under SES-0720961 (Otsu) and University of Alberta School of Business through the Canadian utilities faculty award (Xu).

[†] Cowles Foundation and Department of Economics, Yale University. *Mailing address*: P.O. Box 208281, New Haven, CT 06520-8281, USA. *E-mail address*: taisuke.otsu@yale.edu.

[‡] Department of Economics, Texas A&M University, and University of Alberta School of Business. *Mailing address*: 3-40N Business building, University of Alberta School of Business, Edmonton, Alberta T6G 2R6, Canada. *E-mail address*: keli.xu@ualberta.ca.

analysis particularly in the fields of economics, psychology, and statistics (see e.g. Trochim, 2001, and Imbens and Lemieux, 2008, for surveys). The main purpose of this paper is to propose a new inference approach to RDD analysis based on empirical likelihood.[1]

In the literature of RDD analysis, there are at least two important issues that have attracted substantial attention from researchers. First, although RDD analysis were initially discussed in the context of regression analysis, recent research has focused on deeper understanding of the estimated parameters of interest based on the theory of causal effects (see e.g. Rubin, 1974, Holland, 1986, and Angrist, Imbens and Rubin, 1996). In causal analysis, RDDs are split into two categories, the sharp and fuzzy RDDs. This categorization is based on how the treatment assignments are determined by a covariate (called the forcing variable). For the sharp design, the treatment is completely determined by the forcing variable on the either side of a cutoff value and we can identify and estimate the average causal effect of the treatment at the cutoff value. For the fuzzy design, the treatment is partly determined by the forcing variable and the treatment assignment probability jumps at the cutoff value. In this case, we can identify and estimate the average causal effect of the treatment for the compliers (see Hahn, Todd and van der Klaauw, 2001, and Section 2.1 below). The present paper adopts this framework and focuses on inferences for the average causal effects identified in the sharp and fuzzy RDDs.

The second issue that has attracted researchers' attention is the importance of nonparametric methods in RDD analysis (e.g. Sacks and Ylvisaker, 1978, Knafl, Sacks and Ylvisaker, 1985). Since RDD analysis is concerned with the causal effects locally at some cutoff value of the forcing variable, it is natural to allow flexible functional forms for regression and treatment assignment probability functions. Hahn, Todd and van der Klaauw (2001) and Porter (2003) proposed nonparametric estimators for average causal effects in the sharp and fuzzy RDDs based on local polynomial fitting (Fan and Gijbels, 1996). Their nonparametric estimators possess reasonable convergence rates and are asymptotically normal under certain regularity conditions. However, the asymptotic variances of these estimators, which are required to construct Wald-type confidence sets, are rather complicated due to discontinuities in the conditional mean, variance, and covariance functions. Typically, in order to estimate the asymptotic variances, we need additional nonparametric regressions to estimate the left and right limits of the conditional variances and covariances, and we also need nonparametric density estimation for the forcing variable. In this paper we construct empirical likelihood-based confidence sets which allow for nonparametric regression functions but do not require complicated asymptotic variance estimation.

This circumvention of asymptotic variance estimation for the empirical likelihood-based confidence sets is not only a practical but also theoretical matter. Chen and Qin (2000) showed that the empirical likelihood confidence set for a conditional mean function at boundary points have better higher-order coverage properties than the Wald-type confidence sets. Chen and Qin confirmed that this refinement follows from the fact that empirical likelihood naturally internalizes the estimation of an asymptotic

[1] See Owen (2001) for a review on empirical likelihood.

variance component which needs to be estimated to construct the Wald statistic. This fact makes our application of empirical likelihood particularly attractive since RDD analysis is mainly concerned with inference on conditional mean functions at boundary (or cutoff) points. Indeed, our empirical likelihood construction is an extension of Chen and Qin (2000) to the sharp and fuzzy RDD setups, and we can interpret Chen and Qin's empirical likelihood as a special case of ours (see Section 4.2). We show that the empirical likelihood ratios for the causal effects in the sharp and fuzzy RDDs are asymptotically chi-square distributed. Therefore, similar to the existing papers such as Chen and Qin (2000) and Fan, Zhang and Zhang (2001), we can still observe an analog of Wilks's phenomenon in this nonparametric RDD setup.

The paper is organized as follows. In Section 2 we present the basic setup and construct the empirical likelihood function for the causal effects. Section 3 studies first-order asymptotic properties of the empirical likelihood ratios and confidence sets. Section 4 discusses bandwidth selection methods, higher-order properties, and extensions to incorporate additional covariates and parametric functional forms. The proposed methods are examined in Section 5 through Monte Carlo simulations and an empirical example which evaluates the effect of class size on pupils' scholastic achievements investigated in Angrist and Lavy (1999). Section 6 concludes. Appendix A contains the proofs and lemmas for the main theorems.

2 Setup and Methodology

2.1 Regression Discontinuity Design

We first introduce our basic setup. Let $Y_i(1)$ and $Y_i(0)$ be potential outcomes of unit i with and without exposure to a treatment, respectively. Let $W_i \in \{0,1\}$ be an indicator variable for the treatment. We set $W_i = 1$ if unit i is exposed to the treatment and set $W_i = 0$ otherwise. The observed outcome is $Y_i = (1 - W_i) Y_i(0) + W_i Y_i(1)$ and we cannot observe $Y_i(0)$ and $Y_i(1)$ simultaneously. Our purpose is to make inference on the causal effect of the treatment, or more specifically, probabilistic aspects of the difference of potential outcomes $Y_i(1) - Y_i(0)$. RDD analysis focuses on the case where the treatment assignment W_i is completely or partly determined by some observable covariate X_i, called the forcing variable. For example, to study the effect of class size on pupils' achievements, it is reasonable to consider the following setup: the unit i is school, Y_i is an average exam score, W_i is an indicator variable for the class size ($W_i = 0$ for one class and $W_i = 1$ for two classes), and X_i is the number of enrollments.

Depending on the assignment rule for W_i based on X_i, we have two cases, called the sharp and fuzzy RDDs. In the sharp RDD, the treatment is deterministically assigned based on the value of X_i, i.e.

$$W_i = \mathbb{I}\{X_i \geq c\},$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and c is a known cutoff point. A parameter of interest in this case is the average causal effect at the discontinuity point c,

$$\theta_s = \mathrm{E}\left[\left. Y_i(1) - Y_i(0) \right| X_i = c \right].$$

Since the difference of potential outcomes $Y_i(1) - Y_i(0)$ is unobservable, we need a tractable representation of θ_s in terms of quantities that can be estimated by data. If the conditional mean functions $\mathrm{E}\left[\left. Y_i(1)\right| X_i = x\right]$ and $\mathrm{E}\left[\left. Y_i(0)\right| X_i = x\right]$ are continuous at $x = c$, then the average causal effect θ_s can be identified as a contrast of the right and left limits of the conditional mean $\mathrm{E}\left[\left. Y_i\right| X_i = x\right]$ at $x = c$,

$$\theta_s = \lim_{x \downarrow c} \mathrm{E}\left[\left. Y_i\right| X_i = x\right] - \lim_{x \uparrow c} \mathrm{E}\left[\left. Y_i\right| X_i = x\right]. \tag{1}$$

In contrast to sharp RDD analysis, fuzzy RDD analysis focuses on the case where the forcing variable X_i is not informative enough to determine the treatment W_i but can affect on the treatment probability. In particular, the fuzzy RDD assumes that the conditional treatment probability of W_i jumps at $X_i = c$,

$$\lim_{x \downarrow c} \Pr\left\{\left. W_i = 1\right| X_i = x\right\} \neq \lim_{x \uparrow c} \Pr\left\{\left. W_i = 1\right| X_i = x\right\}.$$

To define a reasonable parameter of interest for the fuzzy case, let $W_i(x)$ be a potential treatment for unit i when the cutoff level for the treatment was set at x, and assume that $W_i(x)$ is non-increasing in x at $x = c$. Using the terminology of Angrist, Imbens and Rubin (1996), unit i is called a complier if her cutoff level is X_i, i.e.[2]

$$\lim_{x \downarrow X_i} W_i(x) = 0, \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

A parameter of interest in the fuzzy RDD, suggested by Hahn, Todd and van der Klaauw (2001), is the average causal effect for compliers at $X_i = c$,

$$\theta_f = \mathrm{E}\left[\left. Y_i(1) - Y_i(0)\right| i \text{ is complier}, X_i = c\right].$$

Hahn, Todd and van der Klaauw (2001) showed that under mild conditions the parameter θ_f can be identified by the ratio of the jump in the conditional mean of Y_i at $X_i = c$ to the jump in the conditional treatment probability at $X_i = c$, i.e.

$$\theta_f = \frac{\lim_{x \downarrow c} \mathrm{E}\left[\left. Y_i\right| X_i = x\right] - \lim_{x \uparrow c} \mathrm{E}\left[\left. Y_i\right| X_i = x\right]}{\lim_{x \downarrow c} \Pr\left\{\left. W_i = 1\right| X_i = x\right\} - \lim_{x \uparrow c} \Pr\left\{\left. W_i = 1\right| X_i = x\right\}}. \tag{2}$$

If additional covariates Z_i are available, the same identification arguments for θ_s and θ_f go through by slightly modifying the assumptions and adding conditioning variables $Z_i = z$ to the conditional means and probabilities above. This paper focuses on how to make inference for these average causal effect parameters θ_s and θ_f in the sharp and fuzzy RDDs.

To estimate the parameters θ_s and θ_f, it is common to apply some nonparametric regression techniques (e.g. Hahn, Todd and van der Klaauw, 2001, and Porter, 2003). For example, the left and

[2] If $\lim_{x \downarrow X_i} W_i(x) = 0$ and $\lim_{x \uparrow X_i} W_i(x) = 0$, then unit i is called a nevertaker. If $\lim_{x \downarrow X_i} W_i(x) = 1$ and $\lim_{x \uparrow X_i} W_i(x) = 1$, then unit i is called an alwaystaker.

right limits of the conditional mean $\alpha_l = \lim_{x \uparrow c} \mathrm{E}\left[Y_i | X_i = x \right]$ and $\alpha_r = \lim_{x \downarrow c} \mathrm{E}\left[Y_i | X_i = x \right]$ can be estimated by local linear regression estimators $\hat{\alpha}_l$ and $\hat{\alpha}_r$, i.e. solutions to the following weighted least square problems with respect to a_l and a_r,

$$\min_{a_l, b_l} \sum_{i: X_i < c} \mathbb{K}\left(\frac{X_i - c}{h} \right) \left(Y_i - a_l - b_l \left(X_i - c \right) \right)^2, \tag{3}$$

$$\min_{a_r, b_r} \sum_{i: X_i \geq c} \mathbb{K}\left(\frac{X_i - c}{h} \right) \left(Y_i - a_r - b_r \left(X_i - c \right) \right)^2,$$

respectively, with a kernel function $\mathbb{K}$ and bandwidth $h = h_n$ satisfying $h \to 0$ as $n \to \infty$. Then from the identification formula (1), the parameter θ_s is estimated by

$$\hat{\theta}_s = \hat{\alpha}_r - \hat{\alpha}_l.$$

In the same manner a nonparametric estimator for θ_f can be obtained as

$$\hat{\theta}_f = \frac{\hat{\alpha}_r - \hat{\alpha}_l}{\hat{\alpha}_{wr} - \hat{\alpha}_{wl}},$$

where $\hat{\alpha}_{wl}$ and $\hat{\alpha}_{wr}$ are estimators for the left and right limits of the conditional treatment probabilities $\alpha_{wl} = \lim_{x \uparrow c} \Pr\left\{ W_i = 1 | X_i = x \right\}$ and $\alpha_{wr} = \lim_{x \downarrow c} \Pr\left\{ W_i = 1 | X_i = x \right\}$, respectively, and are obtained as solutions to the weighted least square problems with respect to a_{wl} and a_{wr},

$$\min_{a_{wl}, b_{wl}} \sum_{i: X_i < c} \mathbb{K}\left(\frac{X_i - c}{h} \right) \left(W_i - a_{wl} - b_{wl} \left(X_i - c \right) \right)^2, \tag{4}$$

$$\min_{a_{wr}, b_{wr}} \sum_{i: X_i \geq c} \mathbb{K}\left(\frac{X_i - c}{h} \right) \left(W_i - a_{wr} - b_{wr} \left(X_i - c \right) \right)^2,$$

respectively. The kernel functions and bandwidths in (3) and (4) can be different. But to simplify the presentation we assume that they are identical.

Porter (2003) derived the asymptotic distributions of the nonparametric estimators $\hat{\theta}_s$ and $\hat{\theta}_f$. For example, under certain regularity conditions the asymptotic distribution of the estimator $\hat{\theta}_s$ using the local linear regressions in (3) is obtained as

$$\sqrt{nh}\left(\hat{\theta}_s - \theta_s \right) \xrightarrow{d} N\left(0, \frac{\sigma_l^2 + \sigma_r^2}{f(c)} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \right), \tag{5}$$

where $\sigma_l^2 = \lim_{x \uparrow c} \mathrm{Var}\left(Y_i | X_i = x \right)$, $\sigma_r^2 = \lim_{x \downarrow c} \mathrm{Var}\left(Y_i | X_i = x \right)$, $f(c)$ is the density function of X_i evaluated at c, $e_1 = (1, 0, \ldots, 0)'$, $\Gamma = \begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_2 \end{pmatrix}$, $\Delta = \begin{pmatrix} \delta_0 & \delta_1 \\ \delta_1 & \delta_2 \end{pmatrix}$, $\gamma_j = \int_0^\infty \mathbb{K}(z) z^j dz$, and $\delta_j = \int_0^\infty \mathbb{K}(z)^2 z^j dz$. The estimator $\hat{\theta}_f$ is also asymptotically normal with the asymptotic variance depending on σ_l^2, σ_r^2, $\lim_{x \uparrow c} \mathrm{Var}\left(W_i | X_i = x \right)$, $\lim_{x \downarrow c} \mathrm{Var}\left(W_i | X_i = x \right)$, $\lim_{x \uparrow c} \mathrm{Cov}\left(Y_i, W_i | X_i = x \right)$, $\lim_{x \downarrow c} \mathrm{Cov}\left(Y_i, W_i | X_i = x \right)$, and $f(c)$. The conventional Wald-type confidence sets for θ_s and θ_f are obtained by estimating these asymptotic variances of $\hat{\theta}_s$ and $\hat{\theta}_f$. Typically, we estimate the above nonparametric components by additional nonparametric regressions and plug those estimated components into the asymptotic variance formulae. The obtained Wald-type confidence set is symmetric around the estimator $\hat{\theta}_s$ or $\hat{\theta}_f$.

This paper proposes alternative confidence sets for the parameters θ_s and θ_f based on empirical likelihood, which circumvent the asymptotic variance estimation issues mentioned above and have data-determined shapes.

2.2 Empirical Likelihood for RDD

We now construct empirical likelihood functions for the average causal effect parameters θ_s and θ_f. We extend the empirical likelihood construction of Chen and Qin (2000) for local linear fitting to the sharp and fuzzy RDD contexts. Let $I_i = \mathbb{I}\{X_i \geq c\}$ be an indicator for whether the forcing variable X_i exceeds the cutoff level c. Note that $W_i = I_i$ in the sharp RDD, but $W_i \neq I_i$ in the fuzzy RDD.

