Regression Discontinuity with Categorical Outcomes*

Ke-Li Xu†

Indiana University Bloomington

July 13, 2017

Abstract

We consider the regression discontinuity (RD) design with categorical outcomes, and exploit the possibility of adapting well-developed microeconometric models to the RD setting. The channels through which the forcing variable affects the potential outcome distributions are constrained to be minimal, to preserve the nonparametric feature of the RD design. Focusing on general categorical outcomes (nominal or ordinal), we develop a new RD estimator based on a nonparametric extension of the well-known multinomial logit model. The key issues of selecting the optimal bandwidth and constructing confidence regions robust to bias correction, of which the solutions only exist so far for the local linear estimator and a single treatment effect, are addressed through the general approach of local likelihood. The proposed estimator and associated inference are easy to implement, and the codes in MATLAB and R are available as a supplement to the paper. They are demonstrated by two empirical applications and simulation experiments.

Keywords: Bandwidth selection; categorical outcomes; local likelihood; multinomial logit model; nonparametric models; regression discontinuity; robust inference.

JEL classification: C14; C21; C25.

*The author thanks Editor, Associate Editor, three anonymous referees, and seminar participants at the 2015 Midwest Econometrics Group meeting at St. Louis, the economics departments at Indiana and NCSU, and Alberta School of Business for helpful discussions and comments. The author is also grateful to Jason Lindo for sharing the datasets used in the paper, and College of Liberal Arts at TAMU for part of financial support under the Rothrock fellowship at an early stage of the research.

†*Address*: Department of Economics, Indiana University, Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405-7104, USA. *E-mail*: kelixu@indiana.edu.

1 Introduction

Regression discontinuity (RD) design, as a quasi-experimental design which exploits exogenous discontinuities in policy assignment, has been popular in policy evaluation. Applications can be found in almost all fields of applied microeconomics and social science,[1] and research on relevant econometric issues has been very active recently.[2]

Among outcomes examined in a typical RD analysis, the categorical outcome is one of the most common types. They can be primary outcomes of interest, or auxiliary outcomes or covariates in which discontinuities should not exist for interpretation or research-design validity checks. They are most common for micro-level data, and examples include dropoff or graduation of students, mortality, election of a political party, enrolling in a program, taking a paternity leave, default on loans, recidivism of drunk driving, winning or losing a game, leaving zero tip, etc.. These variables can be recorded as binary or multiple-valued, e.g. graduation from college within a normal four-year period or longer, infant mortality during different time intervals, causes for a particular outcome (child/adult mortality, hospital utilization for elder patients). Multiple-valued outcomes also naturally arise when concerning consumer choices, like different levels of default suggestions offered to customers or various occupational choices. A long list of recent references are contained in this footnote.[3] Duration outcomes that are coarsely discretized (e.g. in months or weeks) can be also treated as categorical (and ordered), following Han and Hausman (1990) and Sueyoshi (1995). Recent applications of RD and related designs with this type of outcome can be found in Caliendo, Tatsiramos and Uhlendorff (2013) and Landais (2015).

[1] See Imbens and Lemieux (2008), Lee and Lemieux (2010), and Cattaneo and Escanciano (2017) for surveys on the topic.

[2] The econometric issues that have been tackled recently include the window choice (Imbens and Kalyanaraman, 2012, Arai and Ichimura, 2015), robust confidence intervals (Calonico et al., 2014, Calonico et al., 2017), uniform inference (Armstrong and Kolesár, 2015), testing for manipulation (McCrary, 2008, Otsu et al., 2013, Cattaneo et al., 2017), criterion-function-based inference (Otsu, et al., 2015, Xu, 2016a), effects of including covariates (Calonico, Cattaneo, Farrell, and Titiunik, 2016), effects of allowing manipulation (Gerard, et al., 2016), weak identification (Feir et al., 2016, Xu, 2016a, Dong, 2014), discrete running variable (Lee and Card, 2008, Kolesár and Rothe, 2017), quantile effects (Frandsen et al., 2012, Xu, 2016a), measurement errors (Dong, 2015, Barreca et al., 2016), multiple thresholds (Bertanha, 2017), unknown thresholds (Porter and Yu, 2015), clustered data (Bartalotti and Brummet, 2017), among others.

[3] Categorical outcomes are prevalent in RD applications. An incomplete list of articles that are recently published in economics journals and contain examples mentioned in this paragraph include Almond, et al. (2010), Berger and Pope (2011), Clark and Martorell (2014), Clark and Royer (2013), Cohodes and Goodman (2014), Dahl, Loken and Mogstad (2014), Dobbie and Skiba (2013), Haggag and Paci (2014), Hansen (2015), Lindo, et al. (2010), Malamud and Pop-Eleches (2010), Shigeoka (2014).

The methodological literature on RD and related designs has been exclusively (or slightly less so) focused on the local linear or local polynomial estimator which is best suited for continuous outcomes. When the outcome takes on only a few values, which are not necessarily ordered, the direct practice of the existing methodological recommendations can cause unsatisfactory results. This is especially true for multinomial outcomes. We highlight the following points to motivate our approach.

First, considering the simple case of a binary outcome, a nonlinear transformation such that the estimated probability belongs to the unit interval seems natural.[4] The popular bandwidth selecting procedure by Imbens and Kalyanaraman (2012) which is developed for the local linear estimator becomes suboptimal (still with the optimal rate though) for the local nonlinear estimator of the probability function. This is because the nonlinear transformation twists the bias which an econometrician evaluates and balances with the variance for optimal bandwidth determination.

Second, for the same reason above caused by using the transformation, the bias corrector and the robust standard error under bias correction proposed by Calonico, Cattaneo, and Titiunik (2014) no longer apply.

Third, one standard practice to deal with binary outcomes is to first aggregate over different bins along the forcing variable so that the outcomes are transformed to fractions and thus can be treated as continuous. This method involves an additional tuning parameter, the bin size for aggregation, and its determination is often subjective. The automatic optimal plots method by Calonico, Cattaneo, and Titiunik (2015) can be used, however, it remains unclear how this choice of aggregation affects the choice of the window size around the cutoff and the final RD estimate.

Fourth, the concerns above also arise in RD applications with multiple-valued outcomes. In these cases, like those in the traditional regression-based microeconometric analysis, we are interested in the effect on the probability of the outcome belonging to each category. One standard practice is to first generate a binary outcome for each category (against the chosen baseline category) and then apply the standard RD methods for each binary outcome. Besides the issues mentioned above, another important drawback of this practice is that it views each category in isolation (ignoring the

[4] Hall, Wolff and Yao (1999) and Xu and Phillips (2011) discussed the possibilities that the local linear functional estimator may fail to respect the natural range of the estimand (e.g. a probability or volatility function), especially around the boundary of the covariate support. See also discussions of simulation results in Section 6.1.

correlation among responses) and thus does not support simultaneous inference of treatment effects across categories, which is of ultimate policy interest. Furthermore, if the bandwidths are data-dependent they are likely to be chosen differently for response probabilities across categories, which further complicates the inference problem. For example, Shigeoka (2014) exploited the discontinuity in patient cost sharing at age 70 in Japan and found the effect of cost sharing on outpatient visits is likely to be higher for non-life-threatening diagnoses (e.g. those requiring treatment to enhance the quality of life, like skin diseases) than more serious diagnoses (like cancer and heart disease). It thus would be informative to jointly test the significance across all diagnoses or the significance across more serious diagnoses.[5]

In response to the growing volume of applications and spreading empirical concerns among practitioners, in this paper we provide a systematic statistical analysis of the RD design with categorical outcomes. It is built up on a fully-nonparametric extension (Section 2) of the well-known multinomial logit (MNL) model (McCullagh and Nelder, 1989) by permitting the index function of unknown form (instead of being of the linear form). The model facilitates the joint one-step estimation of the treatment effects (Section 3) once the bandwidth around the cutoff is determined, and naturally supports the joint inference across categories. Based on the theoretical analysis, we provide algorithms for optimal bandwidth selection (Section 4) and constructing robust confidence regions for treatment effects when the optimal bandwidth is used (Section 5), thereby extending the work by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) which focus on the local linear estimator of a single treatment effect.

Our proposed RD estimator is easy to implement. In its simplest version, it amounts to fitting a standard multinomial logit model using the observations in the left and right neighborhoods of the cutoff. The core algorithms for implementation, i.e calculating the RD estimators and robust confidence regions with the optimal bandwidths, are summarized in Algorithms I&II with the choice of linear approximation of baseline-category logits and the uniform kernel. The MATLAB and R program codes are available as a supplement to the paper.[6] We also provide two empirical

[5] There is a small literature on system nonparametric and semiparametric models, however, no articles is tailored to RD designs. Henderson, et al. (2015) considered the smooth-coefficient multiple-equation model, which includes nonparametric system model with a single covariate. Their focus on the local constant estimator and the bandwidth selection based on cross validation do not benefit RD applications, in which interests are only in boundary points.

[6] The codes are available on the author's website (sites.google.com/site/xukeli2015).

examples that study the effects of the infant very-low-birth-weight classification and being put in the probation status in college, together with simulation experiments (Section 6). Several relevant and potential extensions of the current work are discussed in Section 7. Technical details are relegated to Appendices A-C.

Recent work by Koch and Racine (2016) also discussed potential issues with the standard procedures applied to the RD design with the multinomial outcome. They proposed a RD estimator which is based on estimating the joint density of the (discrete) multinomial outcome and the (continuous) forcing variable, and they advocated to fit a local constant to ensure the estimated probabilities belong to the unit interval. Like our approach, theirs imposes no structure that the data must obey, and the estimands are nonparametrically identified. However, their estimator does not come with the bandwidth which is optimal for the RD estimand, the robust inference when the optimal bandwidth is used, or any type of simultaneous inference.

2 Categorical outcomes in the RD design

Suppose that for an individual i, the outcome $\widetilde{Y}_i$ takes on the value that belongs to $(J+1)$ mutually exclusive categories, i.e. $\widetilde{Y}_i \in \{0, 1, 2, .., J\}$. Since the outcomes may be unordered, or unevenly-spaced if the outcomes are ordered, $\mathbb{E}(\widetilde{Y}_i)$ and $\text{Var}(\widetilde{Y}_i)$ may not have useful meanings.

Consider the typical (sharp) RD design, in which the binary treatment T_i is driven by the running variable $X_i \in \mathcal{X} \subset \mathbb{R}$ together with a cutoff c, i.e. $T_i = \mathbb{I}(X_i \geq c)$.[7] Assume X_i is continuous. Let $\widetilde{Y}_i(1)$ and $\widetilde{Y}_i(0)$ be two potential outcomes for the treated ($T_i = 1$) and controlled groups ($T_i = 0$), respectively. We observe $\widetilde{Y}_i = T_i\widetilde{Y}_i(1) + (1 - T_i)\widetilde{Y}_i(0)$.

For $j = 0, 1, ..., J$, the conditional (potential) outcome probabilities for two groups are

$$\mathbb{P}(\widetilde{Y}_i(1) = j|X_i = x) = \mu_{+,j}(x) \tag{1}$$

$$\mathbb{P}(\widetilde{Y}_i(0) = j|X_i = x) = \mu_{-,j}(x) \tag{2}$$

where continuous functions $\{\mu_{+,1}(\cdot), ..., \mu_{+,J}(\cdot)\}$ and $\{\mu_{-,1}(\cdot), ..., \mu_{-,J}(\cdot)\}$ are unknown, and $\mu_{+,0}(\cdot)$

[7] Here we follow the convention in the sharp RD literature (e.g. Imbens and Lemieux, 2008) to denote the units on the right side of the cutoff as the treated group. If the units on the left side are easier to interpreted as the treated group (i.e. $T_i = \mathbb{I}(X_i < c)$, as in two empirical examples in Section 6), the treatment effects become the negatives of τ_j's (defined in (4) below).

and $\mu_{-,0}(\cdot)$ (the outcome probability for the base category $j=0$) satisfy $\sum_{j=0}^{J}\mu_{+,j}(x)=1$ and $\sum_{j=0}^{J}\mu_{-,j}(x)=1$, respectively, for any $x\in\mathcal{X}$. As in most of the RD literature, we do not impose global restrictions on unknown functions μ's.