We first consider the sharp RDD case. Observe that the local linear estimators $\hat{\alpha}_l$ and $\hat{\alpha}_r$ defined in (3) satisfy the first-order conditions (see Fan and Gijbels, 1996)

$$\sum_{i=1}^{n} (1 - I_i) K_{li} (Y_i - \hat{\alpha}_l) = 0, \quad \sum_{i=1}^{n} I_i K_{ri} (Y_i - \hat{\alpha}_r) = 0, \tag{6}$$

where

$$\begin{aligned} K_{li} &= \mathbb{K}\left(\frac{X_i - c}{h}\right) \left\{ \begin{array}{c} \frac{1}{nh} \sum_{i=1}^{n} (1 - I_i) \mathbb{K}\left(\frac{X_i - c}{h}\right) \left(\frac{X_i - c}{h}\right)^2 \\ - \left(\frac{X_i - c}{h}\right) \frac{1}{nh} \sum_{i=1}^{n} (1 - I_i) \mathbb{K}\left(\frac{X_i - c}{h}\right) \left(\frac{X_i - c}{h}\right) \end{array} \right\}, \\ K_{ri} &= \mathbb{K}\left(\frac{X_i - c}{h}\right) \left\{ \begin{array}{c} \frac{1}{nh} \sum_{i=1}^{n} I_i \mathbb{K}\left(\frac{X_i - c}{h}\right) \left(\frac{X_i - c}{h}\right)^2 \\ - \left(\frac{X_i - c}{h}\right) \frac{1}{nh} \sum_{i=1}^{n} I_i \mathbb{K}\left(\frac{X_i - c}{h}\right) \left(\frac{X_i - c}{h}\right) \end{array} \right\}. \end{aligned}$$

If we regard (6) as estimating equations for $\mathrm{E}[\hat{\alpha}_l]$ and $\mathrm{E}[\hat{\alpha}_r]$, the empirical likelihood function for $(\mathrm{E}[\hat{\alpha}_r] - \mathrm{E}[\hat{\alpha}_l], \mathrm{E}[\hat{\alpha}_l])$ is defined as

$$L_s(t, a) = \sup_{\{p_i\}_{i=1}^{n}} \prod_{i=1}^{n} p_i, \tag{7}$$

$$\text{s.t. } 0 \leq p_i \leq 1, \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i (1 - I_i) K_{li} (Y_i - a) = 0, \ \sum_{i=1}^{n} p_i I_i K_{ri} (Y_i - t - a) = 0.$$

Also, the log empirical likelihood ratio is defined as $\ell_s(t, a) = -2\{\log L_s(t, a) + n \log n\}$. By applying the Lagrange multiplier method, under mild conditions (see Theorem 2.2 in Newey and Smith, 2004), we can use the dual problem in place of (7). The dual form for $\ell_s(t, a)$ is

$$\ell_s(t, a) = 2 \sup_{\lambda \in \Lambda_n(t,a)} \sum_{i=1}^{n} \log\left(1 + \lambda' g_i(t, a)\right), \tag{8}$$

where $\Lambda_n(t, a) = \left\{\lambda \in \mathbb{R}^2 : \lambda' g_i(t, a) \in V \text{ for } i = 1, \ldots, n\right\}$, V is an open interval containing 0, and

$$g_i(t, a) = [(1 - I_i) K_{li} (Y_i - a), I_i K_{ri} (Y_i - t - a)]'. \tag{9}$$

Also, after profiling out the nuisance parameter a, the concentrated empirical likelihood ratio for $\mathrm{E}\left[\hat{\alpha}_r\right]-\mathrm{E}\left[\hat{\alpha}_l\right]$ is defined as

$$\ell_s\left(t\right)=\min_{a\in\mathcal{A}}\ell_s\left(t,a\right), \tag{10}$$

where $\mathcal{A}$ is a parameter space of α_l.

In practice, we use the dual representations in (8) and (10) to implement empirical likelihood inference. Note that (i) the optimization problem for the Lagrange multiplier λ in (8) is two-dimensional, and (ii) the objective function $\sum_{i=1}^{n}\log\left(1+\lambda' g_i\left(t,a\right)\right)$ for λ is typically concave in λ. Therefore, the computational cost to evaluate the empirical likelihood ratio $\ell_s\left(t,a\right)$ is not expensive.

The above construction gives us the empirical likelihood ratios for $\mathrm{E}\left[\hat{\alpha}_r\right]-\mathrm{E}\left[\hat{\alpha}_l\right]$ and $\mathrm{E}\left[\hat{\alpha}_l\right]$, rather than for $\theta_s=\alpha_r-\alpha_l$ and α_l. However, if we choose a relatively fast decay rate for the bandwidth h (i.e. undersmoothing), the bias components $\theta_s-\left(\mathrm{E}\left[\hat{\alpha}_r\right]-\mathrm{E}\left[\hat{\alpha}_l\right]\right)$ and $\alpha_l-\mathrm{E}\left[\hat{\alpha}_l\right]$ become asymptotically negligible. Therefore, the functions (8) and (10) can be employed as valid empirical likelihood ratios for the parameters θ_s and α_l.

We next consider the fuzzy RDD case. Similar to (7), we consider the following likelihood maximization problem:

$$L_f\left(t,a,a_{wl},a_{wr}\right)=\max_{\{p_i\}_{i=1}^{n}}\prod_{i=1}^{n}p_i, \tag{11}$$

$$\begin{aligned}\text{s.t.}\quad & 0\leq p_i\leq 1,\ \sum_{i=1}^{n}p_i=1,\ \sum_{i=1}^{n}p_i\left(1-I_i\right)K_{li}\left(Y_i-a\right)=0,\ \sum_{i=1}^{n}p_iI_iK_{ri}\left(Y_i-t\left(a_{wr}-a_{wl}\right)-a\right)=0,\\ & \sum_{i=1}^{n}p_i\left(1-I_i\right)K_{li}\left(W_i-a_{wl}\right)=0,\ \sum_{i=1}^{n}p_iI_iK_{ri}\left(W_i-a_{wr}\right)=0.\end{aligned}$$

Note that the last two conditions come from the first-order conditions for the local linear estimators of α_{wl} and α_{wr}. The dual form of the empirical likelihood ratio for $\left(\theta_f,\alpha_l,\alpha_{wl},\alpha_{wr}\right)$ is written as

$$\begin{aligned}\ell_f\left(t,a,a_{wl},a_{wr}\right) &= -2\left\{\log L_f\left(t,a,a_{wl},a_{wr}\right)+n\log n\right\}\\ &= 2\sup_{\lambda\in\Lambda_n\left(t,a,a_{wl},a_{wr}\right)}\sum_{i=1}^{n}\log\left(1+\lambda' h_i\left(t,a,a_{wl},a_{wr}\right)\right),\end{aligned} \tag{12}$$

where $\Lambda_n\left(t,a,a_{wl},a_{wr}\right)=\left\{\lambda\in\mathbb{R}^4:\lambda' h_i\left(t,a,a_{wl},a_{wr}\right)\in V_h\text{ for }i=1,\dots,n\right\}$, V_h is an open interval containing 0, and

$$\begin{aligned}h_i\left(t,a,a_{wl},a_{wr}\right) &= \left[\left(1-I_i\right)K_{li}\left(Y_i-a\right),I_iK_{ri}\left(Y_i-t\left(a_{wr}-a_{wl}\right)-a\right),\right.\\ &\qquad \left.\left(1-I_i\right)K_{li}\left(W_i-a_{wl}\right),I_iK_{ri}\left(W_i-a_{wr}\right)\right]'.\end{aligned} \tag{13}$$

Also, the concentrated empirical likelihood ratio for θ_f is defined as

$$\ell_f\left(t\right)=\min_{\left(a,a_{wl},a_{wr}\right)\in\mathcal{A}\times[0,1]\times[0,1]}\ell_f\left(t,a,a_{wl},a_{wr}\right). \tag{14}$$

3 Asymptotic Properties

This section investigates asymptotic properties of the empirical likelihood ratios proposed in the last section and proposes asymptotically valid empirical likelihood confidence sets for the average causal effects θ_s and θ_f identified from the sharp and fuzzy RDDs.

First, we consider the empirical likelihood ratios $\ell_s(t,a)$ in (8) and $\ell_s(t)$ in (10) for the sharp RDD. We impose the following assumptions.

Assumption 3.1.

(i) $\{Y_i, W_i, X_i\}_{i=1}^n$ *is i.i.d.*

(ii) *There exists a neighborhood $\mathcal{N}$ around c such that (a) the density function f of X_i is continuously differentiable and bounded away from zero in $\mathcal{N}$, (b) $\mathrm{E}[Y_i | X_i = x] - \theta_s \mathbb{I}\{x \geq c\}$ is continuously differentiable in $\mathcal{N} \setminus \{c\}$ and is continuous at c with finite left and right hand derivatives, (c) $\mathrm{E}\left[Y_i^2 \middle| X_i = x\right]$ is continuous in $\mathcal{N} \setminus \{c\}$ and has finite left and right hand limits at c, and (d) $\mathrm{E}\left[|Y_i|^\zeta \middle| X_i = x\right]$ is uniformly bounded on $\mathcal{N}$ for some $\zeta \geq 4$. Also, V_l and V_r defined in (18) are positive.*

(iii) $\mathbb{K}$ *is a symmetric and bounded density function with support $[-k,k]$ for some $k \in (0,\infty)$.*

(iv) *As $n \to \infty$, $h \to 0$, $nh \to \infty$, $nh^5 \to 0$, and $n^{1/\zeta - 1/2} h^{-1/2} \to 0$.*

(v) *$\mathcal{A}$ is compact and $\alpha_l \in int(\mathcal{A})$.*

Assumption 3.1 (i) is on the data structure. Since RDD analysis is typically applied to cross section data, this assumption is reasonable. Assumption 3.1 (ii) restricts the local shape of the data distribution around $x = c$. Note that this assumption allows discontinuity of the conditional moments $\mathrm{E}[Y_i | X_i = x]$, $\mathrm{E}\left[Y_i^2 \middle| X_i = x\right]$, and $\mathrm{E}\left[|Y_i|^\zeta \middle| X_i = x\right]$ at $x = c$. Assumption 3.1 (iii) is on the kernel function $\mathbb{K}$ and imposes that we use a second-order kernel. Assumption 3.1 (iv) is on the bandwidth parameter h. If $h \propto n^{-\eta}$, this assumption is satisfied for $\eta \in \left(\frac{1}{5}, 1 - \frac{2}{\zeta}\right)$. The bandwidth h can be stochastic: in that case, we replace "$\to$" with "$\xrightarrow{p}$" in this assumption. The requirement $nh^5 \to 0$ corresponds to an undersmoothing condition to remove the bias components in the construction of empirical likelihood. See Section 4.1 for further discussion. Assumption 3.1 (v) is required for the concentrated empirical likelihood ratio $\ell_s(\theta_s)$.

Under these assumptions, we obtain the asymptotic distributions of the empirical likelihood ratios $\ell_s(\theta_s, \alpha_l)$ and $\ell_s(\theta_s)$.

Theorem 3.1.

(i) *Under Assumption 3.1 (i)-(iv), $\ell_s(\theta_s, \alpha_l) \xrightarrow{d} \chi^2(2)$.*

(ii) *Under Assumption 3.1, $\ell_s(\theta_s) \xrightarrow{d} \chi^2(1)$.*

See Appendix A.1 for a proof of this theorem. Theorem 3.1 says that the empirical likelihood ratios $\ell_s(\theta_s, \alpha_l)$ and $\ell_s(\theta_s)$ are asymptotically pivotal and converge to chi-square distributions, i.e. Wilks's phenomenon emerges in this nonparametric RDD context. This result can be compared with earlier works which have also demonstrated the Wilks's phenomenon for empirical likelihood in other nonparametric models, such as Chen and Qin (2000), Fan, Zhang and Zhang (2001), Xu (2009), and Chan, Peng and Zhang (2010). Intuitively, the moment restriction $E[g_i(\theta_s, \alpha_l)] \approx 0$ can be viewed as a "localized" moment restriction at $X_i = c$ with an effective sample size nh, instead of n for standard moment restrictions. By undersmoothing, we can neglect the bias in $E[g_i(\theta_s, \alpha_l)]$ from 0, and an adaptation of a standard argument from the empirical likelihood literature for standard moment restrictions implies Wilks's phenomenon in our nonparametric context. Also, based on Theorem 3.1 (ii), the $100(1-\xi)\,\%$ asymptotic empirical likelihood confidence set for the average causal effect parameter θ_s is obtained as

$$ELCS_{s,\xi} = \left\{t : \ell_s(t) \leq \chi^2_{1-\xi}(1)\right\},$$

where $\chi^2_{1-\xi}(1)$ is the $100(1-\xi)\,\%$ critical value for the $\chi^2(1)$ distribution.

We now compare with the conventional Wald-type confidence set

$$WCS_{s,\xi} = \left[\hat{\theta}_s \pm z_{1-\xi/2}\sqrt{\widehat{Asy.Var}\left(\hat{\theta}_s\right)}\right],$$

where $z_{1-\xi/2}$ is the $100(1-\xi/2)\,\%$ standard normal critical value and $\widehat{Asy.Var}\left(\hat{\theta}_s\right)$ is some (typically nonparametric) estimator of the asymptotic variance of $\hat{\theta}_s$ presented in (5). There are at least four important differences. First, the empirical likelihood confidence set does not require the variance estimator $\widehat{Asy.Var}\left(\hat{\theta}_s\right)$, which typically requires additional nonparametric estimation for σ_l^2, σ_r^2, and $f(c)$. In Section 4.2, we argue that in some special case this circumvention of variance estimation can yield a better higher-order coverage property for the empirical likelihood confidence set. Second, the empirical likelihood confidence set is not necessarily symmetric around the point estimator $\hat{\theta}_s$: the shape of the confidence set is determined by that of the empirical likelihood function. Intuitively, the Wald-type confidence set is derived from a quadratic approximation of some criterion function to obtain $\hat{\theta}_s$. The empirical likelihood confidence set is derived directly from the empirical likelihood function without relying on such a quadratic approximation. Third, in finite samples the empirical likelihood confidence set may not be an interval (it could be disjoint or unbounded) but the Wald-type confidence set is always an interval. At first glance, this feature might seem like a drawback to the empirical likelihood approach. However, as Stock and Wright (2000) argued in a GMM context, disjoint or unbounded confidence sets can be viewed as a symptom of weak identification, in which case the GMM or (negative) empirical likelihood criterion function tends to be flat or wiggly around the bottom. Under weak identification, it is known that the Wald-type confidence set can yield highly misleading conclusions (Stock and Wright, 2000). See also Lemieux and Marmer (2009) for a discussion of the weak identification problem in a fuzzy RDD context. Although formal analysis on weak identification in our setup is beyond the scope of this paper, it is at least beneficial to use the empirical likelihood confidence set as a complement to the Wald-type one. Finally, although the empirical likelihood confidence set circumvents asymptotic variance estimation, it requires numerical search to find endpoints for the confidence set satisfying $\ell_s(t) = \chi^2_{1-\xi}(1)$, so it is more computationally expensive than the Wald-type confidence set. Based on these differences, we recommend the empirical likelihood confidence set as a complement to the conventional Wald-type confidence set.

Next, we consider the empirical likelihood ratios $\ell_f(t, a, a_{wl}, a_{wr})$ in (12) and $\ell_f(t)$ in (14) for the fuzzy RDD. For this case, we add the following assumption.

Assumption 3.2.

There exists a neighborhood $\mathcal{N}'$ around c such that $\mathrm{E}[W_i | X_i = x] - (\alpha_{wr} - \alpha_{wl}) \mathbb{I}\{x \geq c\}$ is continuously differentiable in $\mathcal{N}' \setminus \{c\}$ and is continuous at c with finite left and right hand derivatives. Also, $\alpha_{wl}, \alpha_{wr} \in (0, 1)$.

This assumption corresponds to Assumption 3.1 (ii) in the sharp RDD case. The asymptotic properties of the empirical likelihood ratios $\ell_f(\theta_s, \alpha_l, \alpha_{wl}, \alpha_{wr})$ and $\ell_f(\theta_f)$ are presented as follows.

Theorem 3.2.

(i) *Under Assumptions 3.1 (i)-(iv) and 3.2, $\ell_f(\theta_s, \alpha_l, \alpha_{wl}, \alpha_{wr}) \xrightarrow{d} \chi^2(4)$.*

(ii) *Under Assumptions 3.1 and 3.2, $\ell_f(\theta_f) \xrightarrow{d} \chi^2(1)$.*

Since the proof is similar to that of Theorem 3.1, it is omitted. Based on Theorem 3.2 (ii), the $100(1-\xi)\%$ empirical likelihood confidence set for the average causal effect parameter θ_f is

$$ELCS_{f,\xi} = \left\{t : \ell_f(t) \leq \chi^2_{1-\xi}(1)\right\}.$$

Similar comments to Theorem 3.1 apply here. However, we mention that the asymptotic variance of $\hat{\theta}_f$ is more complicated than that of $\hat{\theta}_s$. In addition to σ_l^2, σ_r^2, and $f(c)$, the asymptotic variance of $\hat{\theta}_f$ contains four more nonparametric components: $\lim_{x\uparrow c} \mathrm{Var}(W_i | X_i = x)$, $\lim_{x\downarrow c} \mathrm{Var}(W_i | X_i = x)$, $\lim_{x\uparrow c} \mathrm{Cov}(Y_i, W_i | X_i = x)$, and $\lim_{x\downarrow c} \mathrm{Cov}(Y_i, W_i | X_i = x)$. Also, the Wald-type confidence set relies upon a linear approximation (or delta method) to the ratio $\hat{\theta}_f = \frac{\hat{\alpha}_r - \hat{\alpha}_l}{\hat{\alpha}_{wr} - \hat{\alpha}_{wl}}$.

4 Discussion

4.1 Bandwidth Selection

To implement our empirical likelihood inference, we need to choose the bandwidth h. One way to select the bandwidth is to conduct a higher-order expansion, derive some Edgeworth expansion formula for the coverage probability (say, $\Pr\{ELCS_{s,\xi}\} = 1 - \xi + r(n, h)$ with $r(n, h) \to 0$ as $n \to \infty$ for the sharp RDD case), and then choose h to minimize the dominant term of the coverage error $r(n, h)$.

This approach was adopted by Chen and Qin (2000) for their empirical likelihood confidence interval of the conditional mean. Our setup is more complicated than that of Chen and Qin (2000) due to the existence of more than one moment restriction and additional profile-out steps needed to obtain $\ell_s(\theta_s)$ and $\ell_f(\theta_f)$. Thus, we leave this analysis for future research.

An alternative would be to adopt some bandwidth selection procedure that is effective for point estimation of nonparametric regression functions. Although our interest is on interval estimation or hypothesis testing for θ_s or θ_f, desirable properties for point estimation can reflect favorably on the performance of the empirical likelihood-based inference. For local linear nonparametric regression, Li and Racine (2004) studied data-driven cross-validation methods under a general setup and presented desirable theoretical and simulation evidence. However, there are two difficulties that prevent us from applying Li and Racine's (2004) results to our context. First, the results of Li and Racine (2004) are not directly applicable because we need to choose the bandwidths to estimate the regression functions at the boundary points, such as $\lim_{x\downarrow c} \mathrm{E}[Y_i|X_i = x]$ and $\lim_{x\uparrow c} \mathrm{E}[Y_i|X_i = x]$. Second, to obtain the limiting χ^2 null distributions for the empirical likelihood ratios in Theorems 3.1 and 3.2, we need to undersmooth the bandwidth to satisfy $nh^5 \to 0$ (Assumption 3.1 (iv)), which excludes Li and Racine's (2004) convergence rate $O_p\left(n^{-1/5}\right)$ for their least square cross-validation bandwidth. If we allow $nh^5 \to c$ for some constant c, modified arguments imply the limiting non-central χ^2 null distributions for the empirical likelihood ratios, where the non-centrality parameters depend on c. Although full investigation of these issues is reserved for future work, we suggest the following modified cross-validation bandwidth selection method motivated by Li and Racine (2004): (i) choose the bandwidths for the local linear regressions in (3) and (4) by the cross-validation method discussed in Li and Racine (2004), and then (ii) modify those cross-validated bandwidths by multiplying $n^{-\epsilon}$ (say, $\epsilon = 0.1$) for undersmoothing. Also, as suggested by Imbens and Lemieux (2008), one may implement this procedure for observations which are close enough to the cutoff point (i.e. observations with $|X_i - c| \leq \delta$ for some given $\delta > 0$).

4.2 Higher-order Properties

We present some intuition for why empirical likelihood confidence sets can be theoretically better than Wald-type confidence sets. Consider the sharp RDD case and assume that the right limit $\alpha_r = \lim_{x\downarrow c} \mathrm{E}[Y_i|X_i = x]$ is known. In this case we can concentrate on the inference problem for the left limit $\alpha_l = \lim_{x\uparrow c} \mathrm{E}[Y_i|X_i = x]$. The empirical likelihood ratio for α_l can be written as

$$\tilde{\ell}_s(a) = 2 \sup_{\lambda \in \Lambda_n(a)} \sum_{i:X_i \geq c} \log\left(1 + \lambda \tilde{g}_i(a)\right),$$

where $\Lambda_n(a) = \left\{\lambda \in \mathbb{R} : \lambda \tilde{g}_i(a) \in \tilde{V} \text{ for all } i \text{ with } X_i > c\right\}$, $\tilde{V}$ is an open interval containing 0, and $\tilde{g}_i(t,a) = K_{li}(Y_i - a)$. The same argument to Theorem 3.1 yields $\tilde{\ell}_s(\alpha_l) \xrightarrow{d} \chi^2(1)$, and the empirical likelihood confidence set for α_l is defined as $\widetilde{ELCS}_{s,\xi} = \left\{a : \tilde{\ell}_s(a) \leq \chi^2_{1-\xi}(1)\right\}$. On the other hand, the Wald-type confidence set for α_l based on the local linear estimator $\hat{\alpha}_l$ from the first equation of

(3) is defined as $\widetilde{WCS}_{s,\xi} = \left[\hat{\alpha}_l \pm z_{1-\xi/2}\sqrt{\widehat{Asy.Var}\left(\hat{\alpha}_l\right)}\right]$, where $\widehat{Asy.Var}\left(\hat{\alpha}_l\right)$ is some nonparametric estimator for the asymptotic variance of $\hat{\alpha}_l$. Under this setup with additional regularity conditions, we can directly apply the results of Chen and Qin (2000). Chen and Qin (2000) found that even though both $\widetilde{ELCS}_{s,\xi}$ and $\widetilde{WCS}_{s,\xi}$ are derived from the local linear regression problem, their coverage errors for α_l have different orders, i.e.

$$\Pr\left\{\alpha_l \in \widetilde{ELCS}_{s,\xi}\right\} = 1 - \xi + O\left(nh^5 + h^2 + (nh)^{-1}\right),$$
$$\Pr\left\{\alpha_l \in \widetilde{WCS}_{s,\xi}\right\} = 1 - \xi + O\left(nh^5 + h + (nh)^{-1}\right).$$

For example, if $h = O\left(n^{-1/3}\right)$ (which satisfies Assumption 3.1 (iv)), then the coverage error of $\widetilde{ELCS}_{s,\xi}$ is $O\left(n^{-2/3}\right)$ but the coverage error of $\widetilde{WCS}_{s,\xi}$ is $O\left(n^{-1/3}\right)$. As Chen and Qin (2000) argued, this higher-order difference emerges from the fact that the coverage error of $\widetilde{WCS}_{s,\xi}$ depends on the estimation error of the asymptotic variance of $\hat{\alpha}_l$. Since the empirical likelihood confidence interval is free from such an estimation error, $\widetilde{ELCS}_{s,\xi}$ yields a better higher-order coverage property than $\widetilde{WCS}_{s,\xi}$.[3]

The empirical likelihood ratios presented in Section 2 are more complicated because of additional moment functions and profile-out manipulations. Therefore, formal higher-order analysis is beyond the scope of the paper. However, it is reasonable to conjecture that similar arguments to Chen and Qin (2000) will yield analogous higher-order properties.

4.3 Extensions

In this section we discuss two extensions of the present results: the inclusion of additional covariates and parametric functional forms.

It is often the case that we need to incorporate additional covariates to RDD analysis. We first consider the sharp RDD. Suppose there are m additional covariates Z_i. Then α_l and α_r are estimated by $\bar{\alpha}_l$ and $\bar{\alpha}_r$, which are solutions of the weighted least square problems with respect to a_l and a_r,

$$\min_{a_l,b_l,d_l} \sum_{i:X_i<c} \mathbb{K}\left(\frac{X_i-c}{h}\right)\left(Y_i - a_l - b_l\left(X_i-c\right) - d_l' Z_i\right)^2,$$
$$\min_{a_r,b_r,d_r} \sum_{i:X_i\geq c} \mathbb{K}\left(\frac{X_i-c}{h}\right)\left(Y_i - a_r - b_r\left(X_i-c\right) - d_r' Z_i\right)^2,$$

respectively. Solving the minimization problems gives $\bar{\alpha}_l = \sum_{i:X_i<c}\pi_{li}Y_i$, where $\pi_{li} = e_1' A_l^{-1}\mathbb{K}\left(\frac{X_i-c}{h}\right)\left[1, X_i-c, Z_i'\right]'$ with $e_1 = (1,0,\cdots,0)' \in \mathbb{R}^{m+2}$ and

$$A_l = \sum_{i:X_i<c}\mathbb{K}\left(\frac{X_i-c}{h}\right)\begin{pmatrix} 1 & X_i-c & Z_i' \\ X_i-c & (X_i-c)^2 & Z_i'\left(X_i-c\right) \\ Z_i & Z_i\left(X_i-c\right) & Z_iZ_i' \end{pmatrix},$$

[3] Chen and Qin (2000) also proposed Bartlett correction for $\widetilde{ELCS}_{s,\xi}$, which provides even smaller coverage errors.

and $\bar{\alpha}_r = \sum_{i:X_i \geq c} \pi_{ri} Y_i$, where A_r (therefore π_{ri}) is similarly defined with the summation taken for the observations with $X_i \geq c$.[4] Since it can be shown that $\sum_{i:X_i<c} \pi_{li} = \sum_{i:X_i \geq c} \pi_{ri} = 1$, the estimating equations for α_l and α_r can be written as

$$\sum_{i=1}^{n} (1 - I_i) \pi_{li} (Y_i - \bar{\alpha}_l) = 0, \quad \sum_{i=1}^{n} I_i \pi_{ri} (Y_i - \bar{\alpha}_r) = 0.$$

The empirical likelihood ratio can be obtained by replacing the estimating functions $g_i(t, a)$ in (9) with

$$g_i(t, a) = [(1 - I_i) \pi_{li} (Y_i - a), I_i \pi_{ri} (Y_i - t - a)]'.$$

For the fuzzy RDD case, the estimating functions $h_i(t, a, a_{wl}, a_{wr})$ in (13) are correspondingly modified as

$$\begin{aligned} h_i(t, a, a_{wl}, a_{wr}) &= [(1 - I_i) \pi_{li} (Y_i - a), I_i \pi_{ri} (Y_i - t (a_{wr} - a_{wl}) - a), \\ &\quad (1 - I_i) \pi_{li} (W_i - a_{wl}), I_i \pi_{ri} (W_i - a_{wr})]'. \end{aligned}$$

The concentrated empirical likelihood ratios $\ell_s(\theta_s)$ and $\ell_f(\theta_f)$ are similarly defined and they are asymptotically chi-square distributed at the true parameter values.

In the previous sections, we do not impose any parametric functional form on the conditional mean $\mathrm{E}[Y_i | X_i = x]$ and conditional treatment probability $\Pr\{W_i = 1 | X_i = x\}$. The empirical likelihood approach can naturally accommodate parametric functional forms. For example, in a fuzzy RDD with the cutoff value $c = 0$, one specifies the regression model for Y_i as

$$Y_i = \beta_0 + \beta_1 I_i + \beta_l' P_l(X_i)(1 - I_i) + \beta_r' P_r(X_i) I_i + u_i, \quad \mathrm{E}[u_i | X_i, I_i] = 0, \tag{15}$$

where $P_l(X_i)$ and $P_r(X_i)$ are finite dimensional vectors of polynomials of X_i without constant terms. This specification allows the regression functions to have different left and right limits at the threshold $x = c = 0$. In this case, the numerator of θ_f in (2) is identified by $\lim_{x \downarrow 0} \mathrm{E}[Y_i | X_i = x] - \lim_{x \uparrow 0} \mathrm{E}[Y_i | X_i = x] = \beta_1$. The model (15) can be estimated by the two stage least squares with instrumental variables V_i, for example. Typical candidates for V_i are the indicator variable I_i and polynomials of X_i. To incorporate parametric information in (15), we can modify the estimating function $h_i(t, a, a_{wl}, a_{wr})$ in (13) as follows:

$$\begin{aligned} h_i(t, b_0, b_l, b_r, a_{wl}, a_{wr}) &= \left[V_i' \left(Y_i - b_0 - t (a_{wr} - a_{wl}) I_i - b_l' P_l(X_i)(1 - I_i) - b_r' P_r(X_i) I_i\right), \right. \\ &\quad \left. (1 - I_i) K_{li} (W_i - a_{wl}), I_i K_{ri} (W_i - a_{wr})\right]'. \end{aligned}$$

The empirical likelihood ratios and their asymptotic chi-square null distributions can be obtained under analogous conditions. By applying the same argument, it is also possible to incorporate parametric information on the conditional treatment probability $\Pr\{W_i = 1 | X_i = x\}$, such as the logit or probit functional form.

[4] Note that if there are no additional covariates, π_{li} and π_{ri} reduce to K_{li} and K_{ri}, respectively.

5 Numerical Examples

In this section we study the finite sample performance of the proposed empirical likelihood methods through simulations and an empirical application, and compare with the conventional Wald or t-test based on the asymptotic normality of the average causal effect estimators $\hat{\theta}_s$ and $\hat{\theta}_f$.

5.1 Simulations

We consider the following data generating process of the sharp RDD:

$$Y_i = \mu(X_i) + \theta_s W_i + \sigma(X_i)\varepsilon_i, \tag{16}$$

where $\mu(x) = x^2$, $W_i = \mathbb{I}\{X_i \geq c\}$, $X_i \sim iid\ Uniform[-2,2]$, $\varepsilon_i \sim iid\ N(0,1)$, and

$$\sigma(x) = 2.5\exp(-|x|)\mathbb{I}\{x \geq c\} + \sqrt{1.4}(1 - \mathbb{I}\{x \geq c\}). \tag{17}$$

The cutoff point is set to $c = 0.5$ so that the conditional mean $\mathrm{E}[Y_i|X_i = x]$ jumps at $x = 0.5$ from $\alpha_l = 0.25$ to $\alpha_r = 3.25$. Thus, in this setup, the average causal effect is $\theta_s = \alpha_r - \alpha_l = 3$. The conditional variance function $\mathrm{Var}(Y_i|X_i = x) = \sigma^2(x)$ is homoskedastic for $x < c$ and heteroskedastic for $x \geq c$. This specification of $\sigma^2(x)$ is adopted to assess the impact of heteroskedasticity. A representative sample with 100 observations is displayed in Figure 1 (a).

We consider two kinds of t-tests based on different estimators for the asymptotic variance of $\hat{\theta}_s$: (i) Porter's (2003) residual-based kernel estimator of the variance function on boundaries (denoted as AN1), and (ii) its improved version based on local linear estimators of the variance function as in Ruppert *et al.* (1997) and Fan and Yao (1998) (denoted as AN2).[5] We compare these t-tests for the null hypothesis $H_0 : \theta_s = 3$ with the empirical likelihood test (denoted as EL) introduced in this paper.

To implement these tests, we need to choose the kernel function $\mathbb{K}$ and bandwidth h. In our experiments, we use the Epanechnikov kernel function $\mathbb{K}(z) = \frac{3}{4}(1 - z^2)\mathbb{I}\{|z| \leq 1\}$ and six fixed bandwidths ranging from $h = 0.8$ to $h = 1.3$ when the sample size is 100 and from $h = 0.7$ to $h = 1.2$ when the sample size is 200. We also consider a data-dependent bandwidth selected via least square cross-validation, in which we discard 50% of the observations on each side far from the cutoff value, as recommended by Imbens and Lemieux (2008, Section 5.1). Figure 1 (b) plots the distribution (over replications) of the data-driven bandwidths selected for the two sample sizes.