The objects of primary interest here are τ_j's defined in (3) below:

$$\begin{aligned}\tau_j &\triangleq \mathbb{P}(\widetilde{Y}_i(1)=j|X_i=c)-\mathbb{P}(\widetilde{Y}_i(0)=j|X_i=c) &(3)\\ &= \mu_{+,j}(c)-\mu_{-,j}(c) &(4)\\ &= \lim_{x\to c+}\mathbb{P}(\widetilde{Y}_i=j|X_i=x)-\lim_{x\to c-}\mathbb{P}(\widetilde{Y}_i=j|X_i=x), &(5)\end{aligned}$$

for $j=1,...,J$. They are interpreted as the ATEs (average treatment effects) at c of the treatment T_i, i.e. the change of the probability of belonging to category j by the treatment, for units taking X_i-value at c. They are identified as in (5) under standard RD assumptions, i.e. $\mu_{+,j}(x)$ and $\mu_{-,j}(x)$ are continuous at $x=c$ (Assumptions A1-A2 in Appendix A); see also Hahn, et al. (2001), Imbens and Lemieux (2008). In what follows, we develop a procedure of estimating τ_j's and conducting hypothesis testing (e.g. their joint significance). Let $\tau=(\tau_1,...,\tau_J)^\top$.

In view of (5), for each j, we are to estimate two conditional probability functions at boundaries. We find the following MNL (multinomial logit) transformation convenient. There exist functions $\{g_{+,j}(x),g_{-,j}(x):1\le j\le J\}$ (which may be referred to as baseline-category logits) such that (in a right and left neighborhood $x=c$, respectively)

$$\begin{aligned}\mu_{+,j}(x) &= \frac{\exp(g_{+,j}(x))}{1+\sum_{j=1}^{J}\exp(g_{+,j}(x))} &(6)\\ \mu_{-,j}(x) &= \frac{\exp(g_{-,j}(x))}{1+\sum_{j=1}^{J}\exp(g_{-,j}(x))} &(7)\end{aligned}$$

and $\mu_{+,0}(x)=[1+\sum_{j=1}^{J}\exp(g_{+,j}(x))]^{-1}$, $\mu_{-,0}(x)=[1+\sum_{j=1}^{J}\exp(g_{-,j}(x))]^{-1}$. Their existence is guaranteed by Assumption A3 in Appendix A.[8] It reduces to the familiar binary logistic transformation when the outcome is binary.

[8] Under Assumption A3, the functions $g_{+,j}(x)$ and $g_{-,j}(x)$ are nonparametrically identified as (for $1\le j\le J$)

$$\log\frac{\mu_{+,j}(x)}{\mu_{+,0}(x)}=g_{+,j}(x) \text{ and } \log\frac{\mu_{-,j}(x)}{\mu_{-,0}(x)}=g_{-,j}(x),$$

for x in a right and left neighborhood $x=c$, respectively.

The transformations in (6) and (7) are purely statistically motivated (McCullagh and Nelder, 1989, Ch.5), and they do not impose additional structure on the data generating process.[9] Moreover, they do not require the outcomes to be ordered. Since an unknown function is involved in each category probability, the model is most computationally tractable when J is small. When J is moderate or large, a more structured yet flexible model for the RD design might be used, e.g. exploring the ordinal structure of the outcome when available; see Section 7 for further discussions.

The models (1)-(2) and (6)-(7) extend the conventional linear MNL model for treated and controlled groups (in which the functions g's take the linear form, McCullagh and Nelder, 1989) to allow flexible functional forms of g's. As in the linear MNL model, a computational benefit of transforming μ's to g's using (6)-(7) is that the likelihood maximization, which is now implemented over the space of functions g, is unconstrained (always producing outcome probabilities within the unit interval).[10]

3 Local multinomial logit estimator

In this section we describe the local likelihood estimator of τ_j in (4), which is formally defined in (12) below. The multinomial outcome can be written as a sequence of binary outcomes $Y_{ij} = \mathbb{I}(\widetilde{Y}_i = j)$. By mutual exclusion $\sum_{j=0}^{J} Y_{ij} = 1$ for any i. Denote $Y_i = (Y_{i1}, ..., Y_{iJ})^\top$.

Denote the ν-th order (for $\nu \geq 0$) derivatives as $g_{+,j}^{(\nu)}(x)$ and $g_{-,j}^{(\nu)}(x)$. We will write $g_{+,j}^{(\nu)}(c)$ and $g_{-,j}^{(\nu)}(c)$ simply as $g_{+,j}^{(\nu)}$ and $g_{-,j}^{(\nu)}$.

For a value $X \in \mathcal{X}$ which is in a right neighborhood of c, given that the function is smooth enough (Assumption A2), $g_{+,j}(X)$ can be approximated by a p-th ($p \geq 0$) order polynomial

$$g_{+,j}(X) \approx \overline{g}_{+,j}(c, X) = g_{+,j} + (X-c)g_{+,j}^{(1)} + ... + (X-c)^p g_{+,j}^{(p)}/p! = \overline{X}^\top \beta_{+,j}^*, \tag{8}$$

where $\overline{X} = (1, X-c, ..., (X-c)^p)^\top$ and $\beta_{+,j}^* = (\beta_{+,j,0}^*, \beta_{+,j,1}^*, ..., \beta_{+,j,p}^*)^\top = (g_{+,j}, g_{+,j}^{(1)}, ..., g_{+,j}^{(p)}/p!)^\top$.

Let $K(\cdot) \geq 0$ be a symmetric kernel function supported on $[-1, 1]$ and $h > 0$ be a bandwidth

[9] As in the RD literature we here follow a reduced-form approach; we view (6) and (7) nothing more than one-to-one mappings, as opposed to being generated by any economic theory.

[10] Another application of the local likelihood approach is explored by Otsu et al. (2013). They considered estimation and testing of the discontinuity in density, and found the approach works remarkably well when compared with the conventional binning approach, in the context of testing for manipulation in the RD design (McCrary, 2008).

parameter (as in standard nonparametric estimation; Fan and Gijbels, 1996). Denote

$$\widehat{\beta}_+ = (\widehat{\beta}_{+,1}^\top, ..., \widehat{\beta}_{+,J}^\top)^\top = \arg \max_{\beta \in \mathbb{R}^{(p+1)J}} L_+(\beta), \tag{9}$$

where $\widehat{\beta}_{+,j} = (\widehat{\beta}_{+,j,0}, \widehat{\beta}_{+,j,1}, ..., \widehat{\beta}_{+,j,p})^\top \in \mathbb{R}^{p+1}$, and

$$L_+(\beta) = \sum_{i=1}^n \ell(\overline{X}_i^\top \beta_1, ..., \overline{X}_i^\top \beta_J; Y_{i1}, ..., Y_{iJ}) K((X_i - c)/h) I_i, \tag{10}$$

with $I_i = \mathbb{I}(X_i \geq c)$ and

$$\ell(g_1, ..., g_J; y_1, ..., y_J) = \sum_{j=1}^J y_j g_j - \log(1 + \sum_{j=1}^J \exp(g_j)).$$

Here we have used $\sum_{j=0}^J Y_{ij} = 1$. The criterion function $L_+(\beta)$ is obtained by using localization (in the right h-neighborhood of the cutoff) and the approximation (8) in the infeasible log likelihood $\sum_{i=1}^n \sum_{j=0}^J Y_{ij} \log \mu_{+,j}(X_i)$.

Alternatively, the global series-based method can also be used to approximate the functions $g_{+,j}(\cdot)$ but it is more computationally taxing (than the local method) in the likelihood setting here with multiple unknown functions.

Similarly, denote

$$\widehat{\beta}_- = \arg \max_{\beta \in \mathbb{R}^{(p+1)J}} L_-(\beta), \tag{11}$$

where

$$L_-(\beta) = \sum_{i=1}^n \ell(\overline{X}_i^\top \beta_1, ..., \overline{X}_i^\top \beta_J; Y_{i1}, ..., Y_{iJ}) K((X_i - c)/h)(1 - I_i).$$

Local likelihood estimators of $g_{+,j}^{(\nu)}$ and $g_{-,j}^{(\nu)}$ (for $\nu = 0, 1, ..., p$) are defined as $\widehat{g}_{+,j}^{(\nu)} = \nu! \widehat{\beta}_{+,j,\nu}$ and $\widehat{g}_{+,j}^{(\nu)} = \nu! \widehat{\beta}_{-,j,\nu}$. The associated estimator of τ (referred to as local MNL or local likelihood estimator with order p):

$$\widehat{\tau}_j = \widehat{\mu}_{+,j} - \widehat{\mu}_{-,j}, \tag{12}$$

where $\widehat{\mu}_{+,j} = [1 + \sum_{j=1}^J \exp(\widehat{g}_{+,j})]^{-1} \exp(\widehat{g}_{+,j})$ and $\widehat{\mu}_{-,j} = [1 + \sum_{j=1}^J \exp(\widehat{g}_{-,j})]^{-1} \exp(\widehat{g}_{-,j})$.

Our recommended estimator of τ is based on $p = 1$. A higher order local likelihood optimization

is often needed in optimal bandwidth determination (Section 4).

We now describe an approximation of the distribution for $\widehat{\tau}$. Let $g_+^{(\nu)} = (g_{+,1}^{(\nu)}, ..., g_{+,J}^{(\nu)})^\top$ and $g_-^{(\nu)} = (g_{-,1}^{(\nu)}, ..., g_{-,J}^{(\nu)})^\top$. Let Υ_+ be $J \times J$ matrix with (j, j')-element $\mu_{+,j}(\delta_{jj'} - \mu_{+,j'})$, where $\delta_{jj'} = \mathbb{I}(j = j')$. Define Υ_- similarly. Let $f(x)$ be the density of X_i at x.

Define $N_p = (\int_0^1 u^{i+j-2} K(u)\mathrm{d}u)_{i,j=1,...,p+1} \in \mathbb{R}^{(p+1)\times(p+1)}$, and the equivalent kernel function (as defined in Fan and Gijbels, 1996, (3.16), p.64) $K_{p,\nu}^*(u) = e_{p+1,\nu+1}^\top N_p^{-1}(1, u, ..., u^p)^\top K(u)$, where $e_{p,\nu}$ is the $p \times 1$ vector with the ν-th element one and others zero. Two sets of kernel-specific constants that will play important roles are $c_{B,p,\nu} = \int_0^1 u^{p+1} K_{p,\nu}^*(u)\mathrm{d}u$ and $c_{V,p,\nu} = \int_0^1 K_{p,\nu}^{*2}(u)\mathrm{d}u$ (for $\nu = 0, 1, ..., p$).

Let $B = c_{B,p,0} h^{p+1}(\Upsilon_+ g_+^{(p+1)} - \Upsilon_- g_-^{(p+1)})/(p+1)!$ and $V = (\Upsilon_+ + \Upsilon_-)c_{V,p,0}/f(c)$. Under the assumptions of Theorem 3 in Appendix A,

$$\widehat{\tau} - B \overset{Appox.}{\sim} \mathcal{N}(\tau, (nh)^{-1}V), \tag{13}$$

where $\overset{Appox.}{\sim}$ denotes "approximately distributed as" in the asymptotic sense ($n \to \infty$ and $h = h(n) \to 0$). In other words, the distribution of $\widehat{\tau}$ is approximately normal around the true value with the bias B and the variance $(nh)^{-1}V$.

A good balance between the computational tractability and reasonable bias and variance properties leads to the popular choice of the polynomial order $p = 1$; as p increases, the bias gets smaller (as a higher order of h) and the variance gets larger (as $c_{V,p,0}$ increases with p). The correlations among $\widehat{\tau}_j$'s are reflected by off-diagonal elements of V.

The asymptotic properties of the local likelihood estimator were studied by Fan, Heckman and Wand (1995) within the likelihood class of the canonical exponential family with the univariate response. Our analysis here extends their results to multivariate responses.

4 Bandwidth selection

It is an empirical norm that the RD estimate is sensitive to the choice of the smoothing bandwidth. The robustness check with a range of bandwidths is useful but a single treatment effect estimate (optimal in a sense) is definitely needed to quantify the results. In this section we consider practical

bandwidth determination, and focus on the local likelihood estimator of τ based on local linear ($p = 1$) approximation.