Tables 1 and 2 report the rejection rates of the two t-tests (AN1 and AN2) and the empirical likelihood test (EL) over 1000 replications with the nominal sizes 5% and 10%, when the sample sizes are 100 and 200, respectively. In addition, we report the averages and standard errors (over replications)

[5] Local linear fitting is generally preferred in estimating nonparametric functions at boundary points because of automatic boundary bias correction. But in finite samples the local linear fitting may give negative estimates of variances occasionally (see e.g. Xu and Phillips, 2009). In our simulations, the percentages of negative local linear estimates of $\hat{\sigma}_r^2(c)$ or $\hat{\sigma}_l^2(c)$ over replications range from 5.8% to 0.8% for six bandwidths considered when $n = 100$, and from 1.1% to 0.1% when $n = 200$. But we did not observe negative estimates for $\hat{\sigma}_r^2(c) + \hat{\sigma}_l^2(c)$.

of the estimates $\hat{\alpha}_r$ and $\hat{\alpha}_l$ of the right and left limits of the conditional mean, and those of the estimates $\hat{\sigma}_r^2(c)$ and $\hat{\sigma}_l^2(c)$ of the right and left limits of the conditional variance. We also record the averages and variances (over replications) of the estimate $\hat{\theta}_s$ in the columns labeled as "$\hat{\theta}_s$" and "$var\left(\hat{\theta}_s\right)$". The column labeled as "$\widehat{var\left(\hat{\theta}_s\right)}$" gives the averages and standard errors (over replications) of the estimated asymptotic variances, where $\sigma_r^2(c)$ and $\sigma_l^2(c)$ are estimated by the kernel (AN1) or the local linear method (AN2). It should be compared with $var\left(\hat{\theta}_s\right)$, the true value of the asymptotic variance of $\hat{\theta}_s$ presented in (5).

Several observations are in order. The three tests (AN1, AN2, and EL) for $H_0 : \theta_s = 3$ are generally oversized. Over all bandwidths considered including the data-driven one, EL appears to have the least amount of size distortion among the three tests. Using the cross-validated bandwidth does not help much to reduce size distortions. When the larger sample size is used, the empirical sizes of the three tests are closer to the nominal ones, with the largest improvement observed for the EL test. Noticeable biases are observed for $\hat{\alpha}_r$ and $\hat{\alpha}_l$, especially when large bandwidths are used. On the other hand, these estimates happen to be biased in the same direction so that the bias of their difference $\hat{\theta}_s = \hat{\alpha}_r - \hat{\alpha}_l$ is negligible. The variance of $\hat{\theta}_s$ is quite close to the sum of the variances of $\hat{\alpha}_r$ and $\hat{\alpha}_l$. Marked size distortions of AN1 are largely explained by the fact that the variance of $\hat{\theta}_s$ is poorly estimated when $\hat{\sigma}_r^2(c)$ and $\hat{\sigma}_l^2(c)$ are estimated using the kernel method. In particular, $\hat{\sigma}_r^2(c)$ is seriously biased, with the average (over replications) just about half of the true value of $\sigma_r^2(c)$. On the other hand, $\sigma_l^2(c)$ appears to be estimated satisfactorily. Take the case when $n = 100$ and $h = 1.0$ for example. The average (over replications) of $\hat{\sigma}_r^2(c)$ is 1.17 with standard error 0.48, which is far below the true value $\sigma_r^2(c) = 2.3$, while the average (over replications) of $\hat{\sigma}_l^2(c)$ is 1.38 with standard error 0.47, which is fairly close to the true value $\sigma_l^2(c) = 1.4$. Consequently, the average (over replications) of the estimated asymptotic variances of $\hat{\theta}_s$ is 0.46 with standard error 0.13, which underestimates the true value $var\left(\hat{\theta}_s\right) = 0.63$. This explains the serious over-rejection of AN1. Similar comments apply for other bandwidths and for the case of $n = 200$. This is not surprising in view of our design of the variance function (with significant non-zero derivatives on the right side but zero derivatives of any order on the left side). In contrast, the estimates of $var\left(\hat{\theta}_s\right)$ are considerably improved when we use the local linear estimators for $\sigma_r^2(c)$ and $\sigma_l^2(c)$ (still with appreciable downward bias for $\hat{\sigma}_r^2(c)$). This is consistent with the better size property of AN2 compared to that of AN1.[6]

[6] The performance of the t-tests AN1 and AN2 can be alternatively improved by using the standard error estimated via bootstrap. To be concrete, generate B bootstrap samples by resampling the pairs (X_i, Y_i) and for each bootstrap sample we obtain the estimate of θ_s, denoted by $\hat{\theta}_s^*(b)$, where $b = 1, \ldots, B$. Define the test statistic $\left(\hat{\theta}_s - \theta_s\right)/se^*\left(\hat{\theta}_s\right)$, where $se^*\left(\hat{\theta}_s\right)$ is the standard error of the estimates $\hat{\theta}_s^*(b)$ over B bootstrap replications. Although this test statistic avoids nonparametric regressions to estimate the asymptotic variance of $\hat{\theta}_s$, it is computationally more expensive. In our experiments, the bootstrap test takes about ten times longer than the EL test if the number of bootstrap replications is $B = 399$. Our preliminary simulation results (not reported here) show that (i) the bootstrap method has smaller estimation errors for $var\left(\hat{\theta}_s\right)$ than those of AN1 and AN2; and (ii) the bootstrap test shows similar size properties to the EL test. To our best knowledge, there is no theoretical study on bootstrap methods in the RDD context and further

The P-value plots (Davidson and MacKinnon, 1998) displayed in Figure 2 compare the actual null rejection rates of each of the two squared t-tests and the EL test with a range of nominal null rejection rates from 0.2%-25%, when $(n, h) = (100, 1.0)$ and $(200, 0.9)$. The P-value discrepancy plots (Figure 3) show the differences of actual and nominal null rejection rates. These plots are useful to evaluate the quality of asymptotic approximations for the test statistics in finite samples. It is clear from these figures that all p-values of the EL test are closer to the nominal null rejection rates than those of the t-tests. This means that the $\chi^2(1)$ distribution serves as a better approximation for the finite sample distribution of the EL test statistic than that of the two (squared) t-test statistics. Similar results are obtained for other bandwidths.

Figures 4 and 5 show the calibrated powers of the three tests under the alternative $H_A : \theta_s = \theta_A$. These calibrated powers are computed by using adjusted critical values (see Table 3) at which the null rejection rates are 10% under the data generating process in (16).[7] We observe that all tests are more powerful when a larger bandwidth is used. AN1 and AN2 generally have similar power properties except that AN2 is less powerful for small bandwidths due to the relatively higher variability of the local linear variance estimates. It is clear from the figures that EL has dominant power for all bandwidth values except when the value of θ_A is on the far right side of the null hypothesis. This exception disappears when the sample size is 200. In this case, for all values of θ_A, EL has the highest power among all tests considered.

Overall, our simulation result suggests that the empirical likelihood method is very promising because the resulting test has better size and power properties than the conventional Wald or t-tests.

5.2 Empirical Application

We use the data of Angrist and Lavy (1999) to study the effect of the number of classes on pupils' scholastic achievement. In Israeli public schools, Maimonides's rule, which stipulates that a class should be split when it has more than 40 students, has been used to determine the division of enrollment cohorts into classes. Here we only consider schools which have one or two classes and focus on 4th graders, although Angrist and Lavy's original analysis involved schools with up to six classes and studied 3rd, 4th, and 5th graders. We end up with a sample with 1177 observations (after removing 2 observations with missing values), with 307 schools having only one class (the controlled group) and 870 schools having two classes (the treated group).

Plots of average math scores and verbal scores (outcome variables) against enrollment sizes (forcing variable) are displayed in Figures 7 and 8, respectively. The round circles represent the controlled group and the pentagrams represent the treated group. Actual class size may not be the same as what would be predicted by a strict application of Maimonides's rule. It is clear from the figures that there are

research is needed.

[7] The calibrated powers we report here are often misnamed as "size-adjusted" powers in the literature, as pointed out by Horowitz and Savin (2000) and Davidson and MacKinnon (2006). Following their arguments, we emphasize that such calibrated critical values are only useful in our specific Monte Carlo studies and are irrelevant to empirical applications.

schools with enrollments near the cutoff point 40 appearing both in the treated and controlled groups. In other words, this is an fuzzy RDD. Local linear fits are also plotted for the two groups. We use the bandwidth $h = 10$ for illustration, which is close to the one selected via least square cross-validation. The jump size for the average verbal scores seems to be larger than that for the average math scores. The local linear estimate of the propensity score function (i.e. $\Pr\{W_i = 1 | X_i = x\}$) is plotted in Figure 6 with treatment assignments (jiggled with small random noises so that overlapped observations are distinguishable). A discontinuity at the enrollment count $c = 40$ is clearly visible.

We construct confidence sets for the average causal effect θ_f in (2) for the fuzzy RDD by the Wald test (AN CSs) and the empirical likelihood test (EL CSs) with confidence level 90%. Figure 9 (a) presents the estimates and confidence sets for the discontinuity size in the propensity score function (i.e. $\alpha_{wr} - \alpha_{wl}$), which can be obtained by applying our method for the sharp RDD to the dependent variable W_i. The estimates of $\alpha_{wr} - \alpha_{wl}$ are between 0.54 and 0.70 and the EL CSs for $\alpha_{wr} - \alpha_{wl}$ are wider than the AN CSs in both the lower and upper tails. Figures 9 (b) and 10 present the AN and EL CSs together with the local linear point estimates using a group of bandwidths for the math score and the verbal score, respectively. Depending on the choice of the bandwidth, the estimate for the average causal effect θ_f ranges from 1.8 to 7.4 for the math score and from 5.0 to 12.0 for the verbal score. The AN CSs are symmetric around the point estimates by construction. In contrast, the EL CSs are typically asymmetric around the point estimates and wider than the AN CSs. This result is consistent with the simulation evidence in Section 5.1 that the AN CSs are potentially subject to under-coverage (or over-rejection). For both the math and verbal scores, the AN and EL CSs have similar lower endpoints. On the other hand, these two CSs yield rather different upper endpoints. For example, if we take $h = 10$, which is close to the one selected via least square cross-validation, the upper endpoints of the AN and EL CSs for the verbal scores are considerably different: around 19 and 30, respectively. This contrast suggests that compared to the lower endpoints, we may not have enough sample information to determine the upper endpoints of the confidence set for θ_f.

For further graphical illustration, in Figures 11 and 12 we plot the values of the Wald and EL test statistics for a range of candidate parameter values for the jump in the propensity score and the causal effect. The critical values at different confidence levels are also marked. These plots show how the empirical likelihood confidence sets are constructed via inversion of the test statistics. Also they show how the EL CIs are asymmetric around the point estimates. Both AN and EL CSs show that splitting a large class into two small classes has a significant impact to improve the pupils' verbal scores, but not to improve their math scores. Also, from Figure 12, we can see that the empirical likelihood function is relatively flat for the right tail. This result indicates that we may not have strong sample information to determine the upper endpoint of the confidence set of θ_f. Note that the Wald approach never provides such additional information. This difference demonstrates that the empirical likelihood approach can provide useful information in practice that is not available by the conventional Wald approach. In practice, the Wald approach tends to yield too small confidence sets. On the other hand, the empirical likelihood approach tends to yield relatively larger confidence sets. Thus, the researcher can feel confident in her results if she obtains the same conclusion from both approaches (e.g. $\theta_f > 0$ in the verbal score example). Meanwhile, if she obtains different conclusions from these approaches (e.g. $\theta_f > 15$ in the verbal score example with the 5% significance level), she needs to be cautious about whether she has enough sample information to extract a definitive conclusion.

6 Conclusion

This paper proposes empirical likelihood inference methods for average causal effects in regression discontinuity designs. Our methods allow for sharp and fuzzy regression discontinuity designs and do not need to specify parametric functional forms on the regression functions. Compared to the conventional Wald-type confidence sets, our empirical likelihood confidence sets do not require asymptotic variance estimation and can be asymmetric around the point estimates. Monte Carlo simulations and an empirical example evaluating the effect of class size on pupils' performance are used to illustrate the benefits of the proposed methods.

A Mathematical Appendix

In the appendix, we provide mathematical proofs of the main results. Define

$$\begin{aligned} s_{l,j_1j_2} &= f(c)\int_{-k}^{0}\mathbb{K}(z)^{j_1} z^{j_2}dz, \quad s_{r,j_1j_2} = f(c)\int_{0}^{k}\mathbb{K}(z)^{j_1} z^{j_2}dz, \\ V_l &= \sigma_l^2\left(s_{l,12}^2 s_{l,20} - 2s_{l,12}s_{l,11}s_{l,21} + s_{l,11}^2 s_{l,22}\right), \\ V_r &= \sigma_r^2\left(s_{r,12}^2 s_{r,20} - 2s_{r,12}s_{r,11}s_{r,21} + s_{r,11}^2 s_{r,22}\right), \\ V &= \begin{pmatrix} V_l & 0 \\ 0 & V_r \end{pmatrix}. \end{aligned} \tag{18}$$

A.1 Proof of Theorem 3.1

Proof of (i). From Lemma A.1 (iii), the first-order condition for $\hat{\lambda}(\theta_s, \alpha_l)$, which solves the optimization problem in (8), satisfies

$$0 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i(\theta_s,\alpha_l)}{1+\hat{\lambda}(\theta_s,\alpha_l)' g_i(\theta_s,\alpha_l)} = \frac{1}{nh}\sum_{i=1}^{n} g_i(\theta_s,\alpha_l) - \hat{V}_1\hat{\lambda}(\theta_s,\alpha_l), \tag{19}$$

w.p.a.1 (with probability approaching one), where $\hat{V}_1 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i(\theta_s,\alpha_l)g_i(\theta_s,\alpha_l)'}{\left(1+\dot{\lambda}' g_i(\theta_s,\alpha_l)\right)^2}$, the second equality follows from an expansion around $\hat{\lambda}(\theta_s,\alpha_l) = 0$, and $\dot{\lambda}$ is a point on the line joining $\hat{\lambda}(\theta_s,\alpha_l)$ and 0. Since $\left|\hat{V}_1 - V\right| \leq \max_{1\leq i\leq n}\left|\frac{1}{1+\dot{\lambda}' g_i(\theta_s,\alpha_l)}\right|^2 \left|\frac{1}{nh}\sum_{i=1}^{n} g_i(\theta_s,\alpha_l) g_i(\theta_s,\alpha_l)' - V\right| \xrightarrow{p} 0$ (by Lemma A.1 (ii) and (iii)) and V is positive definite (Assumption 3.1 (ii)), $\hat{V}_1$ is invertible w.p.a.1. Thus, we have $\hat{\lambda}(\theta_s,\alpha_l) = \hat{V}_1^{-1}\frac{1}{nh}\sum_{i=1}^{n} g_i(\theta_s,\alpha_l)$ w.p.a.1, and a second-order expansion of $\ell_s(\theta_s,\alpha_l) = 2\sum_{i=1}^{n}\log\left(1+\hat{\lambda}(\theta_s,\alpha_l)' g_i(\theta_s,\alpha_l)\right)$ w.p.a.1 (by Lemma A.1 (iii)) around $\hat{\lambda}(\theta_s,\alpha_l) = 0$ yields

$$\begin{aligned} \ell_s(\theta_s,\alpha_l) &= 2\hat{\lambda}(\theta_s,\alpha_l)'\sum_{i=1}^{n} g_i(\theta_s,\alpha_l) - \hat{\lambda}(\theta_s,\alpha_l)'\hat{V}_2\hat{\lambda}(\theta_s,\alpha_l) \\ &= \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i(\theta_s,\alpha_l)\right)'\left[2\hat{V}_1^{-1} - \hat{V}_1^{-1}\hat{V}_2\hat{V}_1^{-1}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i(\theta_s,\alpha_l)\right), \end{aligned} \tag{20}$$

w.p.a.1, where $\hat{V}_2 = \frac{1}{nh}\sum_{i=1}^{n}\frac{g_i(\theta_s,\alpha_l)g_i(\theta_s,\alpha_l)'}{\left(1+\ddot{\lambda}' g_i(\theta_s,\alpha_l)\right)^2}$ and $\ddot{\lambda}$ is a point on the line joining $\hat{\lambda}(\theta_s,\alpha_l)$ and 0. Since $\left|\hat{V}_2 - V\right| \xrightarrow{p} 0$ by the same argument to $\hat{V}_1$, we have $2\hat{V}_1^{-1} - \hat{V}_1^{-1}\hat{V}_2\hat{V}_1^{-1} \xrightarrow{p} V^{-1}$. Therefore, Lemma A.1 (ii) implies the conclusion.

Proof of (ii). Let $\hat{\alpha} = \arg\min_{a\in\mathcal{A}}\ell_s(\theta_s, a)$. Based on Lemma A.2, we can apply the same argument to derive (20), which yields

$$\ell_s(\theta_s) = \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i(\theta_s,\hat{\alpha})\right)'\left[2\tilde{V}_1^{-1} - \tilde{V}_1^{-1}\tilde{V}_2\tilde{V}_1^{-1}\right]\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n} g_i(\theta_s,\hat{\alpha})\right), \tag{21}$$

w.p.a.1., where $\tilde{V}_1 = \frac{1}{nh}\sum_{i=1}^n \frac{g_i(\theta_s,\hat{\alpha})g_i(\theta_s,\hat{\alpha})'}{\left(1+\dot{\lambda}'g_i(\theta_s,\hat{\alpha})\right)^2}$, $\tilde{V}_2 = \frac{1}{nh}\sum_{i=1}^n \frac{g_i(\theta_s,\hat{\alpha})g_i(\theta_s,\hat{\alpha})'}{\left(1+\ddot{\lambda}'g_i(\theta_s,\hat{\alpha})\right)^2}$, and $\dot{\lambda}$ and $\ddot{\lambda}$ are points on the line joining $\hat{\lambda}(\theta_s,\hat{\alpha})$ and 0. Also, Lemma A.2 implies $2\tilde{V}_1^{-1} - \tilde{V}_1^{-1}\tilde{V}_2\tilde{V}_1^{-1} \xrightarrow{p} V^{-1}$.

We now derive the asymptotic distribution of $\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\hat{\alpha})$. From Lemma A.2 (ii), $\hat{\lambda}(\theta_s,\hat{\alpha})$ satisfies the first-order condition

$$0 = \frac{1}{nh}\sum_{i=1}^n \frac{g_i(\theta_s,\hat{\alpha})}{1+\hat{\lambda}(\theta_s,\hat{\alpha})'g_i(\theta_s,\hat{\alpha})}, \tag{22}$$

w.p.a.1. Since the derivative of this condition with respect to $\hat{\lambda}(\theta_s,\hat{\alpha})$ converges in probability to the positive definite matrix V (by Lemma A.2), we can apply the implicit function theorem, i.e. $\hat{\lambda}(\theta_s,a)$ is continuously differentiable with respect to a in a neighborhood of $\hat{\alpha}$ w.p.a.1. Let $\frac{\partial g_i(\theta_s,a)}{\partial a} = -\left((1-I_i)K_{li}, I_iK_{ri}\right)' = -G_i$. The envelope theorem implies

$$0 = \frac{1}{nh}\sum_{i=1}^n \frac{-G_i'\hat{\lambda}(\theta_s,\hat{\alpha})}{1+\hat{\lambda}(\theta_s,\hat{\alpha})'g_i(\theta_s,\hat{\alpha})} = -\hat{G}_1'\hat{\lambda}(\theta_s,\hat{\alpha}), \tag{23}$$

w.p.a.1, where $\hat{G}_1$ is implicitly defined. On the other hand, an expansion of (22) around $\left(\hat{\alpha},\hat{\lambda}(\theta_s,\hat{\alpha})\right) = (\alpha_l, 0)$ yields

$$\begin{aligned} 0 &= \frac{1}{nh}\sum_{i=1}^n g_i(\theta_s,\alpha_l) + \frac{1}{nh}\sum_{i=1}^n \frac{-G_i(\hat{\alpha}-\alpha_l)}{1+\tilde{\lambda}'g_i(\theta_s,\tilde{\alpha})} - \frac{1}{nh}\sum_{i=1}^n \frac{g_i(\theta_s,\tilde{\alpha})g_i(\theta_s,\tilde{\alpha})'}{\left(1+\tilde{\lambda}'g_i(\theta_s,\tilde{\alpha})\right)^2}\hat{\lambda}(\theta_s,\hat{\alpha}) \\ &= \frac{1}{nh}\sum_{i=1}^n g_i(\theta_s,\alpha_l) - \hat{G}_2(\hat{\alpha}-\alpha_l) - \hat{V}_3\hat{\lambda}(\theta_s,\hat{\alpha}), \end{aligned} \tag{24}$$

where $\left(\tilde{\alpha},\tilde{\lambda}\right)$ is a point on the line joining $\left(\hat{\alpha},\hat{\lambda}(\theta_s,\hat{\alpha})\right)$ and $(\alpha_l,0)$, and $\hat{G}_2$ and $\hat{V}_3$ are implicitly defined. Combining (23) and (24),

$$0 = \begin{pmatrix} 0 \\ \frac{1}{nh}\sum_{i=1}^n g_i(\theta_s,\alpha_l) \end{pmatrix} + \hat{M}\begin{pmatrix} \hat{\alpha}-\alpha_l \\ \hat{\lambda}(\theta_s,\hat{\alpha}) \end{pmatrix}, \quad \text{where } \hat{M} = \begin{pmatrix} 0 & -\hat{G}_1' \\ -\hat{G}_2 & -\hat{V}_3 \end{pmatrix}. \tag{25}$$

Lemma A.2 implies $\hat{V}_3 \xrightarrow{p} V$, $\hat{G}_1 \xrightarrow{p} G$, and $\hat{G}_2 \xrightarrow{p} G$, where $G = \frac{f(c)}{2}(1,1)'$. Thus, $\hat{M}$ is invertible w.p.a.1. By solving (25) for $\sqrt{nh}(\hat{\alpha}-\alpha_l)$, we have $\sqrt{nh}(\hat{\alpha}-\alpha_l) = \left(G'V^{-1}G\right)^{-1}G'V^{-1}\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\alpha_l) + o_p(1)$. From this and an expansion of $\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\hat{\alpha})$ around $\hat{\alpha} = \alpha_l$,

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\hat{\alpha}) = \left[I - G\left(G'V^{-1}G\right)^{-1}G'V^{-1}\right]\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\alpha_l) + o_p(1). \tag{26}$$

From (21), (26), and $\frac{1}{\sqrt{nh}}\sum_{i=1}^n g_i(\theta_s,\alpha_l) \xrightarrow{d} N(0,V)$ (by Lemma A.1 (ii)),

$$\begin{aligned} \ell_s(\theta_s) &\xrightarrow{d} \phi'V^{1/2}\left[I - G\left(G'V^{-1}G\right)^{-1}G'V^{-1}\right]'V^{-1}\left[I - G\left(G'V^{-1}G\right)^{-1}G'V^{-1}\right]V^{1/2}\phi \\ &= \phi'\left[I - A\left(A'A\right)^{-1}A'\right]\phi = \chi^2(1), \end{aligned}$$

where $\phi \sim N(0,I)$ and $A = V^{-1/2}G$. Therefore, the conclusion is obtained.

A.2 Lemmas

Define

$$\begin{aligned} S_{ln,j} &= \frac{1}{nh}\sum_{i=1}^{n}\left(1-I_i\right)\mathbb{K}\left(\frac{X_i-c}{h}\right)\left(\frac{X_i-c}{h}\right)^j, \quad S_{rn,j}=\frac{1}{nh}\sum_{i=1}^{n}I_i\mathbb{K}\left(\frac{X_i-c}{h}\right)\left(\frac{X_i-c}{h}\right)^j, \\ \mu\left(x\right) &= \mathrm{E}\left[\left.Y_i\right|X_i=x\right]-\theta_s\mathbb{I}\left\{x\geq c\right\}, \quad \mu_l'=\lim_{\delta\uparrow 0}\frac{\mu\left(c+\delta\right)-\mu\left(c\right)}{\delta}, \quad \mu_r'=\lim_{\delta\downarrow 0}\frac{\mu\left(c+\delta\right)-\mu\left(c\right)}{\delta}. \end{aligned}$$

Note that $K_{li}=\mathbb{K}\left(\frac{X_i-c}{h}\right)\left\{S_{ln,2}-\left(\frac{X_i-c}{h}\right)S_{ln,1}\right\}$ and $K_{ri}=\mathbb{K}\left(\frac{X_i-c}{h}\right)\left\{S_{rn,2}-\left(\frac{X_i-c}{h}\right)S_{rn,1}\right\}$.

Lemma A.1. *Suppose that Assumption 3.1 (i)-(iv) holds. Then*

(i) $S_{ln,1}-s_{l,11}=O_p\left(\left(nh\right)^{-1/2}\right)+O\left(h\right)$, $S_{ln,2}-s_{l,12}=O_p\left(\left(nh\right)^{-1/2}\right)+O\left(h\right)$, $S_{rn,1}-s_{r,11}=O_p\left(\left(nh\right)^{-1/2}\right)+O\left(h\right)$, *and* $S_{rn,2}-s_{r,12}=O_p\left(\left(nh\right)^{-1/2}\right)+O\left(h\right)$,

(ii) $\frac{1}{nh}\sum_{i=1}^{n}g_i\left(\theta_s,\alpha_l\right)g_i\left(\theta_s,\alpha_l\right)'\xrightarrow{p}V$, *and* $\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}g_i\left(\theta_s,\alpha_l\right)\xrightarrow{d}N\left(0,V\right)$,

(iii) *there exists* $\hat{\lambda}\left(\theta_s,\alpha_l\right)\in int\left(\Lambda_n\left(\theta_s,\alpha_l\right)\right)$ *satisfying*
$\sum_{i=1}^{n}\log\left(1+\hat{\lambda}\left(\theta_s,\alpha_l\right)'g_i\left(\theta_s,\alpha_l\right)\right)=\sup_{\lambda\in\Lambda_n\left(\theta_s,\alpha_l\right)}\sum_{i=1}^{n}\log\left(1+\lambda'g_i\left(\theta_s,\alpha_l\right)\right)$ *w.p.a.1,*
$\left|\hat{\lambda}\left(\theta_s,\alpha_l\right)\right|=O_p\left(\left(nh\right)^{-1/2}\right)$, *and* $\max_{1\leq i\leq n}\left|\hat{\lambda}\left(\theta_s,\alpha_l\right)'g_i\left(\theta_s,\alpha_l\right)\right|\xrightarrow{p}0$.

Proof of (i). We only prove the first statement. The other statements can be shown in the same manner. By the change of variables and an expansion $f\left(c+hz\right)$ around $hz=0$,

$$\mathrm{E}\left[S_{ln,1}\right]-s_{l,11}=\int_{-k}^{0}\mathbb{K}\left(z\right)zf\left(c+hz\right)dz-s_{l,11}=h\int_{-k}^{0}\mathbb{K}\left(z\right)z^2f'\left(c_z\right)dz=O\left(h\right),$$

where c_z is a point on the line joining c and $c+hz$ and the last equality follows from Assumption 3.1 (ii) and (iii). Also, a similar argument yields

$$\mathrm{Var}\left(S_{ln,1}\right)\leq\frac{1}{nh^2}\mathrm{E}\left[\left(1-I_i\right)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2\left(\frac{X_i-c}{h}\right)^2\right]\frac{1}{nh}\int_{-k}^{0}\mathbb{K}\left(z\right)^2z^2f\left(c+hz\right)dz=O\left(\left(nh\right)^{-1}\right).$$

Therefore, Lyapunov's central limit theorem implies $S_{ln,1}-\mathrm{E}\left[S_{ln,1}\right]=O_p\left(\left(nh\right)^{-1/2}\right)$. Combining these results, the conclusion is obtained.

Proof of (ii). Proof of the first statement. It is sufficient to show that

$$\frac{1}{nh}\sum_{i=1}^{n}\left(1-I_i\right)K_{li}^2\left(Y_i-\alpha_l\right)^2\xrightarrow{p}V_l, \quad \frac{1}{nh}\sum_{i=1}^{n}I_iK_{ri}^2\left(Y_i-\theta_s-\alpha_l\right)^2\xrightarrow{p}V_r.$$

Since the proofs are similar, we only show the first statement. By the definition of K_{li}^2,

$$\begin{aligned}
&\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)K_{li}^2(Y_i-\alpha_l)^2 \\
= \;& S_{ln,2}^2\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2(Y_i-\alpha_l)^2+S_{ln,1}^2\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2\left(\frac{X_i-c}{h}\right)^2(Y_i-\alpha_l)^2 \\
&-2S_{ln,2}S_{ln,1}\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l)^2. \qquad (27)
\end{aligned}$$

By the same argument to the proof of Part (i) of this lemma,

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2(Y_i-\alpha_l)^2\right] &\to \sigma_l^2 s_{l,20}, \\
\mathrm{Var}\left(\frac{1}{nh}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)^2(Y_i-\alpha_l)^2\right) &\to 0, \qquad (28)
\end{aligned}$$

Thus, from Chebyshev's inequality and Lemma A.1 (i), the probability limit of the first term in (27) is $\sigma_l^2 s_{l,12}^2 s_{l,20}$. By applying the same argument to the second and third terms of (27), we obtain the conclusion.

Proof of the second statement. From the definition of $g_i(\theta_s,\alpha_l)$, it is sufficient to show that

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(1-I_i)K_{li}(Y_i-\alpha_l)\xrightarrow{d}N(0,V_l), \quad \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}I_iK_{ri}(Y_i-\theta_s-\alpha_l)\xrightarrow{d}N(0,V_r).$$

Since the proofs are similar, we only show the first statement. From the definition of K_{li},

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(1-I_i)K_{li}(Y_i-\alpha_l) \\
= \;& (S_{ln,2}-s_{l,12})\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l) \\
&-(S_{ln,1}-s_{l,11})\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l) \\
&+\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left\{\begin{array}{c}(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)\left\{s_{l,12}-\left(\frac{X_i-c}{h}\right)s_{l,11}\right\}(Y_i-\alpha_l) \\ -\mathrm{E}\left[(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)\left\{s_{l,12}-\left(\frac{X_i-c}{h}\right)s_{l,11}\right\}(Y_i-\alpha_l)\right]\end{array}\right\} \\
&+\sqrt{\frac{n}{h}}\mathrm{E}\left[(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)\left\{s_{l,12}-\left(\frac{X_i-c}{h}\right)s_{l,11}\right\}(Y_i-\alpha_l)\right] \\
= \;& T_1-T_2+T_3+T_4.
\end{aligned}$$

For T_1, Lyapunov's central limit theorem implies

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left\{(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l)-\mathrm{E}\left[(1-I_i)\mathbb{K}\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l)\right]\right\}\xrightarrow{d}N\left(0,\sigma_l^2 s_{l,20}\right),$$

and the change of variables and Assumption 3.1 (ii)-(iv) imply

$$\mathrm{E}\left[(1-I_i)\,\mathbb{K}\left(\frac{X_i-c}{h}\right)(Y_i-\alpha_l)\right]=h\int_{-k}^{0}\mathbb{K}(z)\left(E\left[Y_i|\,X_i=c+hz\right]-\alpha_l\right)f(c+hz)\,dz=h^2\mu_l' s_{l,10}+O\left(h^3\right).$$

Thus, from Lemma A.1 (i) and Assumption 3.1 (iv), we have $T_1=o_p(1)$. Similarly, we can show that $T_2=o_p(1)$. For T_4, the change of variables and Assumption 3.1 (ii)-(iv) yield

$$\begin{aligned} T_4 &= \sqrt{nh}\int_{-k}^{0}\mathbb{K}(z)\left(s_{l,12}-s_{l,11}z\right)\left(\mathrm{E}\left[Y_i|\,X_i=c+hz\right]-\alpha_l\right)f(c+hz)\,dz \\ &= \sqrt{nh}h\mu_l'\left(s_{l,12}s_{l,10}-s_{l,11}^2\right)+O\left(\sqrt{nh}h^2\right)\to 0. \end{aligned}$$

For T_3, note that

$$\begin{aligned} \mathrm{E}\left[T_3^2\right] &= \int_{-k}^{0}\mathbb{K}(z)^2\left(s_{l,12}-s_{l,11}z\right)^2\mathrm{E}\left[\left.(Y_i-\alpha_l)^2\right|X_i=c+hz\right]f(c+hz)\,dz \\ &\quad -h\left(\int_{-k}^{0}\mathbb{K}(z)\left(s_{l,12}-s_{l,11}z\right)\left(\mathrm{E}\left[Y_i|\,X_i=c+hz\right]-\alpha_l\right)f(c+hz)\,dz\right)^2 \\ &\to \sigma_l^2\left(s_{l,12}^2s_{l,20}-2s_{l,12}s_{l,11}s_{l,21}+s_{l,11}^2s_{l,22}\right)=V_l, \end{aligned}$$

where the convergence follows from a similar argument to (28). Therefore, Lyapunov's central limit theorem implies $T_3\xrightarrow{d}N(0,V_l)$. Combining these results, we obtain the conclusion.