We assume a common bandwidth for both sides of the cutoff. The bandwidth is selected in an optimal way jointly for response probabilities of J categories, i.e. it minimizes the norm of the vector AMSE (asymptotic mean squared error $C(h)$, obtained from (13)). By definition, $h_{\text{opt}} = \arg\min_{h>0} C(h)$, where

$$\begin{aligned} C(h) &= ||B||^2 + \text{trace}(V) \\ &= c_{B,1,0}^2 h^4 ||\Upsilon_+ g_+^{(2)} - \Upsilon_- g_-^{(2)}||^2/4 + \text{trace}(\Upsilon_+ + \Upsilon_-) c_{V,1,0}/[nhf(c)], \end{aligned}$$

with $||\cdot||$ being the Euclidean norm for a vector and trace$(\cdot)$ denoting the trace of a matrix. Thus

$$h_{\text{opt}} = \left(\frac{c_{V,1,0}}{c_{B,1,0}^2 nf(c)} \frac{\text{trace}(\Upsilon_+ + \Upsilon_-)}{||\Upsilon_+ g_+^{(2)} - \Upsilon_- g_-^{(2)}||^2} \right)^{1/5}. \tag{14}$$

With unknown quantities in (14) being estimated, we obtain the feasible optimal bandwidth

$$\widehat{h} = \left(\frac{c_{V,1,0}}{c_{B,1,0}^2 n\widehat{f}(c)} \frac{\text{trace}(\widetilde{\Upsilon}_+ + \widetilde{\Upsilon}_-)}{||\widetilde{\Upsilon}_+ \widehat{g}_+^{(2)} - \widetilde{\Upsilon}_- \widehat{g}_-^{(2)}||^2} \right)^{1/5}. \tag{15}$$

Here the estimates $\widetilde{\Upsilon}_+$, $\widetilde{\Upsilon}_-$, $\widehat{g}_+^{(2)}$ and $\widehat{g}_-^{(2)}$ are obtained by preliminary local likelihood regressions with pilot bandwidths, as described in the following algorithm (Steps 4 and 5). For computational simplicity, we use the uniform kernel $K(u) = \mathbb{I}_{[-1,1]}(u)/2$.

Algorithm I (Optimal Bandwidth)

1. Apply the standard estimator to obtain $\widehat{f}(c)$ (using continuity of $f(x)$ at $x = c$; see e.g. Imbens and Kalyanaraman, 2012, page 941, Step 1).

2. Run global cubic ($p = 3$) MNL fits to obtain preliminary estimates $\breve{\Upsilon}_+$, $\breve{\Upsilon}_-$, $\breve{g}_+^{(2)}$, $\breve{g}_-^{(2)}$, $\breve{g}_+^{(3)}$ and $\breve{g}_-^{(3)}$.

3. Calculate pilot bandwidths

$$\widetilde{h}_+ = \left(\frac{c_{V,1,0}}{c_{B,1,0}^2 n\widehat{f}(c)} \frac{\text{trace}(\breve{\Upsilon}_+)}{||\breve{\Upsilon}_+ \breve{g}_+^{(2)}||^2} \right)^{1/5}, \quad \widetilde{h}_- = \left(\frac{c_{V,1,0}}{c_{B,1,0}^2 n\widehat{f}(c)} \frac{\text{trace}(\breve{\Upsilon}_-)}{||\breve{\Upsilon}_- \breve{g}_-^{(2)}||^2} \right)^{1/5} \tag{16}$$

$$\widehat{b}_+ = \left(\frac{90 c_{V,2,2}}{c_{B,2,2}^2 n \widehat{f}(c)} \frac{\text{trace}(\breve{\Upsilon}_+^{-1})}{||\breve{g}_+^{(3)}||^2} \right)^{1/7}, \quad \widehat{b}_- = \left(\frac{90 c_{V,2,2}}{c_{B,2,2}^2 n \widehat{f}(c)} \frac{\text{trace}(\breve{\Upsilon}_-^{-1})}{||\breve{g}_-^{(3)}||^2} \right)^{1/7}. \tag{17}$$

For the uniform kernel, $(c_{V,1,0}/c_{B,1,0}^2)^{1/5} = 2.702$ and $(90 c_{V,2,2}/c_{B,2,2}^2)^{1/7} = 3.557$.

4. Run the local likelihood regression with linear ($p = 1$) approximation, using pilot bandwidths $\widetilde{h}_+$ and $\widetilde{h}_-$, to obtain $\widetilde{\Upsilon}_+$ and $\widetilde{\Upsilon}_-$.

5. Run the local likelihood regression with quadratic ($p = 2$) approximation, using pilot bandwidths $\widehat{b}_+$ and $\widehat{b}_-$, to obtain $\widehat{g}_+^{(2)}$ and $\widehat{g}_-^{(2)}$.

6. Obtain the feasible optimal bandwidth $\widehat{h}$ according to (15).

The optimal bandwidth in (15) and pilot bandwidths in (16) and (17) for the local MNL estimator with a general order p is given in Appendix A. Note that the pilot bandwidths in Step 3 have the same rates as the optimal bandwidths to estimate Υ_+, Υ_-, $g_+^{(2)}$ and $g_-^{(2)}$, and the scale constants are reasonably approximated (although they may not be consistently estimated, since the quantities in Step 2 are in general inconsistently estimated). The formulas in (16) and (17) follow from the balances between biases and variances when estimating g's (thereby Υ's) and their second derivatives (see Theorem 2 in Appendix A).

We show (see Theorem 4 in Appendix A) that $\widehat{h}$ is asymptotically as good as h_{opt} with an error rate $O_p(n^{-1/7})$, i.e.

$$(\widehat{h} - h_{\text{opt}})/h_{\text{opt}} = O_p(n^{-1/7}). \tag{18}$$

An alternative way to select the bandwidth which does not completely rely on the asymptotic arguments above might be achieved following the line of Fan, Farmen and Gijbels (1998). In their approach, the estimated bias and variance involve the Jacobian and Hessian of the objective function in which preliminary estimators of g and its second derivative are required. Pilots bandwidths are thus also needed, which is a non-trivial undertaking. We find the asymptotic method we take here is easier to understand and establish optimal properties.

Focusing on the local linear estimator of the RD design, Arai and Ichimura (2015) allowed different bandwidths on each side of the cutoff. The selected optimal pair of bandwidths yields the AMSE of a smaller order of n (than the one under the assumption of one single bandwidth) when the derivatives of two regression functions at the cutoff have the same sign. This method is

extendable to our setting with the local MNL estimator when $J = 1$.

However, when the outcome has more than two discrete levels ($J \geq 2$), there does not exist a ratio of two bandwidths such that the leading bias components are removed unless the vector $\Upsilon_+ g_+^{(2)} - \Upsilon_- g_-^{(2)}$ contains identical elements. Thus the idea of selecting bandwidths based on a smaller order bias component (which produces the AMSE gain of an order of n), as pursued by Arai and Ichimura (2015), does not apply here when $J \geq 2$. The contrast is due to the multi-dimensional nature of the outcome when $J \geq 2$ while the smoothing is implemented in only one dimension, as evident in (10). Having mentioned that, there could be potential AMSE gain (up to a fixed constant) by allowing different bandwidths on each side when $J \geq 2$. We assume the same bandwidth in our development for its implemental simplicity.

5 Robust inference

In this section we construct the (simultaneous) inference of τ based on its local MNL estimator with order p. If the optimal bandwidth developed above is used, the bias has to been corrected. A simple approach is thus based on the result (13) and consistent estimation of the terms B and V.

Let $\widehat{B}_{g,+} = c_{B,p,0} h^{p+1} \widehat{g}_+^{(p+1)}/(p+1)!$, where $\widehat{g}_+^{(p+1)}$ is obtained by maximizing the local likelihood of order $(p+1)$ using the bandwidth b_+. The term $\widehat{B}_{g,-}$ is similarly defined with the bandwidth b_-. Suppose b_+ and b_- are selected with the optimal rate in estimating $g_+^{(p+1)}$ and $g_-^{(p+1)}$ respectively. Let $\widehat{\Upsilon}_+$ and $\widehat{\Upsilon}_-$ be the estimators of Υ_+ and Υ_- respectively using the optimal bandwidth $\widehat{h}$ in Algorithm I. Construct the correction term $\widehat{B}$ for the bias in $\widehat{\tau}$ by

$$\widehat{B} = \widehat{\Upsilon}_+ \widehat{B}_{g,+} - \widehat{\Upsilon}_- \widehat{B}_{g,-}. \tag{19}$$

Then (13) becomes

$$\widehat{\tau} - \widehat{B} \overset{Appox.}{\sim} \mathcal{N}(\tau, (nh)^{-1} V). \tag{20}$$

The approximation (20) can be justified (in Appendix) in our setting which ensures $h/\min(b_+, b_-) \to 0$. Inference based on (20) can be implemented by replacing V with the estimator $\widehat{V} = \widehat{f}(c)^{-1}(\widehat{\Upsilon}_+ + \widehat{\Upsilon}_-) c_{V,p,0}$.

It is well known that the approximation (20) may perform poorly in finite samples since the

variance does not accommodate the added variability caused by replacing B with $\widehat{B}$.[11] Extending the idea pursued by Calonico et al. (2014), we here provide a new approximation which is suitable for such bias correction.[12]

Let $K^*_{p,0}(\cdot)$ be the equivalent kernel function defined in Section 3. Denote

$$r_+ = c_{V,p,0} + \left(\frac{h}{b_+}\right)^{2p+3} \varphi_1 + \left(\frac{h}{b_+}\right)^{p+2} \varphi_2, \tag{21}$$

where $\varphi_1 = c_{V,p+1,p+1}[(p+1)!]^2 c^2_{B,p,0}$ and $\varphi_2 = -2[K(0)e^\top_{p+2,1}N^{-1}_{p+1}e_{p+2,p+2}]c_{B,p,0}$. Similarly r_- is defined. The constants φ_1 and φ_2 only depend on the kernel function and p. Their values for $p = 1$ and three commonly used kernels are given in Table 1, as well as other kernel constants which are useful in implementation. Define $\Omega = f(c)^{-1}(\Upsilon_+ r_+ + \Upsilon_- r_-)$.

Under the assumptions of Theorem 6 in Appendix A,

$$\widehat{\tau} - \widehat{B} \overset{Appox.}{\sim} \mathcal{N}(\tau, (nh)^{-1}\Omega). \tag{22}$$

The additional two terms in (21) (without which we would have $r_+ = c_{V,p,0}$, thus $V = \Omega$) capture the smaller-order variance which is caused by bias correction. (22) is operational if Ω is replaced by the estimator $\widehat{\Omega} = \widehat{f}(c)^{-1}(\widehat{\Upsilon}_+ r_+ + \widehat{\Upsilon}_- r_-)$. Inference based on (22) (which is referred to as robust inference) is expected to be more accurate than that based on (20) (referred to as standard inference) in finite samples.[13]

Focusing on the local linear estimator, Calonico et al. (2014) provided the variance estimate which is robust to a wider range of b_+ and b_- (e.g. b_+ and b_- are allowed to be of same magnitude as or even smaller than h). We here only consider the bandwidths (used to estimate the $(p+1)$-th derivatives) with the MSE-optimal rate which naturally leads to larger bandwidths than h.

We summarize the results and provide below the algorithm for empirical implementation using

[11] For this reason, it is generally recommended in the literature of kernel smoothing and local polynomial methods to ignore the bias, which is arguably justified by undersmoothing (thus with costs of subjectivity and forfeiting the optimal bandwidth), when conducting inference; see e.g. Owen (2001, Ch. 5), Imbens and Lemieux (2008, p. 630).

[12] In recent work, Calonico, Cattaneo, and Farrell (2017) showed that, for local polynomial estimator, the robust bias-corrected confidence interval of Calonico, et al. (2014) leads to a faster coverage error decay rate than the confidence interval based on undersmoothing.

[13] Calonico et al. (2017) discussed other benefits of bias-corrected confidence intervals, using appropriate standard errors, based on local polynomial estimators in a general setting of nonparametric regression. They also developed a coverage-error optimal bandwidth. It would be worthwhile to investigate whether their approach extends to the current local likelihood setting in future work.

Table 1: Kernel constants (φ_1 and φ_2 are for $p = 1$).

	$K(u)$	$c_{B,1,0}$	$c_{V,1,0}$	$c_{B,2,2}$	$c_{V,2,2}$	φ_1	φ_2
Uniform	$\frac{1}{2}\mathbb{I}_{[-1,1]}(u)$	-0.167	4	1.5	180	20	10
Triangular	$(1-\lvert u\rvert)\mathbb{I}_{[-1,1]}(u)$	-0.100	4.8	1.286	308.6	12.34	12
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbb{I}_{[-1,1]}(u)$	-0.116	4.5	1.328	266.6	14.30	11.15

$p = 1$ and the uniform kernel.

Algorithm II (Point Estimator, Confidence region and Wald test)

1. Use the optimal bandwidth $\widehat{h}$ (which is determined in Algorithm I) to solve the optimization problems in (9) and (11) to obtain $\widehat{g}_+$ and $\widehat{g}_-$.