Proof of (iii). Since the proof is similar to Newey and Smith (2004, Lemmas A1 and A2), it is omitted.

Lemma A.2. *Suppose that Assumption 3.1 holds. Then*

(i) $S_{ln,0}-s_{l,10}=O_p\left((nh)^{-1/2}\right)+O(h)$, *and* $S_{rn,0}-s_{r,10}=O_p\left((nh)^{-1/2}\right)+O(h)$,

(ii) $\frac{1}{nh}\sum_{i=1}^{n}g_i(\theta_s,\hat{\alpha})\,g_i(\theta_s,\hat{\alpha})'\xrightarrow{p}V$, *and* $\left|\frac{1}{nh}\sum_{i=1}^{n}g_i(\theta_s,\hat{\alpha})\right|=O_p\left((nh)^{-1/2}\right)$,

(iii) *there exists* $\hat{\lambda}(\theta_s,\hat{\alpha})\in int\left(\Lambda_n(\theta_s,\hat{\alpha})\right)$ *satisfying*
$\sum_{i=1}^{n}\log\left(1+\hat{\lambda}(\theta_s,\hat{\alpha})'g_i(\theta_s,\hat{\alpha})\right)=\sup_{\lambda\in\Lambda_n(\theta_s,\hat{\alpha})}\sum_{i=1}^{n}\log\left(1+\lambda' g_i(\theta_s,\hat{\alpha})\right)$ *w.p.a.1,*
$\left|\hat{\lambda}(\theta_s,\hat{\alpha})\right|=O_p\left((nh)^{-1/2}\right)$, *and* $\max_{1\le i\le n}\left|\hat{\lambda}(\theta_s,\hat{\alpha})'g_i(\theta_s,\hat{\alpha})\right|\xrightarrow{p}0$.

Detailed proofs are available from the authors upon request. The proof of Lemma A.2 (i) is similar to that of Lemma A.1 (i). The second statement of Lemma A.2 (ii) follows from a similar argument to the proof of Newey and Smith (2004, Lemma A3) combined with Lemma A.1. Since this statement implies the weak consistency of $\hat{\alpha}$ to α_l, Lemma A.1 (ii) implies the first statement of Lemma A.2 (ii). Also, given the consistency of $\hat{\alpha}$ and Lemma A.2 (ii), a similar argument to the proof of Newey and Smith (2004, Lemma A2) implies Lemma A.2 (iii).

References

[1] Angrist, J. D. and Lavy, V. (1999) Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 533-575.

[2] Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444-472.

[3] Chan, N. H., Peng, L. and Zhang, D. (2010) Empirical likelihood based confidence intervals for conditional variance in heteroskedastic regression models. Forthcoming in *Econometric Theory.*

[4] Chen, S. X. and Qin, Y. S. (2000) Empirical likelihood confidence intervals for local linear smoothers. *Biometrika*, 87, 946-953.

[5] Davidson, R. and MacKinnon, J. G. (1998) Graphical methods for investigating the size and power of test statistics. *The Manchester School*, 66, 1-26.

[6] Davidson, R. and MacKinnon, J. G. (2006) The power of bootstrap and asymptotic tests. *Journal of Econometrics*, 133, 421-441.

[7] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall.

[8] Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, 29, 153-193.

[9] Fan, J. and Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85, 645-660.

[10] Hahn, J., Todd, P. and van der Klaauw, W. (2001) Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica*, 69, 201-209.

[11] Holland, P. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945-970.

[12] Horowitz, J. L. and Savin, N. E. (2000) Empirically relevant critical values for hypothesis tests. Journal of Econometrics, 95, 375–389.

[13] Imbens, G. W. and Lemieux, T. (2008) Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, 142, 615-635.

[14] Knafl, G., Sacks, J. and Ylvisaker, D. (1985) Confidence bands for regression functions. *Journal of the American Statistical Association*, 80, 683-691.

[15] Lemieux, T. and Marmer, V. (2009) Weak identification in fuzzy regression discontinuity design. Working paper, Department of Economics, University of British Columbia.

[16] Li, Q. and Racine, J. (2004) Cross-validated local linear nonparametric regression. *Statistica Sinica,* 14, 485-512.

[17] Newey, W. K. and Smith, R. J. (2004) Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219-255.

[18] Owen, A. B.(1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

[19] Owen, A. B. (2001) *Empirical Likelihood.* New York: Chapman & Hall.

[20] Porter, J. (2003) Estimation in the regression discontinuity model. Working paper, Department of Economics, University of Wisconsin.

[21] Rubin, D. (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 51, 309-317.

[22] Ruppert, D., Wand, M. P., Holst, U. and Hössjer, O. (1997) Local polynomial variance function estimation. *Technometrics*, 39, 262-273.

[23] Sacks, J. and Ylvisaker, D. (1978) Linear estimates for approximately linear models. *Annals of Statistics*, 6, 1122-1138.

[24] Stock, J. H. and Wright, J. H. (2000) GMM with weak identification. *Econometrica*, 68, 1055-1096.

[25] Thistlethwaite, D. and Campbell, D. (1960) Regression-discontinuity analysis: an alternative to the ex-post factor experiment. *Journal of Educational Psychology*, 51, 309-317.

[26] Trochim, W. (2001) Regression-discontinuity design. In N. J. Smelser and P. B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, vol. 19, pp. 12940–12945, Oxford, UK: Elsevier.

[27] Xu, K.-L. (2009) Empirical likelihood based inference for recurrent nonparametric diffusions. *Journal of Econometrics,* 153, 65-82.

[28] Xu, K.-L. and Phillips, P.C.B. (2009) Tilted nonparametric estimation of volatility functions with empirical applications. *Cowles Foundation Discussion Paper* 1612R, Yale University.

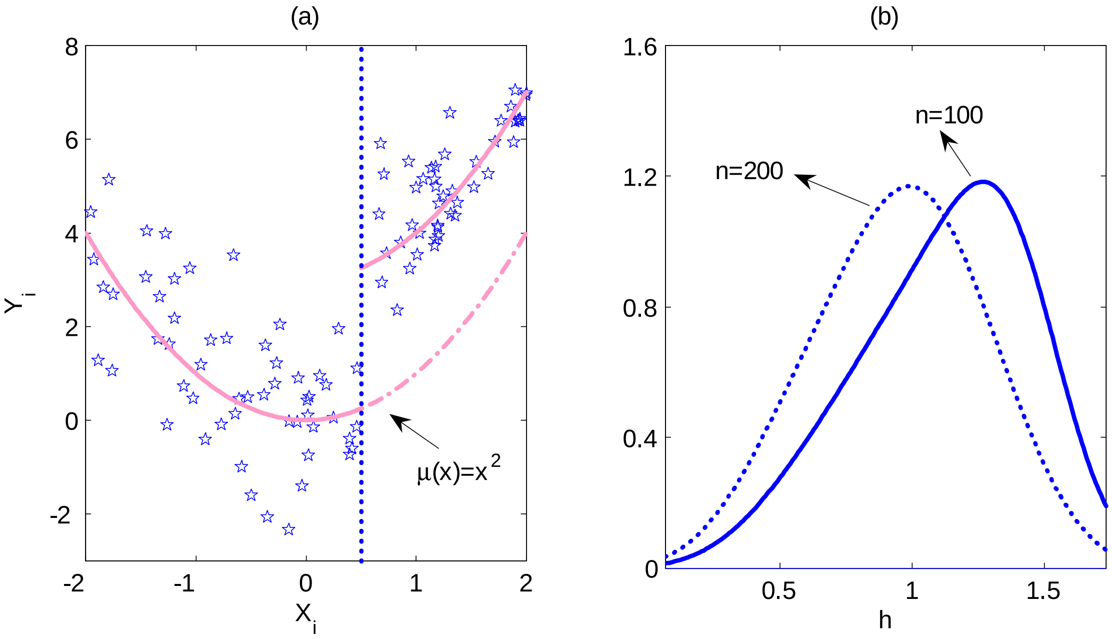

Figure 1: (a) A representative sample with 100 observations; (b) The distributions of the bandwidths selected by cross validation over 1000 replications when the sample sizes are 100 and 200.

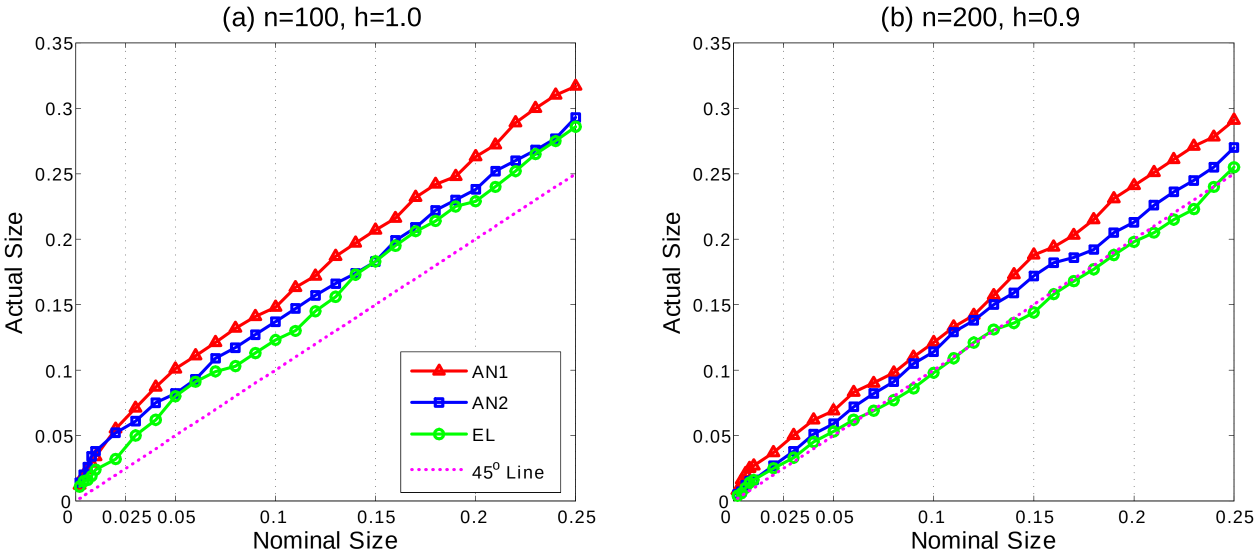

Figure 2: P-value plots (Davidson and MacKinnon, 1998) for the two squared t-test statistics (AN1 and AN2) and empirical likelihood-based test statistic (EL).

Table 1: The rejection rates (under the null) of two t-tests and the empirical likelihood-based test with various fixed bandwidths and the one selected via cross validation, when the nominal sizes are 5% and 10% and the sample size is 100. (Standard errors are in the parentheses.)

Bandwidth	*Tests*	5% *Sizes*	10% *Sizes*	$\widehat{\alpha}_r$	$\widehat{\alpha}_l$	$\widehat{\theta}_s$	$var(\widehat{\theta}_s)$	$\widehat{var(\widehat{\theta}_s)}$	$\widehat{\sigma}_r^2$	$\widehat{\sigma}_l^2$
True values of parameters				3.25	0.25	3			2.30	1.4
$h = 0.8$	AN1	0.092	0.154	3.15 (0.69)	0.20 (0.59)	2.95	0.78	0.60 (0.19)	1.29 (0.57)	1.32 (0.52)
	AN2	0.103	0.163					0.70 (0.38)	1.83 (1.14)	1.24 (0.99)
	EL	0.084	0.135							
$h = 0.9$	AN1	0.093	0.149	3.17 (0.62)	0.16 (0.57)	3.01	0.70	0.53 (0.15)	1.22 (0.53)	1.35 (0.48)
	AN2	0.087	0.138					0.62 (0.31)	1.83 (1.09)	1.22 (0.88)
	EL	0.079	0.124							
$h = 1.0$	AN1	0.103	0.158	3.13 (0.59)	0.13 (0.52)	2.99	0.63	0.46 (0.13)	1.17 (0.48)	1.38 (0.47)
	AN2	0.088	0.146					0.57 (0.25)	1.82 (0.97)	1.30 (0.79)
	EL	0.075	0.125							
$h = 1.1$	AN1	0.097	0.166	3.11 (0.55)	0.12 (0.54)	2.99	0.59	0.42 (0.11)	1.16 (0.47)	1.39 (0.46)
	AN2	0.084	0.143					0.52 (0.22)	1.79 (0.90)	1.35 (0.77)
	EL	0.068	0.122							
$h = 1.2$	AN1	0.087	0.140	3.07 (0.52)	0.06 (0.48)	3.01	0.49	0.39 (0.09)	1.14 (0.44)	1.42 (0.43)
	AN2	0.057	0.113					0.49 (0.19)	1.87 (0.88)	1.39 (0.80)
	EL	0.053	0.102							
$h = 1.3$	AN1	0.082	0.147	3.05 (0.50)	0.06 (0.45)	2.99	0.44	0.36 (0.08)	1.12 (0.42)	1.45 (0.41)
	AN2	0.069	0.116					0.45 (0.18)	1.82 (0.89)	1.37 (0.75)
	EL	0.048	0.098							
h_{cv}	AN1	0.105	0.173	3.11 (0.63)	0.11 (0.58)	3.00	0.70	0.47 (0.25)	1.22 (0.51)	1.39 (0.47)
	AN2	0.085	0.142					0.55 (0.34)	1.84 (1.00)	1.33 (0.83)
	EL	0.076	0.122							

Table 2: The rejection rates (under the null) of two t-tests and the empirical likelihood-based tests with various fixed bandwidths and the one selected via cross validation, when the nominal sizes are 5% and 10% and the sample size is 200. (Standard errors are in the parentheses.)