2. Obtain the point estimate of ATE $\widehat{\tau}$ according to (12).

3. Calculate $\widehat{B}_{g,+}$ and $\widehat{B}_{g,-}$, where $\widehat{g}_+^{(2)}$ and $\widehat{g}_-^{(2)}$ are obtained as in Algorithm I (Step 5, with the bandwidth $\widehat{b}_+$), and the constant $c_{B,1,0} = -1/6$. Obtain $\widehat{B}$ in (19), where $\widehat{\Upsilon}_+$ and $\widehat{\Upsilon}_-$ are obtained using the estimates in Step 1.

4. Obtain $\widehat{\Omega}$, where $\widehat{f}(c)$ is obtained as in Algorithm I (Step 1), $r_+ = 4+20(\widehat{h}/\widehat{b}_+)^5+10(\widehat{h}/\widehat{b}_+)^3$ and $r_- = 4 + 20(\widehat{h}/\widehat{b}_-)^5 + 10(\widehat{h}/\widehat{b}_-)^3$.

5. Obtain the robust Wald test of $\mathcal{H}_0 : R\tau = q$, where R and q contain known constants, with dimensions $d_q \times J$ and $d_q \times 1$ respectively. The test rejects $\mathcal{H}_0$ at the level α if

$$\mathcal{W} = (nh)(R\widehat{\tau} - q - R\widehat{B})^\top (R\widehat{\Omega}R^\top)^{-1}(R\widehat{\tau} - q - R\widehat{B}) > \chi^2_{1-\alpha}(d_q), \tag{23}$$

where $\chi^2_{1-\alpha}(d_q)$ is the $(1-\alpha)$-th quantile of the chi-square distribution with the degree of freedom d_q.

6. Obtain the robust $100(1-\alpha)\%$ confidence region for $R\tau$ as $\mathrm{CR}_{R\tau} = \{q \in \mathbb{R}^{d_q} : \mathcal{W} \leq \chi^2_{1-\alpha}(d_q)\}$, where $\mathcal{W}$ is in (23).

Steps 5 and 6 can be used for inference of individual τ_j and simultaneously for τ. The standard (or non-robust) confidence region and the Wald test are obtained by modifying the algorithm above

using $\widehat{V}$ in place of $\widehat{\Omega}$ in Steps 5 and 6 (i.e. using $r_+ = r_- = c_{V,1,0} = 4$ in Step 4).

The algorithm above also utilizes the fact that the conditional variance matrix (Υ_+ or Υ_-) of the outcome depends only on the conditional probabilities (a feature offered by the multinomial outcome), which guarantees the conditional variance is estimated with the optimal bandwidth. Conditional variance estimation is known to cause some discretion in the formation of confidence region for the ATE. Imbens and Kalyanaraman (2012, equations (12) and (13)) simply used the bandwidth which is designed for estimation of the unknown density function. Calonico et al. (2014, Section 5) noticed the sensitivity of conditional variance estimation (thereby inference of the ATE) to the bandwidth and found that the method of nearest neighborhoods might be more robust, although the number of neighbors remains to be determined in an empirical application.

Complementary to the *pointwise* confidence region (like the one constructed above for individual τ_j or simultaneously for τ) for a given bandwidth, Armstrong and Kolesár (2015) proposed, for sensitivity analysis, forming the confidence band for the ATE which is uniformly valid over a range of bandwidths (usually around the optimal bandwidth $\widehat{h}$). The idea might be also exploited in the current setting although it requires nontrivial work for two reasons. First, the verification of the high-level conditions in their paper applies only to the local polynomial estimator instead of the local likelihood estimator as we pursue here. Second, the high-level conditions in their paper are given for a univariate outcome thus have to be extended to facilitate uniform confidence region construction in the Euclidean space as in our setting.

6 Applications

In this section, we consider two empirical applications with categorical outcomes in the RD design, together with some simulation experiments, to illustrate the finite-sample performance and the practical relevance of proposed statistical procedures.

6.1 Infant mortality vs. very-low-birth-weight status

The background of this quasi-experimental design can be found in Almond, Doyle, Kowalski and Williams (2010). The design exploits the variation in medical inputs generated by the very-low-birth-weight (VLBW) classification of newborns at 1.5 kilograms. The forcing variable is thus the

birth weight. We here consider two mutually exclusive types of one-year mortality as the outcome variable: infants that survive for more than one day (but less than one year) ($j = 2$) and infants that do not ($j = 1$). Infants that survive for more than one year belong to the base category (so $J = 2$).

Table 2 gives the local MNL estimates of mortality effects of being classified as VLBW status for two most populated states in the U.S., California and Texas, during the years of 2000-2001. We restrict the sample to infants with birth weights between 0.7 and 2.3 kilograms. This consideration leaves us the largest possible samples for two states which contain observations symmetric around the cutoff 1.5 kilograms, in line with our theoretical assumption of the same bandwidth used on both sides. We end up with sample sizes $n = 40,750$ and $n = 34,489$ for California and Texas, respectively. The optimal bandwidths are estimated as 0.237 and 0.336 for two states, using Algorithm I.

For California, the mortality effects (of VLBW status) are estimated as 0.25% and -1.56% for two categories of mortality.[14] The effect is insignificant (with the p-value 0.64) for the one-day mortality, and is highly significant (with the p-value 0.02) for the later-term mortality (i.e. being VLBW classified significantly reduces the later-term mortality), according to robust standard errors. The usual standard error (based on (20)) tends to produce more significant results, but as warned in our simulation experiments reported below, it is likely to be too small (or the associated tests tend to over-reject).

The mortality effects for Texas are similar in terms of being in opposite directions for two categories, but are less contrastive in magnitude, compared to California. For both states, the joint effects of the VLBW status across categories are significant at 10% level (but not at 5% level), again using the more reliable robust tests. The 95% confidence regions (non-robust and robust) for (τ_1, τ_2) are plotted in Figure 3 for each state to visualize the results. Table 2 also shows that the equivalence of effects for earlier-term and later-term mortality rates is rejected at 5% level for each state.

The results indicate that medical treatments are (significantly) more effective in improving (i.e. lowering) the later-term mortality. They are much less beneficial when the death occurs within

[14] The values of τ's are reported in Table 2, and their negatives are interpreted as VLBW treatment effects since $T_i = \mathbb{I}(X_i < c)$.

24 hours (the riskiest period of a child's life), possibly due to the reasons with which additional medical treatments can not help much, such as prematureness[15] (e.g. for mothers who do not get access to quality prenatal care or go to the doctor on a regular basis), the fatal illness at birth or lack of time for effective medical treatments. Such a result appears to be absent from previous studies on medical treatment effects on infant mortality. The negative τ-estimate (i.e. positive one-day mortality treatment effect) seems counter-intuitive, however, it is more or less understandable as a finite-sample effect if the true effect is close to zero, given that the mortality rate is likely to be a decreasing function of the infant birth weight (Figures 1 and 2 for two states) and the estimating sample bandwidth is fairly large. This reinforces the statistical insignificance of the one-day mortality rate effect. It is noteworthy that first-day infant deaths account for 40.4% and 31.0% of total deaths for two states respectively within the samples local to the cutoff, so the results bear important policy indications.

We also perform the robustness check by dropping the observations with the birth weights between 1.485 and 1.515 kilograms (those at the cutoff and within its 15-gram neighborhoods, about 1.1% and 1.3% of total observations for two states respectively); see Table 3. Such donut RD estimates, as advocated by Barreca, Guldi, Lindo and Waddell (2011), are useful to examine robustness to non-random bunching at the cutoff (due to rounding errors for potentially low-quality hospitals) and manipulation by hospitals or parents. The significant effect (at 10% level) of the later-term mortality and insignificant effect of the first-day mortality remain (although the joint significance is reduced, compared to that when all observations in windows are used).

This application also serves well as an example showing that combining categories may obscure the significance of effects. Table 4 provides the results when only one-year mortality is considered (thus two non-reference categories considered above are combined). Although the effects of VLBW classification are still sizable in improving mortality rates for both states (especially for California, approximately 1%), they become insignificant. Almond et al. (2010) used a sequence of binary outcomes which combine categories. For example, in the two-category (mutually exclusive) setting here, the approach they used focuses on the one-day and one-year mortalities, the latter of which essentially combines two categories we consider. Their approach is thus not able to uncover the

[15] Prematureness is often cited as the primary season that the United States has one of the highest first-day infant death rates out of all the industrialized countries in the world (Annual State of the World's Mothers report, 2012).

significant treatment effect on the one-day-to-one-year mortality (in particular, no effects is significant for two categories the approach considers, for California or Texas, using the robust standard error; Table 2 (Panel A) and Table 4), or perform any type of joint inference across categories, for neither of the two states over the given period.

We conclude this application by comparing with the *marginal* approach to causal effects τ_1 and τ_2. When the local MNL method applied to the binary outcome for each category, the results for each causal effect marginally are quantitatively slightly different from yet are qualitatively similar to the joint approach reported earlier; τ_1 is insignificant for both states and τ_2 is more (positively) significant with p-value between 5% and 10% when robust t-tests are used. The prevalent approach based on local linear fitting (with the data-determined optimal bandwidth) can be used for each binary outcome, and the results are reported in Table 6. It estimates the two effects more contrastively for Texas sample, however, it fails to find any significance (with standard or robust t-tests) for two effects for neither states.[16] We emphasize that any type of simultaneous inference is lacking in this approach no matter the local MNL or local linear fitting is used.

To recapitulate, although the significance of two types of infant-mortality effects depends on the standard error used, whether a likely-to-manipulate sample is dropped, whether a local linear v.s. local non-linear and joint v.s. marginal approach is taken, and the sample of interest (through all of which the joint test of significance is also affected), the result that the VLBW status causes a bigger and more significant reduction of the later-term mortality than the earlier-term mortality appears to be fairly robust in our analysis.

Simulations: DGPs

We now evaluate the performance of our estimation and inference procedures via simulation experiments. The designs use parametric models, aiming to mimic the infant mortality application above, with an outcome taking value from three categories. For all pseudo-data generating processes (DGPs 1-4) below, the running variable X_i is drawn from the gamma distribution with shape and scale parameters 14.25 and 0.1272.

[16] The local-linear-fitting results are obtained in R, with the command `rdrobust` in the package `rdrobust`.

We first consider the parametric MNL model. For each unit $i = 1, .., n$, let

$$\begin{aligned} Z_{i0} &= e_{i0} \\ Z_{ij} &= \mu_j(X_i) + e_{ij}, \text{for } j = 1, 2, \end{aligned}$$

where the error e_{ij} follows independent standard Type I extreme value distribution, for $j = 0, 1, 2$. The mean functions $\mu_1(x)$ and $\mu_2(x)$ follow one of the following forms:

$$\begin{aligned} \text{DGP 1: } \mu_1(x) &= \begin{cases} 15.79 - 47.25x + 34.28x^2 - 7.68x^3 \sin x, \text{if } x < 1.5; \\ -0.58 - 2.67x + 0.57x^2 - 0.20x^3 \sin x, \text{if } x \geq 1.5, \end{cases} \\ \mu_2(x) &= \begin{cases} 3.04 - 6.14x - 0.34x^2 + 0.94x^3 \sin x, \text{if } x < 1.5; \\ 41.31 - 44.38x + 4.98x^2 + 3.20x^3 \sin x, \text{if } x \geq 1.5, \end{cases} \end{aligned}$$

$$\begin{aligned} \text{DGP 2: } \mu_1(x) &= \begin{cases} 9.40 - 27.08x + 16.26x^2 - 2.72x^3 \sin x, \text{if } x < 1.5; \\ 8.76 - 12.04x + 1.07x^2 + 0.76x^3 \sin x, \text{if } x \geq 1.5, \end{cases} \\ \mu_2(x) &= \begin{cases} -6.52 + 20.61x - 21.94x^2 + 6.34x^3 \sin x, \text{if } x < 1.5; \\ 6.80 - 10.18x + 2.01x^2 + 0.16x^3 \sin x, \text{if } x \geq 1.5, \end{cases} \end{aligned}$$

where the numerical values are obtained from global multinomial fits of data in California and Texas for DGPs 1 and 2 respectively. The functional form for μ's is chosen arbitrarily and is found to fit the data very well (μ functions are plotted in Figure 4). The non-polynomial term $x^3 \sin x$ is designed so that none of steps in Algorithms 1 and 2 are based on correct global specification. The outcome is then $\widetilde{Y}_i = j$ if $j = \max_{0 \leq j' \leq 2} Z_{ij'}$.