Bandwidth	*Tests*	5% *Sizes*	10% *Sizes*	$\widehat{\alpha}_r$	$\widehat{\alpha}_l$	$\widehat{\theta}_s$	$var(\widehat{\theta}_s)$	$\widehat{var(\widehat{\theta}_s)}$	$\widehat{\sigma}_r^2$	$\widehat{\sigma}_l^2$
True values of parameters				3.25	0.25	3			2.30	1.4
$h = 0.7$	AN1	0.077	0.143	3.18 (0.51)	0.19 (0.43)	3.00	0.45	0.35 (0.08)	1.39 (0.45)	1.34 (0.39)
	AN2	0.075	0.131					0.42 (0.16)	1.94 (0.88)	1.30 (0.69)
	EL	0.061	0.116							
$h = 0.8$	AN1	0.086	0.142	3.16 (0.47)	0.18 (0.41)	2.98	0.38	0.31 (0.07)	1.32 (0.42)	1.37 (0.38)
	AN2	0.060	0.114					0.37 (0.13)	1.96 (0.84)	1.34 (0.67)
	EL	0.058	0.113							
$h = 0.9$	AN1	0.080	0.130	3.14 (0.42)	0.18 (0.37)	2.96	0.31	0.27 (0.05)	1.26 (0.38)	1.38 (0.36)
	AN2	0.068	0.118					0.33 (0.11)	1.98 (0.82)	1.34 (0.61)
	EL	0.050	0.105							
$h = 1.0$	AN1	0.062	0.122	3.14 (0.39)	0.15 (0.36)	2.99	0.27	0.24 (0.05)	1.23 (0.35)	1.40 (0.34)
	AN2	0.074	0.120					0.30 (0.09)	1.92 (0.72)	1.37 (0.58)
	EL	0.047	0.096							
$h = 1.1$	AN1	0.087	0.148	3.12 (0.40)	0.11 (0.35)	3.01	0.28	0.21 (0.04)	1.16 (0.33)	1.44 (0.33)
	AN2	0.057	0.094					0.28 (0.08)	1.97 (0.70)	1.39 (0.58)
	EL	0.056	0.097							
$h = 1.2$	AN1	0.079	0.143	3.05 (0.37)	0.09 (0.32)	2.96	0.24	0.20 (0.03)	1.15 (0.31)	1.46 (0.30)
	AN2	0.052	0.098					0.25 (0.07)	1.95 (0.70)	1.43 (0.54)
	EL	0.048	0.099							
h_{cv}	AN1	0.111	0.170	3.17 (0.51)	0.14 (0.45)	3.02	0.33	0.29 (0.16)	1.28 (0.47)	1.40 (0.37)
	AN2	0.080	0.130					0.36 (0.19)	1.96 (0.87)	1.38 (0.63)
	EL	0.070	0.119							

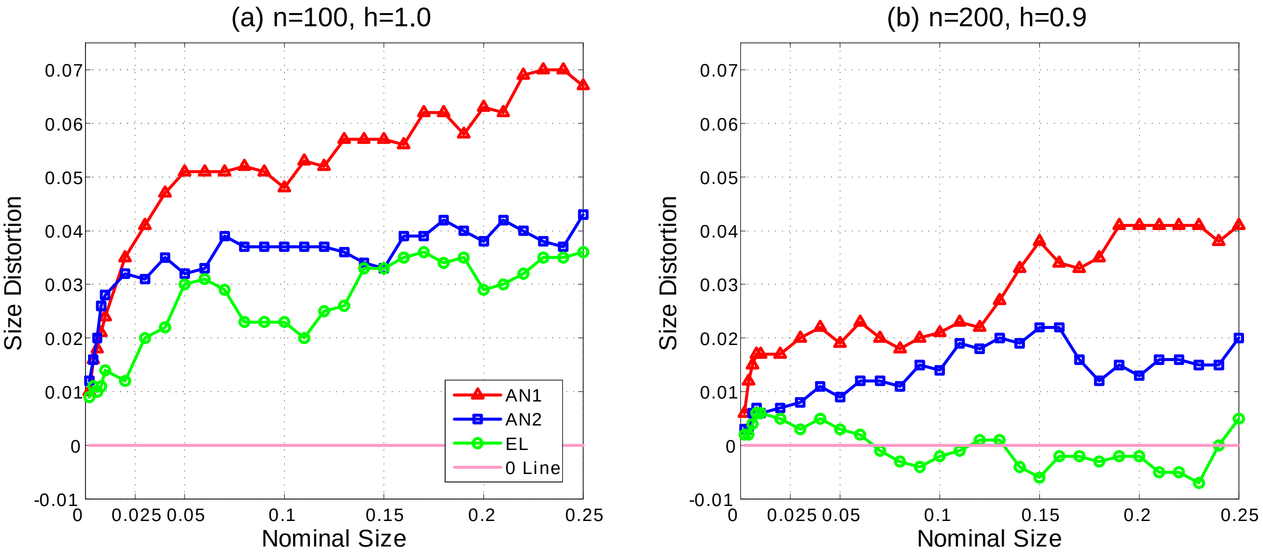

Figure 3: P-value discrepancy plots (Davidson and MacKinnon, 1998) for the two squared t-test statistics (AN1 and AN2) and empirical likelihood-based test statistic (EL).

Table 3: The calibrated 10% critical values (used to obtain the calibrated powers) of the two *squared* $t-$tests (AN1 and AN2) and the EL test.

		$n = 100$			
	$h = 0.8$	$h = 1.0$	$h = 1.2$	h_{cv}	uncalibrated
AN1	5.186	5.632	5.167	5.862	2.706
AN2	6.208	4.968	4.038	5.078	2.706
EL	3.335	3.368	2.785	3.432	2.706
		$n = 200$			
	$h = 0.7$	$h = 0.9$	$h = 1.1$	h_{cv}	uncalibrated
AN1	4.876	5.024	5.301	5.408	2.706
AN2	4.812	4.627	4.014	5.194	2.706
EL	3.123	2.803	2.893	3.331	2.706

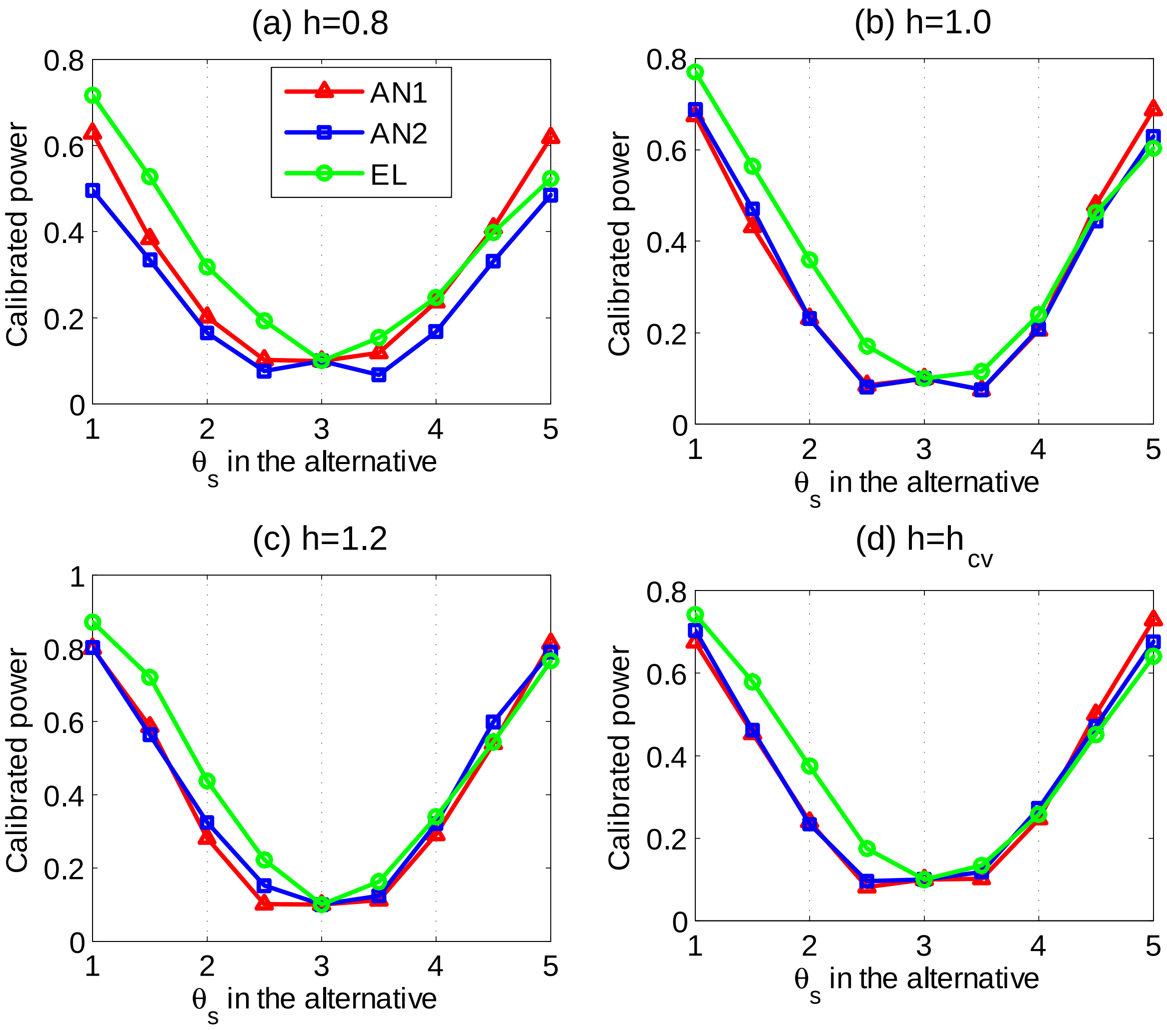

Figure 4: The calibrated powers of the two t-tests (AN1 and AN2) and the empirical likelihood-based test (EL) for various fixed bandwidths and the one selected via cross validation, when the nominal size is 10% and the sample size is 100.

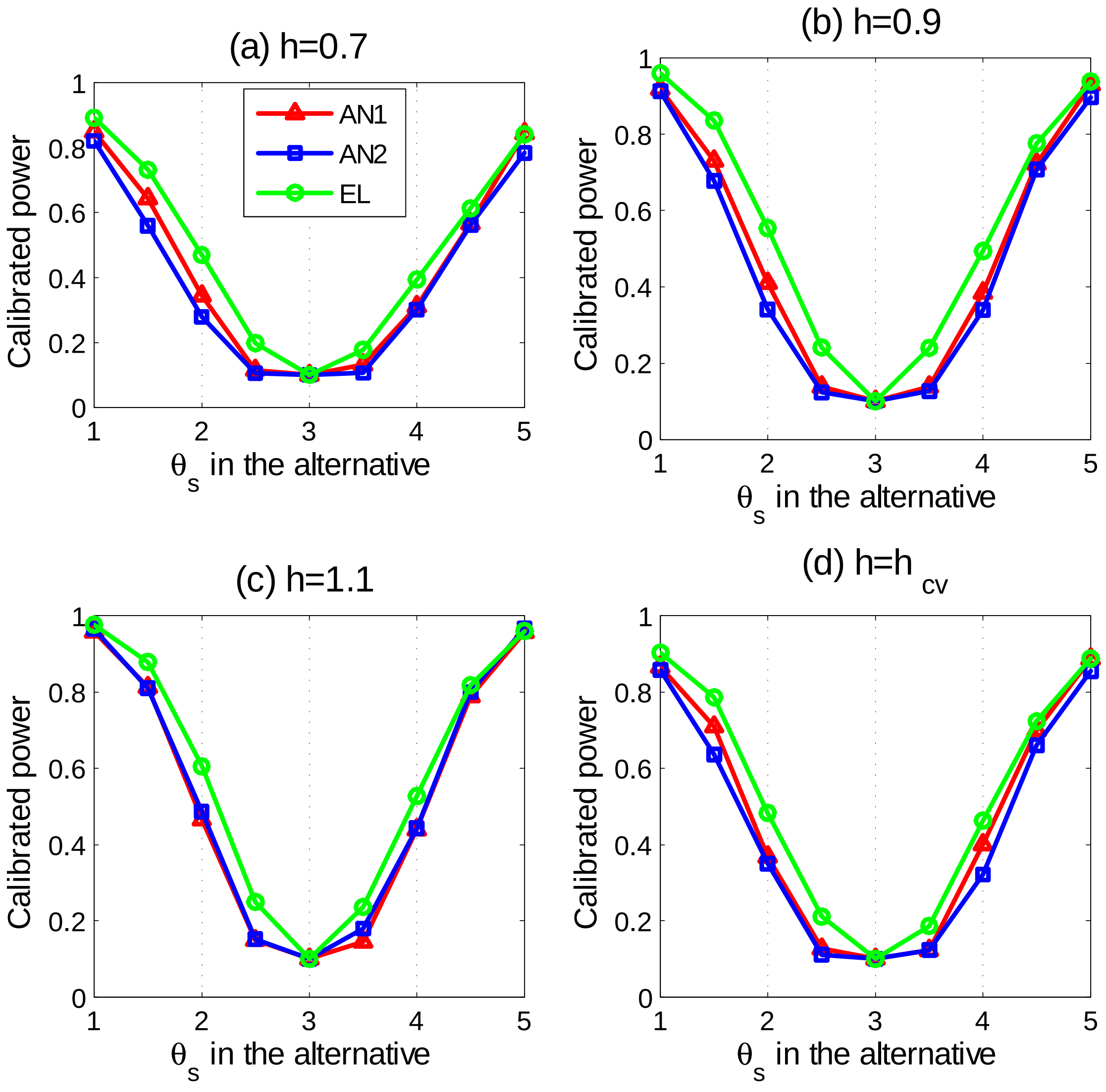

Figure 5: The calibrated powers of the two t-tests (AN1 and AN2) and the empirical likelihood-based test (EL) for various fixed bandwidths and the one selected via cross validation, when the nominal size is 10% and the sample size is 200.

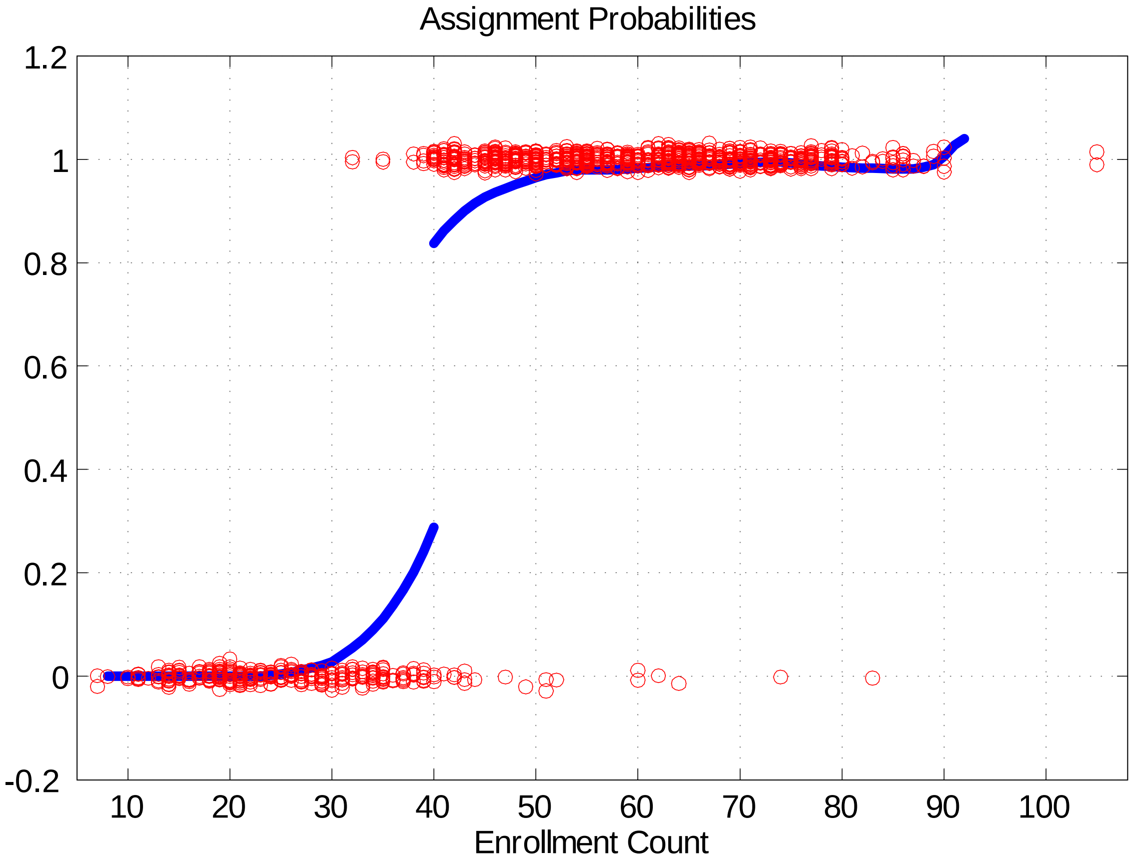

Figure 6: The plot of the assignment (with imposed random noises) by the enrollment count, and the local linear estimates of the conditional probabilities of getting treated (splitting into two classes) given the enrollment counts for the controlled sample (enrollment$\leq$ 40) and the treatment sample (enrollment$>$ 40).