We then consider an ordered probit model. For each unit $i = 1, .., n$, let $Z_i = \varsigma(X_i) + e_i$, where the error $e_i \sim \mathcal{N}(0, 1)$. For each of two DGPs (DGPs 3-4), the function $\varsigma(x)$ and the determination

of the outcome $\widetilde{Y}_i$ are specified as follows:

$$\text{DGP 3} \quad : \quad \varsigma(x) = \begin{cases} -14.43x + 9.29x^2 - 1.87x^3 \sin x, \text{if } x < 1.5; \\ -11.91x + 1.44x^2 + 0.77x^3 \sin x, \text{if } x \geq 1.5, \end{cases}$$

$$\widetilde{Y}_i \quad = \quad \begin{cases} \mathbb{I}(Z_i \leq 4.75) + 2\mathbb{I}(4.75 < Z_i \leq 5.36), \text{ if } x < 1.5; \\ \mathbb{I}(Z_i \leq 9.98) + 2\mathbb{I}(9.98 < Z_i \leq 10.41), \text{ if } x \geq 1.5. \end{cases}$$

and

$$\text{DGP 4} \quad : \quad \varsigma(x) = \begin{cases} -1.23x - 1.72x^2 + 0.96x^3 \sin x, \text{if } x < 1.5; \\ -4.59x + 0.80x^2 + 0.12x^3 \sin x, \text{if } x \geq 1.5, \end{cases}$$

$$\widetilde{Y}_i \quad = \quad \begin{cases} \mathbb{I}(Z_i \leq 0.17) + 2\mathbb{I}(0.17 < Z_i \leq 0.82), \text{ if } x < 1.5; \\ \mathbb{I}(Z_i \leq 2.43) + 2\mathbb{I}(2.43 < Z_i \leq 2.98), \text{ if } x \geq 1.5. \end{cases}$$

Numerical values in DGPs 3 and 4 are obtained from fitting the global parametric ordered-probit model to California and Texas data respectively.

We are able to calculate the implied value for the RD estimand, denoted by τ^* (the true value of $\tau = (\tau_1, \tau_2)$), under each of the four DGPs. While the parametric MNL estimate of τ is fairly close to the nonparametric estimate reported in the empirical analysis above for each of the two states, the parametric ordered-probit estimate is quite different from the nonparametric one. The difference comes from the restriction on functions $\mu_{+,j}(x)$ and $\mu_{-,j}(x)$ imposed by the ordered model, e.g. $\Phi^{-1}(\mu_{+,1}(x)) - \Phi^{-1}(\mu_{+,2}(x))$ being a constant, where $\Phi(\cdot)$ is the standard normal CDF. Nevertheless, DGPs 3 and 4 are useful to check the robustness of proposed nonparametric procedures.

Simulations: Results

The results are reported in Tables 7 and 8, for unordered and ordered models (DGPs 1-2, DGPs 3-4, respectively), where sample sizes considered are $n = 4000$ and $n = 8000$. Simulation results are based on 5000 replications. For all DGPs, the point estimate $\widehat{\tau}$ is quite close to the true value, with mild yet reasonable bias. Although the accuracy and precision are better for one category than the other, they generally improve with a larger sample size. Figure 5 shows two loss functions

(RMSEs, square root of mean square errors, and MADs, mean absolute deviations) for $\widehat{\tau}_1$ and $\widehat{\tau}_2$ evaluated over replications for a variety of sample sizes. They shrink as the sample size increases for all DGPs (Figure 5 only shows for DGPs 1 and 3), which shows evidence for consistency of point estimates. Compared to point estimates, the estimated bandwidth h is less accurate with a quite large standard error, as expected by its low convergence rate.

Tables 7 and 8 also report the τ^*-coverage rates of the 90% marginal confidence intervals for τ_1 and τ_2, and the joint confidence region for (τ_1, τ_2), using bias correction, with the usual standard error and the robust standard error. The robust inference is mostly quite precise, with coverage rate ranging from about 85%-90%. The occasional mild under-coverage seems reasonable given that the optimal bandwidth is estimated for each repetition with a slow convergence rate. In sharp contrast, the usual standard error yields confidence regions which are tighter than the robust ones but seriously undercover, ranging from about 68%-79%.

As in the empirical applications above and below, we adopt a uniform kernel. We also experimented with the triangular kernel, and found the performance of the point estimate is acceptable but not as good as the one with the uniform kernel (in terms of both bias and variance). This is probably due to the lack of a simple correspondence to the parametric model with local data and the potential instability of a generic optimization algorithm required, which discourages the use of such non-uniform kernel in the current context. We have also confirmed in simulations that the popular local linear estimator might generate insensible results. For example, for DGP 1, the proportion of replications in which at least one of four local linear estimators (with data-determined optimal bandwidths) of $\mu_{1,+}(c)$, $\mu_{1,-}(c)$, $\mu_{2,+}(c)$ and $\mu_{2,-}(c)$ fall out of the natural range of a probability [0,1] is about 25%, 7% and 1% for the sample size 2000, 4000 and 8000, respectively.

6.2 College dropout/graduation vs. probation status

The second application is concerned about the effects of being placed in the probation status after the first year in college on the immediate dropout and future graduation, as studied in Lindo, Sanders and Oreopoulos (2010). The design exploits the discontinuity caused by the first year GPA (the forcing variable) passing the threshold. We use the same dataset and focus on categorical outcomes. Part of the purpose of this analysis is to re-examine their results using the local likeli-

hood approach coupled the data-determined bandwidth and the associated robust standard error introduced in earlier sections. Moreover, the new multinomial approach applied to the graduation outcome uncovers the new effect of extended graduation. All together, the results here reinforce the short-term and long-term discouraging effects for students who are placed on probation after the first year in college.

The results on the immediate dropout effect are contained in Table 9. With the bandwidths adapted to the feature of the dependent variable, the neighborhoods range over about 0.4-0.9 GPA points across different samples (male, female, and all students), which are not too far from the homogeneous bandwidth $h = 0.6$ on which the main results of Lindo et al. (2010) are built on. Due to different bandwidths adopted and the different local fitting (local logit versus local linear) methods, the numerical results could be substantially different. We bring an even larger contrast of the increased dropout-probability effect (caused by being placed on probation) on male students to that on female students; 5.6%, 0.1% and 1.2% for male, female and overall groups, respectively (compared to 3.7%, 0.6% and 1.8% as reported in their Table 4). Nonetheless, our qualitative results are largely in line with theirs, confirming that the probation strongly discourages male students to return to school, while the effect is little and insignificant (using the robust standard error) for female students or the overall sample.

We now consider the effect on graduation rates. As explained in Lindo et al. (2010), the effect on graduation is hard to predict since the probation status simultaneously causes some students to leave school, and others to improve their subsequent performance. To facilitate a multinomial analysis, we focus on the subsample which has information on graduation in six years. This allows us to divide the outcome in four categories ($J = 3$) for each student: graduated within 4 years ($j = 1$), graduated in the 5th year ($j = 2$), graduated in the 6th year ($j = 3$), and the base category, graduated after more than 6 years or dropped out from school at some point. The results are contained in Table 10 and Figure 6.

The probation status in the first year significantly (at the 5% level) reduces the probability of students that would graduate within 4 years by 4.8%, and at the same time, increases the probability of finishing the degree with an extended period, i.e. graduating in their 5th and 6th years by 1.0% and 0.5%, respectively, and graduating more than 6 years or be dropped out from school at some

point by 3.3%. It is clear from Figure 6 that the results can be sensitive to bandwidths. In Table 10 we also report the results for a larger bandwidth $h_{LSO} = 0.6$ (as used in Lindo et al., 2010) and such an extended graduation effect remains. The probabilities of graduating within 4 years and in the 5th year are reduced by 1.3% and 2.0%, while the probabilities of graduating in the 6th year or after and leaving the college at some point are increased by 1.1% and 2.2%. The results motivate one to combine the two categories of graduating earlier and combine the other two (thus to generate one single binary outcome, graduated in 5 years), hoping to uncover the largest extended graduation effect. This explains the most significance observed in their Table 6 (Panel B), compared to cases of other binary dependent variables, when the binary outcome of graduating in 5 years is used.

Another benefit of our multinomial analysis is the identification of the group of students who are most likely to graduate one or two years later than the normal four years [Panels (b) and (c) in Figure 6]. Interestingly, they happen to be students whose first-year GPAs are around the probation cutoff, and might be the target group for which the incentive of the probation standard is designed. Our results under the sharp RD design, which concern the same group, thus have important policy implications.

7 Conclusion

This paper complements the rapidly-growing methodological literature on the regression discontinuity design by providing estimation and inference procedures which explicitly address the discrete nature of the outcome variable. The outcome can be binary or multiple-valued, and nominal or ordinal. Our approach is fully nonparametric and accommodates the natural range of the estimand. When the outcome is polytomous, our approach, in contrast to existing ones, considers all categories holistically and exploits the interaction of categories with each other, rather than viewing each category in isolation. The proposed statistical procedures are easy to implement, and are illustrated by two empirical applications and simulation experiments.

Several extensions are possible. First, the local MNL approach proposed here allows a nonparametric function for each category, and is thus computationally taxing if J is large. One example is the discrete duration outcome, which may span many cells but is often too coarse to be treated as continuous. Xu (2016b) pursued a semi-nonparametric approach by imposing a separability

structure on the underlying hazard function to balance issues of tractability and flexibility, as an alternative to the fully nonparametric approach considered here.

Second, we have focused on the sharp RD design. In some applications, the treatment status T_i is only partially determined by the threshold-crossing indicator (T_i is not equal to, but is correlated with, $\mathbb{I}(X_i \geq c)$). In such so-called fuzzy RD design, τ_j in (3) is identified as

$$\tau_j = \frac{\lim_{x \to c+} \mathbb{P}(\widetilde{Y}_i = j | X_i = x) - \lim_{x \to c-} \mathbb{P}(\widetilde{Y}_i = j | X_i = x)}{\lim_{x \to c+} \mathbb{P}(T_i = 1 | X_i = x) - \lim_{x \to c-} \mathbb{P}(T_i = 0 | X_i = x)},$$

if $\lim_{x \to c+} \mathbb{P}(T_i = 1 | X_i = x) \neq \lim_{x \to c-} \mathbb{P}(T_i = 0 | X_i = x)$, and $\mu_{+,j}(x)$ and $\mu_{-,j}(x)$ are continuous at $x = c$. Following our proposal, τ_j can be estimated by the sharp RD estimator in (12) divided by the estimated jump in the treatment probability at c, which can be obtained by using the local MNL estimator for a binary outcome. The optimal bandwidth and robust inference, however, are more involved and worth further investigation. Another related design is based on the discontinuity in the derivative of the mean function (instead of the mean function itself); See Card et al. (2015). The local likelihood approach we pursued here would be also useful.

Third, empirical researchers often observe additional covariates which are not expected to be affected by the treatment. While incorporating such covariates is not necessary for identification of the causal effect, they are often included in estimation to improve efficiency (in the similar way as in randomized experiments). Imbens and Lemieux (2008, p.625) and Calonico, Cattaneo, Farrell and Titiunik (2016) discussed how. The latter paper also showed including covariates has certain effects on the bandwidth choice and the subsequent robust inference, with the focus on the local polynomial RD estimator. Extension to the local MNL framework requires non-trivial additional work.

Fourth, although our theory applies to general local polynomial fitting (of the transformed probability functions), we assume the polynomial order is pre-specified and we use a local linear approximation when implemented. Hall and Racine (2015) has argued that the appropriate order of the polynomial depends on the data generating process and proposed a cross-validation approach to jointly select the polynomial order and the bandwidth. Although their theory does not directly extend to the current setting which concerns the mean function only at a boundary point, the idea is worth being pursued further.

8 Appendix A: Technical details

Appendix A provides the main technical results and their proofs. Lemmas are contained in Appendix B and additional details are presented in Appendix C.

The theoretical results presented here extend earlier work of Fan, Heckman and Wand (1995) by considering the case of multiple-category nominal or ordinal outcomes, optimal bandwidth selection and robust inference under bias correction.

8.1 Assumptions and theorems

The following Assumption A is assumed throughout the paper.

Assumption A.

A1. Assume $f(x)$ is continuous at $x = c$, and $f(c) > 0$.

A2. For $j = 1, ..., J$, $\mu_{+,j}(x)$ and $\mu_{-,j}(x)$ are $(p+2)$ times $(p \geq 0)$ continuously differentiable at $x = c$.

A3. For $j = 1, ..., J$, $0 < \mu_{+,j}(x) < 1$ and $0 < \mu_{-,j}(x) < 1$, for x in a neighborhood of c.

A4. $\Upsilon_+ g_+^{(p+1)} \neq \Upsilon_- g_-^{(p+1)}$.

A5. $h \to 0$, as $n \to \infty$.

Theorem 1. Under Assumptions A1-A3 and A5, $\widehat{\tau} \xrightarrow{p} \tau$.

Theorem 2. For $\nu = 0, 1, ..., p$, assume $(nh)^{1/2} h^{p+1} = O(1)$ and $(nh)^{1/2} h^{\nu} \to \infty$. Then

$$(nh)^{1/2} h^{\nu} (\widehat{g}_+^{(\nu)} - g_+^{(\nu)} - B_{g,+}) \xrightarrow{d} \mathcal{N}(0, \underset{J \times J}{V_{g,+}}),$$

where $B_{g,+} = c_{B,p,\nu} h^{p+1-\nu} g_+^{(p+1)} \nu!/(p+1)!$ and $V_{g,+} = \Upsilon_+^{-1} c_{V,p,\nu} (\nu!)^2 / f(c)$. The similar result holds for $\widehat{g}_-^{(\nu)}$.

Theorem 3. Assume $(nh)^{1/2} h^{p+1} = O(1)$ and $nh \to \infty$. Then

$$(nh)^{1/2} (\widehat{\tau} - \tau - B) \xrightarrow{d} \mathcal{N}(0, V),$$

where B and V are given in Section 3 (below the equation (13)).

Theorem 4. (18) holds.

Theorem 5. Let $\widehat{B}_{g,+}$ be defined in Section 5. Assume $b_+ + h/b_+ + n^{-1}b_+^{-(2p+3)} \to 0$. Then under the assumptions in Theorem 3,

$$(nh)^{1/2}\widehat{\Omega}_{g,+}^{-1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+}) = (nh)^{1/2}\Omega_{g,+}^{-1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+}) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_J), \tag{24}$$

where I_J is the $J \times J$ identity matrix, $\Omega_{g,+} = f(c)^{-1}\Upsilon_+^{-1}r_+$ and $\widehat{\Omega}_{g,+} = \widehat{f}(c)^{-1}\widehat{\Upsilon}_+^{-1}r_+$ with r_+ defined in (21). The similar result holds for $\widehat{g}_-$.

Theorem 6. Let $\widehat{B}$, Ω and $\widehat{\Omega}$ be defined in Section 5. Then under the assumptions in Theorem 5,

$$(nh)^{1/2}\widehat{\Omega}^{-1/2}(\widehat{\tau} - \tau - \widehat{B}) = (nh)^{1/2}\Omega^{-1/2}(\widehat{\tau} - \tau - \widehat{B}) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_J).$$

8.2 Proofs of Theorems

The proofs presented below rely on the preliminary results shown in Appendices B and C (Lemma 2 and the proof of Lemma 1). In this section, "*s.o.*" means smaller order terms (in probability). LLN, LIE and CLT mean the law of large numbers, law of iterated expectation and central limit theorem, respectively.

Proof of Theorem 1. The result follows from consistency of $\widehat{\mu}_+$ and $\widehat{\mu}_-$, which holds from consistency of $\widehat{g}_+$ and $\widehat{g}_-$, and (6)-(7). Consistency of $\widehat{g}_+$ and $\widehat{g}_-$ follows from the proof of Lemma 1 in which is demonstrated that the mean and variance both converge to zero.

Proof of Theorem 2. It follows from Lemma 2.

Proof of Theorem 3. It follows from Theorem 2 (with $\nu = 0$), multivariate delta method and $\Psi'(g_+) = \Upsilon_+$ (Ψ is defined in (33)).

Proof of Theorem 4. Using the pilot bandwidths, we can obtain the $n^{-2/5}$-consistency of $\widetilde{\Upsilon}_+, \widetilde{\Upsilon}_-$,and $\widehat{f}(c)$, and $n^{-1/7}$-consistency $\widehat{g}_+^{(2)}, \widehat{g}_-^{(2)}$. Then Theorem 4 holds by the delta method (i.e. for a vector of parameters θ and its consistent estimator $\widehat{\theta}$, the rate of $\varphi(\widehat{\theta}) - \varphi(\theta_0)$ or $\varphi(\widehat{\theta})/\varphi(\theta_0) - 1$ is determined by the entry of $\widehat{\theta}$ with the slowest rate converging to its true value if $\varphi(\cdot)$ is differentiable and $\varphi(\theta_0)\varphi'(\theta_0) \neq 0$).

Proof of Theorem 5. We are to show

$$(nh)^{1/2}\Omega_{g,+}^{-1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+}) \xrightarrow{d} \mathcal{N}(0, \mathrm{I}_J), \tag{25}$$

and then the other part of (24) follows from consistent estimation of $\Omega_{g,+}$ under the bandwidth conditions.

Let $S_0 = \mathrm{I}_J \otimes e_{p+1,1}^\top$ and $S_{p+1} = \mathrm{I}_J \otimes e_{p+2,p+2}^\top$ (be selection matrices). We use the same notations as in the proof of Lemma 1 (which is laid out for local likelihood estimator of order p) specialized to the MNL model. We use here the bold symbols to denote the versions that apply to local likelihood estimator of order $p+1$ (with the bandwidth b_+). For examples, $\widehat{\boldsymbol{\gamma}}$ is $J(p+2)$-dim while $\widehat{\gamma}$ is $J(p+1)$-dim, $\mathbf{Z}_i = (1, (X_i - c)/b_+, ..., (X_i - c)^{p+1}/b_+^{p+1})^\top$ while $Z_i = (1, (X_i - c)/h, ..., (X_i - c)^p/h^p)^\top$, etc..

The proof of (25) proceeds by seeking for the leading terms of $\mathrm{Var}[(nh)^{1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+})]$. Note that

$$\begin{aligned}
& (nh)^{1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+}) \\
= \; & (nh)^{1/2}[\widehat{g}_+ - g_+ - (\widehat{B}_{g,+} - B_{g,+})] - (nh)^{1/2} B_{g,+} \\
= \; & (nh)^{1/2}(\widehat{g}_+ - g_+) - (nh)^{1/2} h^{p+1}(\widehat{\mathbf{g}}_+^{(p+1)} - g_+^{(p+1)}) c_{B,p,0}/(p+1)! - (nh)^{1/2} B_{g,+} \\
= \; & (nh)^{1/2}(\widehat{g}_+ - g_+) - (h/b_+)^{p+3/2}(nb_+)^{1/2} b_+^{p+1}(\widehat{\mathbf{g}}_+^{(p+1)} - g_+^{(p+1)}) c_{B,p,0}/(p+1)! - (nh)^{1/2} B_{g,+} \\
= \; & S_0\widehat{\gamma} - (h/b_+)^{p+3/2} S_{p+1}\widehat{\boldsymbol{\gamma}} c_{B,p,0} - (nh)^{1/2} B_{g,+} \\
\overset{(40)}{=} \; & -S_0\Sigma_+^{-1} W_{n,+} + (h/b_+)^{p+3/2} S_{p+1}\boldsymbol{\Sigma}_+^{-1}\mathbf{W}_{n,+} c_{B,p,0} - (nh)^{1/2} B_{g,+} + s.o. \\
\overset{(41)}{=} \; & -S_0\Sigma_+^{-1}(A_1 + A_2) + (h/b_+)^{p+3/2} S_{p+1}\boldsymbol{\Sigma}_+^{-1}(\mathbf{A_1} + \mathbf{A_2}) c_{B,p,0} - (nh)^{1/2} B_{g,+} + s.o. \\
= \; & -S_0\Sigma_+^{-1} A_1 + (h/b_+)^{p+3/2} S_{p+1}\boldsymbol{\Sigma}_+^{-1}(\mathbf{A_1} + \mathbf{A_2}) c_{B,p,0} + s.o.,
\end{aligned} \tag{26}$$

where the last line follows from (43). Thus

$$\mathrm{Var}[(nh)^{1/2}(\widehat{g}_+ - g_+ - \widehat{B}_{g,+})] = \mathrm{Var}[-S_0\Sigma_+^{-1} A_1 + (h/b_+)^{p+3/2} S_{p+1}\boldsymbol{\Sigma}_+^{-1}\mathbf{A_1} c_{B,p,0}] + s.o., \tag{27}$$

noting that $\mathbf{A_2}$ only contributes to a bias term (which is negligible under bandwidth conditions since $\mathbb{E}\mathbf{A}_2 = O(n^{1/2} b_+^{1/2} b_+^{p+2})$).

We are now to evaluate the variance in the right-hand-side of (27) (noting that the expectation is zero):

$$\mathrm{Var}(S_0\Sigma_+^{-1}A_1)+(\frac{h}{b_+})^{2p+3}\mathrm{Var}(S_{p+1}\mathbf{\Sigma}_+^{-1}\mathbf{A_1})c_{B,p,0}^2-2(\frac{h}{b_+})^{p+3/2}\mathrm{Cov}(S_0\Sigma_+^{-1}A_1,S_{p+1}\mathbf{\Sigma}_+^{-1}\mathbf{A_1})c_{B,p,0}. \tag{28}$$

Note that by the proof of Lemma 1 (specialized to the MNL model),

$$\mathrm{Var}(S_0\Sigma_+^{-1}A_1) = \Upsilon_+^{-1}c_{V,p,0}/f(c)+o_p(1) \tag{29}$$

$$\mathrm{Var}(S_{p+1}\mathbf{\Sigma}_+^{-1}\mathbf{A_1}) = \Upsilon_+^{-1}c_{V,p+1,p+1}((p+1)!)^2/f(c)+o_p(1). \tag{30}$$

Now look at the covariance term in (28). The (j,j')-block [which is $(p+1)\times(p+2)$] of $\mathrm{Cov}(A_1,\mathbf{A_1})$ is, for now assuming $j\neq j'$,

$$\begin{aligned}
& (nh)^{-\frac{1}{2}}(nb_+)^{-\frac{1}{2}}\sum_{i=1}^n \mathbb{E}\nabla_{g_j}\ell(g_+(X_i);Y_i)\nabla_{g_{j'}}\ell(\mathbf{g}_+(X_i);Y_i)K\left(\tfrac{X_i-c}{h}\right)K\left(\tfrac{X_i-c}{b_+}\right)I_iZ_i\mathbf{Z}_i^\top \\
\overset{(6),\,\mathrm{LIE}}{=}\ & -(nh)^{-\frac{1}{2}}(nb_+)^{-\frac{1}{2}}\sum_{i=1}^n \mathbb{E}\frac{\exp(g_{+,j}(X_i)+\mathbf{g}_{+,j'}(X_i))}{[1+\sum_{j=1}^J\exp(g_{+,j}(X_i))][1+\sum_{j=1}^J\exp(\mathbf{g}_{+,j'}(X_i))]}\cdot \\
& \cdot K((X_i-c)/h)K((X_i-c)/b_+)I_iZ_i\mathbf{Z}_i^\top \\
\overset{\mathrm{Taylor}}{=}\ & -(b_+/h)^{-1/2}(nh)^{-1}\sum_{i=1}^n \mathbb{E}\mu_{+,j}(X_i)\mu_{+,j'}(X_i)K\left(\tfrac{X_i-c}{h}\right)K\left(\tfrac{X_i-c}{b_+}\right)I_iZ_i\mathbf{Z}_i^\top+s.o. \\
\overset{\substack{u=\frac{X-c}{h}\\ =\\ h/b_+\to 0}}{}\ & -(b_+/h)^{-1/2}\mu_{+,j}\mu_{+,j'}f(c)\int_0^1\varsigma_p(u)K(u)K(0)\mathrm{d}u\underbrace{(1,0,...,0)}_{p+2}+s.o.
\end{aligned}$$

Similar results hold for (j,j)-blocks of $\mathrm{Cov}(A_1, \mathbf{A_1})$. Putting them together,

$$\begin{aligned}
& \mathrm{Cov}(S_0\Sigma_+^{-1}A_1, S_{p+1}\boldsymbol{\Sigma}_+^{-1}\mathbf{A_1}) \\
= \; & S_0\Sigma_+^{-1}\mathrm{Cov}(A_1, \mathbf{A_1})\boldsymbol{\Sigma}_+^{-1}S_{p+1}^\top \\
= \; & S_0[\Upsilon_+ \otimes f(c)N_p]^{-1}\mathrm{Cov}(A_1, \mathbf{A_1})[\Upsilon_+ \otimes f(c)N_{p+1}]^{-1}S_{p+1}^\top \\
= \; & S_0[\Upsilon_+ \otimes f(c)N_p]^{-1}[(b_+/h)^{-1/2}\Upsilon_+ \otimes f(c)\int_0^1 \varsigma_p(u)K(u)\mathrm{d}uK(0)e_{p+2,1}^\top] \cdot \\
& \cdot[\Upsilon_+ \otimes f(c)N_{p+1}]^{-1}S_{p+1}^\top + s.o. \\
= \; & (b_+/h)^{-1/2}S_0[\Upsilon_+^{-1} \otimes f^{-1}(c)N_p^{-1}\int_0^1 \varsigma_p(u)K(u)\mathrm{d}uK(0)e_{p+2,1}^\top N_{p+1}^{-1}]S_{p+1}^\top + s.o. \\
= \; & (b_+/h)^{-1/2}[\Upsilon_+^{-1} \otimes f^{-1}(c)e_{p+1,1}^\top N_p^{-1}(\int_0^1 \varsigma_p K)K(0)e_{p+2,1}^\top N_{p+1}^{-1}e_{p+2,p+2}] + s.o.. \\
= \; & (b_+/h)^{-1/2}[\Upsilon_+^{-1} \otimes f^{-1}(c)(\int_0^1 K_{p,0}^*)K(0)e_{p+2,1}^\top N_{p+1}^{-1}e_{p+2,p+2}] + s.o.. \qquad (31)
\end{aligned}$$

Then (25) follows from the central limit theorem, (28), (29), (30), (31) and the fact that $\int_0^1 K_{p,0}^* = 1$.