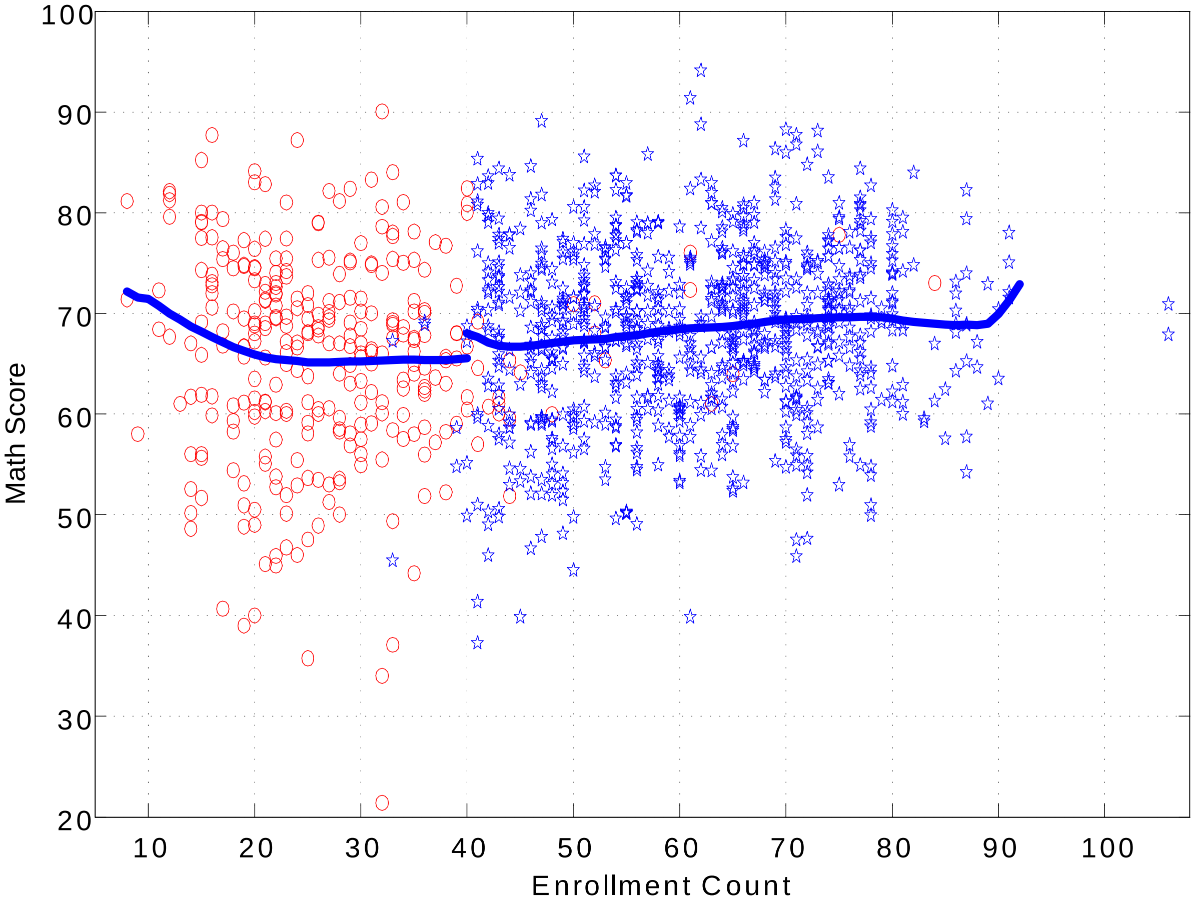

Figure 7: The plot of the average math scores by the enrollment counts, and the local linear fits for the controlled sample (enrollment≤ 40) and the treatment sample (enrollment> 40).

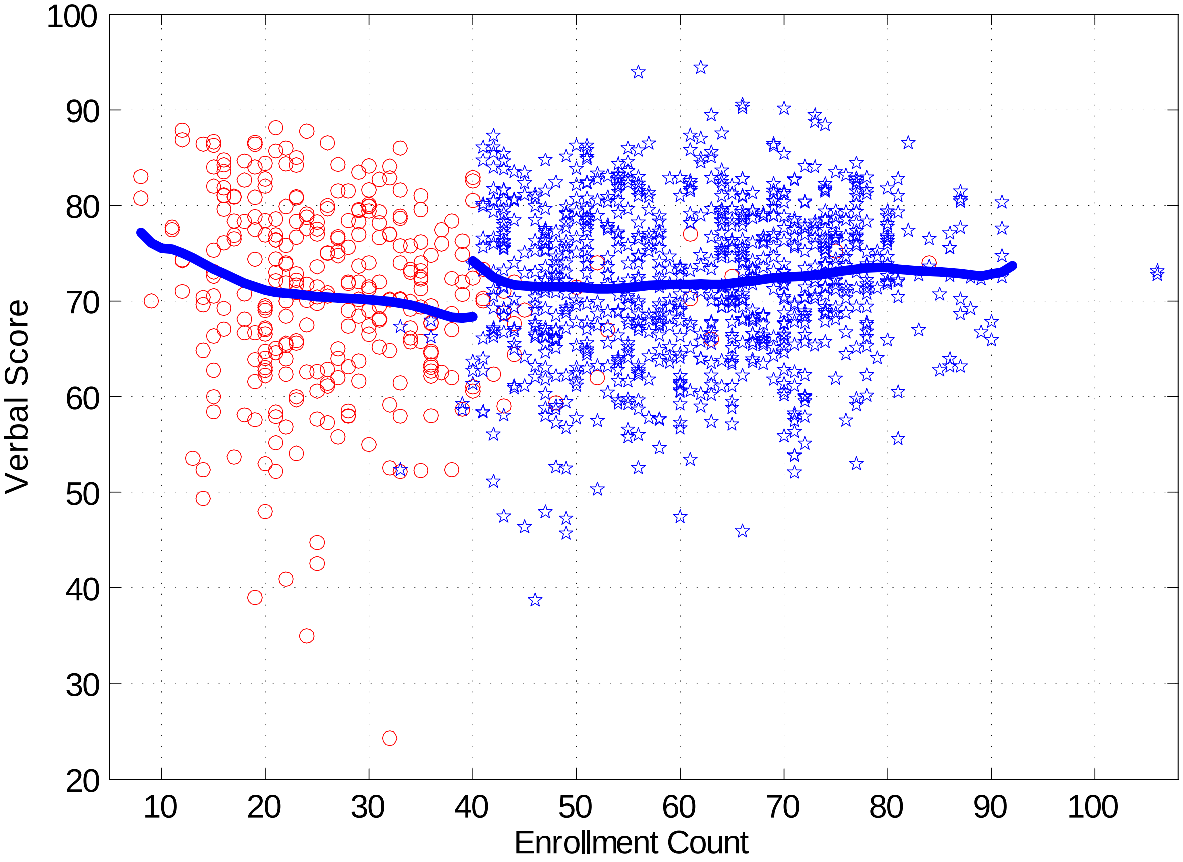

Figure 8: The plot of the average verbal scores by the enrollment counts, and the local linear fits for the controlled sample (enrollment$\leq$ 40) and the treatment sample (enrollment$>$ 40).

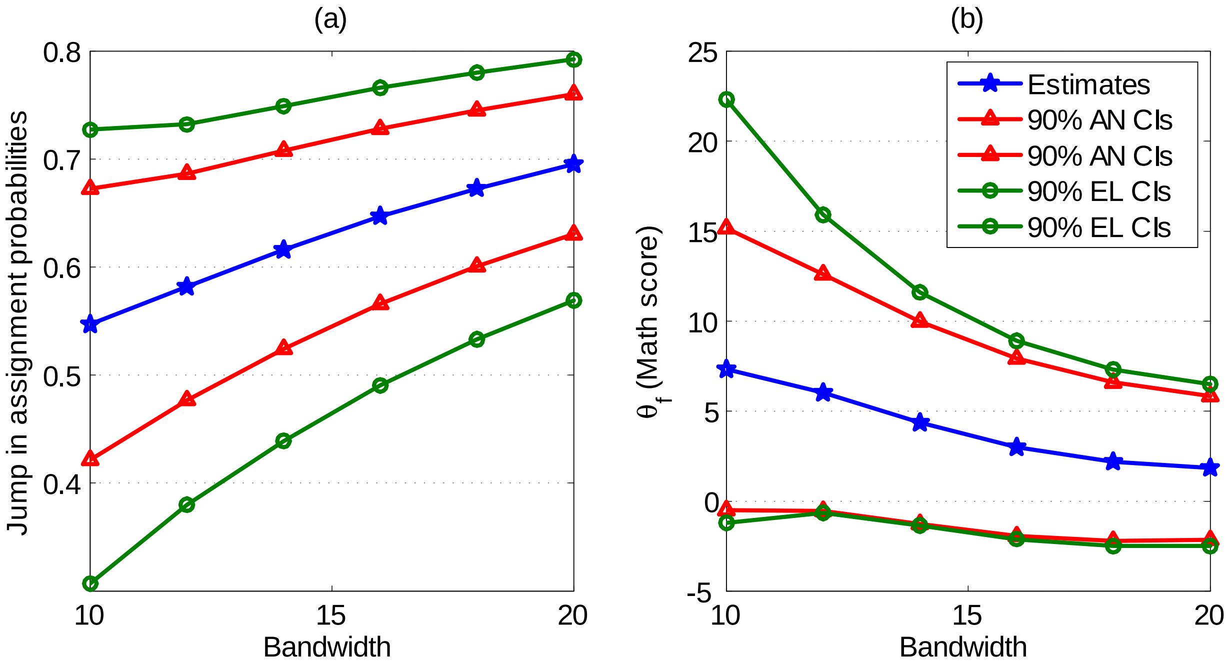

Figure 9: The local linear estimates and the 90% asymptotic normality confidence intervals (AN CIs) and empirical likelihood confidence intervals (EL CIs) of (a) the jump in the propensity score and (b) the average causal treatment effect of splitting into two classes on pupils' *math* score.

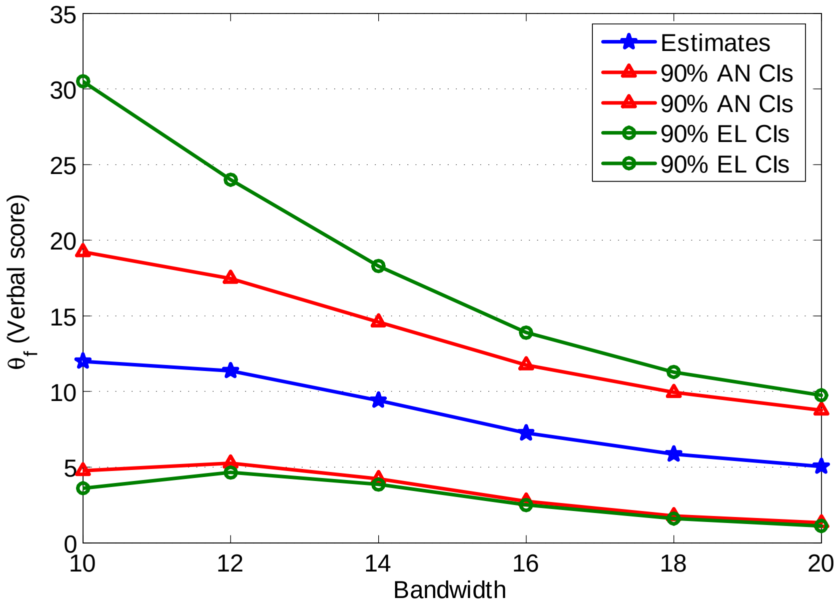

Figure 10: The local linear estimates and the 90% asymptotic normality confidence intervals (AN CIs) and empirical likelihood confidence intervals (EL CIs) of the average causal treatment effect of splitting into two classes on pupils' *verbal* score.

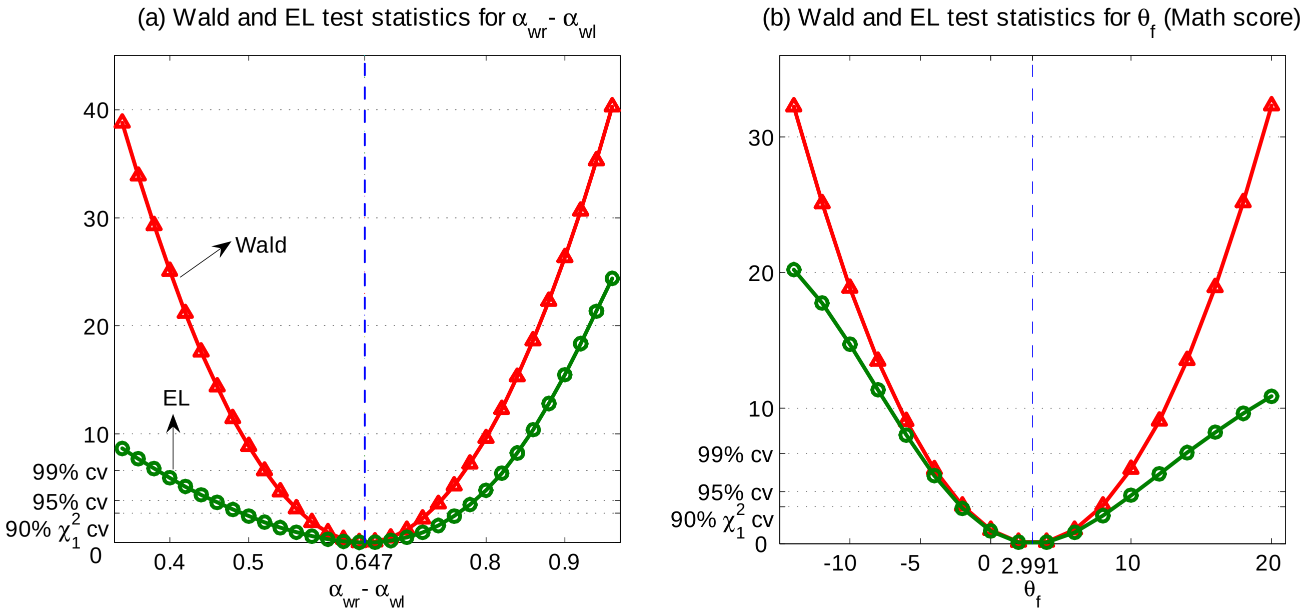

Figure 11: The Wald and empirical likelihood (EL) test statistics for (a) the jump in the propensity score and (b) the average causal treatment effect of splitting into two classes on pupils' *math* score. The smoothing bandwidth $h = 16$ is used. Both test statistics have $\chi^2(1)$ limit distribution and the 90%, 95% and 99% critical values are marked in the figures.

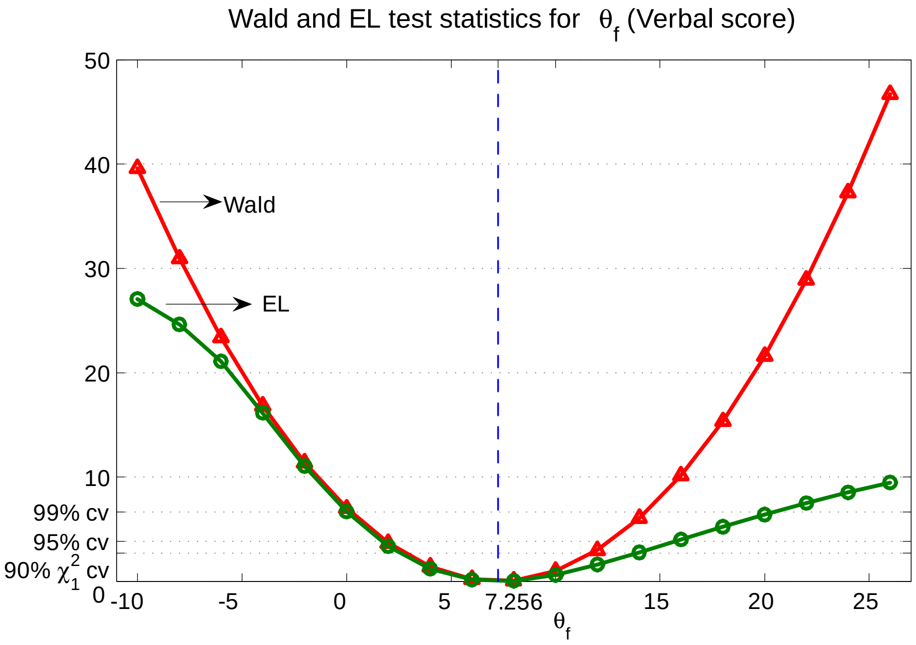

Figure 12: The Wald and EL test statistics for the average causal treatment effect of splitting into two classes on pupils' *verbal* score. The smoothing bandwidth $h = 16$ is used. Both test statistics have $\chi^2(1)$ limit distribution and the 90%, 95% and 99% critical values are marked in the figure.