Proof of Theorem 6. It follows from Theorem 5 and the multivariate delta method.

8.3 Bandwidth selection for general p

The optimal bandwidth in (15) and pilot bandwidths in (16) and (17) for a general p ($p \geq 0$) are presented below:

$$\widehat{h}^{2p+3} = \frac{c_{V,p,0}[(p+1)!]^2}{c_{B,p,0}^2 2(p+1)n\widehat{f}(c)} \frac{\mathrm{trace}(\widetilde{\Upsilon}_+ + \widetilde{\Upsilon}_-)}{||\widetilde{\Upsilon}_+\widetilde{g}_+^{(p+1)} - \widetilde{\Upsilon}_-\widetilde{g}_-^{(p+1)}||^2}$$

$$\widetilde{h}_{\mu_+}^{2p+3} = \frac{c_{V,p,0}[(p+1)!]^2}{c_{B,p,0}^2 2(p+1)n\widehat{f}(c)} \frac{\mathrm{trace}(\breve{\Upsilon}_+)}{||\breve{\Upsilon}_+\breve{g}_+^{(p+1)}||^2}, \quad \widetilde{h}_{\mu_-}^{2p+3} = \frac{c_{V,p,0}[(p+1)!]^2}{c_{B,p,0}^2 2(p+1)n\widehat{f}(c)} \frac{\mathrm{trace}(\breve{\Upsilon}_-)}{||\breve{\Upsilon}_-\breve{g}_-^{(p+1)}||^2},$$

$$\begin{aligned}
\widehat{b}_+^{2p+5} &= \frac{(2p+3)[(p+2)!]^2 c_{V,p+1,p+1}}{2c_{B,p+1,p+1}^2 n\widehat{f}(c)} \frac{\mathrm{trace}(\breve{\Upsilon}_+^{-1})}{||\breve{g}_+^{(p+2)}||^2}, \\
\widehat{b}_-^{2p+5} &= \frac{(2p+3)[(p+2)!]^2 c_{V,p+1,p+1}}{2c_{B,p+1,p+1}^2 n\widehat{f}(c)} \frac{\mathrm{trace}(\breve{\Upsilon}_-^{-1})}{||\breve{g}_-^{(p+2)}||^2},
\end{aligned}$$

where the last two bandwidths $\widehat{b}_+$ and $\widehat{b}_-$ are to estimate $g_+^{(p+1)}$ and $g_-^{(p+1)}$ by a local likelihood with order $(p+1)$. They reduce to (15), (16) and (17) when $p = 1$.

9 Appendix B: Supplementary materials

To prove Theorems 2 and 3, which follow from Lemmas 1 and 2 below, we consider a slightly more general framework. The benefit of the general consideration is that it also allows the analysis of fitting a local ordered outcome model (see (34) below), as well as the MNL model as described in the main text.

9.1 The general model

Let conditional outcome probabilities $\mu_{+,j}(x)$ and $\mu_{-,j}(x)$ be as defined in (1) and (2). Let $\mu_+(x) = (\mu_{+,1}(x), ..., \mu_{+,J}(x))^\top$. The goal is to derive the limit distributions of the estimators of $\mu_+(c)$ and $\mu_-(c)$. We here focus on $\mu_+(c)$, and the similar derivation holds for $\mu_-(c)$.

We consider the model

$$\mu_+(x) = \Psi(\eta_+(x)), \tag{32}$$

for $x \in \mathcal{X}$, where $\Psi : \mathbb{R}^J \mapsto \mathbb{R}^J$. The Jacobian is $\Psi'(\eta_+(x))$.

Partition $\eta_+(x) = (\alpha_+^\top, g_+^\top(x))^\top \in \mathbb{R}^J$, with dimensions d_α and d_g, satisfying $d_\alpha + d_g = J$. The model (32) takes one of the following two forms. In the MNL (as in (6)), $d_\alpha = 0$ and $d_g = J$,

$$\mu_+(x) = \Psi(\eta_+(x)) = \Psi(g_+(x)) = \left(\frac{\exp(g_{+,j}(x))}{1+\sum_{j=1}^J \exp(g_{+,j}(x))} \right)_{1 \le j \le J} \tag{33}$$

In the ordered outcome model, $d_\alpha = J - 1$ and $d_g = 1$,

$$\mu_+(x) = \Psi(\eta_+(x)) = \Psi(\alpha_+, g_+(x)) = \Big(\Phi(\alpha_{+,j} - g_+(x)) - \Phi(\alpha_{+,j-1} - g_+(x)) \Big)_{1 \le j \le J}, \tag{34}$$

with $\alpha_{+,0} = 0$ and $\alpha_{+,J} = \infty$. Here $\Phi(\cdot)$ is a CDF function, and $\alpha_{+,0} = 0$ is an identification assumption (so that the function $g_+(x)$ is unconstrained). As mentioned in the main text, (33) is rather a transformation since it does not impose any structure on the data generating process. In contrast, (34) impose restrictions which explore the ordered structure of the outcome. Note that both models reduce to a local binary logit model when $J = 1$.

For the $J \times 1$ generic argument $\eta = (\alpha^\top, g^\top)^\top$, denote

$$\begin{aligned}
\ell(\eta; y_1, ..., y_J) &= \ell(\alpha, g; y_1, ..., y_J) \\
&= (1 - \sum_{j=1}^{J} y_j) \log(1 - \sum_{j=1}^{J} \mu_j) + \sum_{j=1}^{J} y_j \log \mu_j \\
&= \sum_{j=1}^{J} y_j \log \frac{\mu_j}{1 - \sum_{j=1}^{J} \mu_j} + \log(1 - \sum_{j=1}^{J} \mu_j),
\end{aligned}$$

where $\mu = \Psi(\eta)$.

For a value X that is in the right neighborhood of c, for $k = 1, ..., d_g$, let $g_{+,k}(X)$ be approximated by a p-th order polynomial

$$g_{+,k}(X) \approx \overline{g}_{+,k}(c, X) = g_{+,k} + (X - c) g_{+,k}^{(1)} + ... + (X - c)^p g_{+,k}^{(p)} / p! = \overline{X}^\top \beta_{+,k}^*, \tag{35}$$

where $\overline{X} = (1, X - c, ..., (X - c)^p)^\top$ and $\beta_{+,k}^* = (\beta_{+,k,0}^*, \beta_{+,k,1}^*, ..., \beta_{+,k,p}^*)^\top = (g_{+,k}, g_{+,k}^{(1)}, ..., g_{+,k}^{(p)} / p!)^\top$.

Define the local log-likelihood, writing $\theta = (\alpha^\top, \beta^\top)^\top$ and $\beta = (\beta_1^\top, ..., \beta_{d_g}^\top)^\top \in \mathbb{R}^{(p+1)d_g}$,

$$L_+(\theta) = L_+(\alpha, \beta) = \sum_{i=1}^{n} \ell(\alpha, \overline{X}_i^\top \beta_1, ..., \overline{X}_i^\top \beta_{d_g}; Y_{i1}, ..., Y_{iJ}) K((X_i - c)/h) I_i.$$

Denote

$$\widehat{\theta}_+ = (\widehat{\alpha}_+^\top, \widehat{\beta}_+^\top)^\top = \arg \max_{\theta \in \mathbb{R}^{d_\alpha + (p+1)d_g}} L_+(\theta),$$

where $\widehat{\beta}_+ = (\widehat{\beta}_{+,1}^\top, ..., \widehat{\beta}_{+,d_g}^\top)^\top$. The local likelihood estimator of $\eta_+(c) \in \mathbb{R}^J$ is $\widehat{\eta}_+(c) = (\widehat{\alpha}_+^\top, \widehat{\beta}_{+,1,0}, ..., \widehat{\beta}_{+,d_g,0})^\top$. Thus $\widehat{\mu}_+(c) = \Psi(\widehat{\eta}_+(c))$.

Denote the $J \times 1$ vector $\nabla_\eta \ell(\alpha, g; y_1, ..., y_J) = (\partial \ell(\alpha, g; y_1, ..., y_J)/\partial \alpha^\top, \partial \ell(\alpha, g; y_1, ..., y_J)/\partial g^\top)^\top$. The similar definition is used for the $J \times J$ Hessian matrix $\nabla_{\eta\eta^\top} \ell(\alpha, g; y_1, ..., y_J)$. Conditional Bartlett identities are satisfied at the true value of α and $g(\cdot)$, i.e.

$$\mathbb{E}[\nabla_\eta \ell(\alpha^*, g(X); Y_1, ..., Y_J) I | X] = 0 \tag{36}$$

$$\mathbb{E}[\nabla_\eta \ell(\alpha^*, g(X); Y_1, ..., Y_J) \nabla_\eta \ell(\alpha^*, g(X); Y_1, ..., Y_J)^\top I | X] = -\mathbb{E}[\nabla_{\eta\eta^\top} \ell(\alpha^*, g(X); Y_1, ..., Y_J) I | X], \tag{37}$$

where $I = \mathbb{I}(X \geq c)$.

9.2 Lemmas

We now present asymptotic results for the framework described above. Denote the $(p+1)\times(p+1)$ matrix $T_p = (\int_0^1 u^{i+j-2}K^2(u)\mathrm{d}u)_{i,j=1,...,p+1}$. Define

$$\begin{aligned}\underset{J\times J}{\mathcal{I}_+} &= \lim_{x\to c+}\mathbb{E}[\nabla_{\eta_+}\ell(\alpha_+^*, g_+(X); Y_1, ..., Y_J)\nabla_{\eta_+}\ell(\alpha_+^*, g_+(X); Y_1, ..., Y_J)^\top | X = x] \\ &= \begin{pmatrix} \underset{d_\alpha\times d_\alpha}{\mathcal{I}_{+,\alpha\alpha}} & \underset{d_\alpha\times d_g}{\mathcal{I}_{+,\alpha g^\top}} \\ \underset{d_g\times d_\alpha}{\mathcal{I}_{+,\alpha^\top g}} & \underset{d_g\times d_g}{\mathcal{I}_{+,gg}} \end{pmatrix}\end{aligned}$$

and

$$\underset{d\times d}{\Sigma_+} = \begin{pmatrix} \underset{d_\alpha\times d_\alpha}{\Sigma_{+,\alpha\alpha}} & \underset{d_\alpha\times(p+1)d_g}{\Sigma_{+,\alpha g^\top}} \\ \underset{(p+1)d_g\times d_\alpha}{\Sigma_{+,\alpha^\top g}} & \underset{(p+1)d_g\times(p+1)d_g}{\Sigma_{+,gg}} \end{pmatrix},\ \underset{d\times d}{\Gamma_+} = \begin{pmatrix} \underset{d_\alpha\times d_\alpha}{\Gamma_{+,\alpha\alpha}} & \underset{d_\alpha\times(p+1)d_g}{\Gamma_{+,\alpha g^\top}} \\ \underset{(p+1)d_g\times d_\alpha}{\Gamma_{+,\alpha^\top g}} & \underset{(p+1)d_g\times(p+1)d_g}{\Gamma_{+,gg}} \end{pmatrix},$$

where $d = d_\alpha + (p+1)d_g$, and

$$\begin{aligned}\Sigma_{+,\alpha\alpha} &= \mathcal{I}_{+,\alpha\alpha}f(c)\textstyle\int_0^1 K(u)\mathrm{d}u, & \Sigma_{+,\alpha g^\top} &= \mathcal{I}_{+,\alpha g^\top}f(c)\otimes\textstyle\int_0^1 \varsigma_p^\top(u)K(u)\mathrm{d}u, \\ \Sigma_{+,gg} &= \mathcal{I}_{+,gg}f(c)\otimes N_p, & \Gamma_{+,\alpha\alpha} &= \mathcal{I}_{+,\alpha\alpha}f(c)\textstyle\int_0^1 K^2, \\ \Gamma_{+,\alpha g^\top} &= \mathcal{I}_{+,\alpha g^\top}f(c)\otimes\textstyle\int_0^1 \varsigma_p^\top(u)K^2(u)\mathrm{d}u, & \Gamma_{+,gg} &= \mathcal{I}_{+,gg}f(c)\otimes T_p,\end{aligned}$$

with $\varsigma_p(u) = (1, u, ..., u^p)^\top$ and the Kronecker product $\otimes$.

In what follows, we write $\int_0^1 uK(u)\mathrm{d}u$ and $\int_0^1 \varsigma_p(u)K(u)\mathrm{d}u$ as $\int uK$ and $\int \varsigma_p K$ (omitting the integrator, and integrals always over [0,1]), respectively, for simplicity.

Lemma 1. Let $\widehat{\gamma}_\alpha = (nh)^{1/2}(\widehat{\alpha}_+ - \alpha_+^*)$ and $\widehat{\gamma}_\beta = (\widehat{\gamma}_{\beta,1}^\top, ..., \widehat{\gamma}_{\beta,d_g}^\top)^\top$ where $\widehat{\gamma}_{\beta,k} = (nh)^{1/2}(\widehat{\beta}_{+,k,0} - g_{+,k}, h(\widehat{\beta}_{+,k,1} - g_{+,k}^{(1)}), ..., h^p(\widehat{\beta}_{+,k,p} - g_{+,k}^{(p)}/p!))^\top$ with $1 \leq k \leq d_g$. Let $\widehat{\gamma} = (\widehat{\gamma}_\alpha^\top, \widehat{\gamma}_\beta^\top)^\top$. Then

$$\widehat{\gamma} - B_{\gamma,+} \xrightarrow{d} \mathcal{N}(0, \underset{d\times d}{\Sigma_+^{-1}\Gamma_+\Sigma_+^{-1}}), \tag{38}$$

where $B_{\gamma,+} = \Sigma_+^{-1}\overline{B}_{\gamma,+}$ with

$$\underset{d\times 1}{\overline{B}_{\gamma,+}} = (nh)^{1/2}h^{p+1}f(c)\begin{pmatrix} \mathcal{I}_{+,\alpha g^\top} g_+^{(p+1)} \int u^{p+1}K \\ \mathcal{I}_{+,gg} g_+^{(p+1)} \otimes \int u^{p+1}\varsigma_p K \end{pmatrix} /(p+1)!.$$

Remark. $\widehat{\alpha}_+$ is also biased, although α_+ is assumed globally constant, and the bias comes from estimating g's.

Lemma 2. In the MNL model, (38) holds with

$$\underset{(p+1)J\times 1}{B_{\gamma,+}} = (nh)^{1/2}h^{p+1}(g_+^{(p+1)} \otimes \bar{c}_{B,p})/(p+1)!,$$

where $\bar{c}_{B,p}$ is $(p+1)\times 1$ with elements $c_{B,p,\nu}$, for $\nu = 0,1,...,p$, and $\Sigma_+^{-1}\Gamma_+\Sigma_+^{-1} = \Upsilon_+^{-1} \otimes f(c)^{-1}N_p^{-1}T_pN_p^{-1}$, with Υ_+ being defined in the main text.

Lemma 3. In the ordered model,

$$(nh)^{1/2}[(\widehat{\alpha}_+ - \alpha_+^*, \widehat{g}_+ - g_+) - B_{\eta,+}] \xrightarrow{d} \mathcal{N}(0, \underset{J\times J}{V_{\eta,+}}),$$

where

$$\begin{aligned} B_{\eta,+} &= (nh)^{1/2}h^{p+1}g_+^{(p+1)}\mathrm{diag}(\mathrm{I}_{J-1}, e_{p+1,1}^\top)\begin{pmatrix} \mathcal{I}_{+,\alpha\alpha}\int K & \mathcal{I}_{+,\alpha g}\int \varsigma_p^\top K \\ \mathcal{I}_{+,\alpha^\top g}\int \varsigma_p K & \mathcal{I}_{+,gg}N_p \end{pmatrix}^{-1} \cdot \\ &\quad \cdot \begin{pmatrix} \mathcal{I}_{+,\alpha g^\top}\int u^{p+1}K \\ \mathcal{I}_{+,gg}\int u^{p+1}\varsigma_p K \end{pmatrix} /(p+1)! \end{aligned}$$

and

$$\begin{aligned} V_{\eta,+} &= \mathrm{diag}(\mathrm{I}_{J-1}, e_{p+1,1}^\top)f(c)^{-1}\begin{pmatrix} \mathcal{I}_{+,\alpha\alpha}\int K & \mathcal{I}_{+,\alpha g}\int \varsigma_p^\top K \\ \mathcal{I}_{+,\alpha^\top g}\int \varsigma_p K & \mathcal{I}_{+,gg}N_p \end{pmatrix}^{-1}\begin{pmatrix} \mathcal{I}_{+,\alpha\alpha}\int K^2 & \mathcal{I}_{+,\alpha g}\int \varsigma_p^\top K^2 \\ \mathcal{I}_{+,\alpha^\top g}\int \varsigma_p K^2 & \mathcal{I}_{+,gg}T_p \end{pmatrix} \cdot \\ &\quad \cdot \begin{pmatrix} \mathcal{I}_{+,\alpha\alpha}\int K & \mathcal{I}_{+,\alpha g}\int \varsigma_p^\top K \\ \mathcal{I}_{+,\alpha^\top g}\int \varsigma_p K & \mathcal{I}_{+,gg}N_p \end{pmatrix}^{-1} \mathrm{diag}(\mathrm{I}_{J-1}, e_{p+1,1}). \end{aligned}$$

Thus

$$(nh)^{1/2}[\widehat{\mu}_+ - \mu_+ - \Psi'(\eta_+)B_{\eta,+}] \xrightarrow{d} \mathcal{N}(0, \Psi'(\eta_+)V_{\eta,+}\Psi'(\eta_+)^\top).$$
$$J\times J$$

Remark. The MNL model has simple forms of bias and variance (Lemma 2). It follows from $\mathcal{I}_+$=Υ_+ and $\Psi'(\eta_+) = \Upsilon_+$. Then Σ_+ and Γ_+ in Lemma 1 reduce to $\Sigma_+ = \Upsilon_+ \otimes f(c)N_p$ and $\Gamma_+ = \Upsilon_+ \otimes f(c)T_p$, both of which are $(p+1)J \times (p+1)J$.

10 Appendix C: Proof of Lemma 1

Proof of Lemma 1. Write $K_i = K((X_i - c)/h)$. Let $\gamma = (\gamma_\alpha^\top, \gamma_\beta^\top)^\top$, where $\gamma_\alpha = (nh)^{1/2}(\alpha - \alpha_+^*)$ and $\gamma_\beta = (\gamma_{\beta,1}^\top, ..., \gamma_{\beta,d_g}^\top)^\top$ where $\gamma_{\beta,k} = (nh)^{1/2}(\beta_{k,0} - g_{+,k}, h(\beta_{k,1} - g_{+,k}^{(1)}), ..., h^p(\beta_{k,p} - g_{+,k}^{(p)}/p!))^\top$ with $1 \le k \le d_g$. Write $Z_i = (1, (X_i - c)/h, ..., (X_i - c)^p/h^p)^\top$. Then

$$\widehat{\gamma} = \arg \max_{\gamma \in \mathbb{R}^{d_\alpha + (p+1)d_g}} \overline{L}_+(\gamma),$$

where

$$\begin{aligned}\overline{L}_+(\gamma) &= \underbrace{\sum_{i=1}^n \ell\left(\alpha_+^* + \tfrac{\gamma_\alpha}{(nh)^{1/2}}, \overline{g}_{+,1}(c, X_i) + \tfrac{\gamma_{\beta,1}^\top Z_i}{(nh)^{1/2}}, ..., \overline{g}_{+,d_g}(c, X_i) + \tfrac{\gamma_{\beta,d_g}^\top Z_i}{(nh)^{1/2}}; Y_i\right) K_i I_i}_{=L_+(\alpha,\beta)} \\ &\quad - \sum_{i=1}^n \ell(\alpha_+^*, \overline{g}_+(c, X_i); Y_i) K_i I_i,\end{aligned}$$

with $\overline{g}_{+,k}(c, X_i)$ being defined in (35) (as an approximation of $g_{+,k}(X_i)$). Note that the second term above does not depend on α or β.

Define

$$\underset{[d_\alpha + (p+1)d_g] \times 1}{W_{n,+}} = (nh)^{-1/2} \sum_{i=1}^n \begin{pmatrix} \nabla_\alpha \ell(\alpha_+^*, \overline{g}_+(c, X_i); Y_i) K_i I_i \\ \nabla_g \ell(\alpha_+^*, \overline{g}_+(c, X_i); Y_i) K_i I_i \otimes Z_i \end{pmatrix}.$$

A Taylor expansion gives

$$\begin{aligned}
\overline{L}_+(\gamma) &= (nh)^{-1/2}\sum_{k=1}^{d_\alpha}\sum_{i=1}^{n}\nabla_{\alpha_k}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\gamma_{\alpha,k} \\
&\quad +(nh)^{-1/2}\sum_{k=1}^{d_g}\sum_{i=1}^{n}\nabla_{g_k}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\gamma_{\beta,k}^\top Z_i \\
&\quad +\frac{(nh)^{-1}}{2}\sum_{k,k'=1}^{d_\alpha}\sum_{i=1}^{n}\nabla_{\alpha_k\alpha_{k'}}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\gamma_{\alpha,k}\gamma_{\alpha,k'} \\
&\quad +\frac{(nh)^{-1}}{2}\sum_{k,k'=1}^{d_g}\sum_{i=1}^{n}\nabla_{g_kg_k}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i(\gamma_\beta^\top Z_i)(\gamma_\beta^\top Z_i) \\
&\quad +(nh)^{-1}\sum_{k=1}^{d_\alpha}\sum_{k'=1}^{d_g}\sum_{i=1}^{n}\nabla_{\alpha_kg_{k'}}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\gamma_{\alpha,k}\gamma_{\beta,k'}^\top Z_i + s.o. \\
&= W_{n,+}^\top\gamma + \frac{1}{2}\gamma^\top\Sigma_{n,+}\gamma + s.o. \\
&= W_{n,+}^\top\gamma + \frac{1}{2}\gamma^\top\Sigma_+\gamma + s.o., \qquad (39)
\end{aligned}$$

where

$$\underset{d\times d}{\Sigma_{n,+}} = (nh)^{-1}\sum_{i=1}^{n}\begin{pmatrix} \nabla_{\alpha\alpha^\top}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i & * \\ \nabla_{\alpha g^\top}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\otimes Z_i & \nabla_{gg^\top}\ell(\alpha_+^*,\overline{g}_+(c,X_i);Y_i)K_iI_i\otimes Z_iZ_i^\top \end{pmatrix},$$

and smaller order terms in (39) come from bounded derivatives. We use "$*$" above in a symmetric matrix to simplify the entry. In (39) we have used $\Sigma_{n,+}\xrightarrow{p}\Sigma_+$, which follows from a standard LLN, (35) and bounded third derivatives of ℓ with respect to α and g.

By (39) and the quadratic approximation lemma,

$$\widehat{\gamma} = -\Sigma_+^{-1}W_{n,+} + s.o.. \qquad (40)$$

For $1 \leq k \leq d_\alpha$, the k-th element of $W_{n,+}$ equals

$$\begin{aligned}
(W_{n,+})_k &= (nh)^{-1/2} \sum_{i=1}^{n} \nabla_{\alpha_k} \ell(\alpha_+^*, \overline{g}_+(c, X_i); Y_{i1}, ..., Y_{iJ}) K((X_i - x)/h) I_i \\
&\overset{\text{Taylor}}{\underset{(35)}{=}} \underbrace{(nh)^{-1/2} \sum_{i=1}^{n} \nabla_{\alpha_k} \ell(\alpha_+^*, g_+(X_i); Y_i) K_i I_i}_{=(A_1)_k} \\
&\quad \underbrace{+[(p+1)!]^{-1} (nh)^{-1/2} \sum_{k'=1}^{d_g} \sum_{i=1}^{n} \nabla_{\alpha_k g_{k'}} \ell(\alpha_+^*, g_+(X_i); Y_i) g_{+,k'}^{(p+1)}(X_i - c)^{p+1} K_i I_i}_{=(A_2)_k} \\
&\quad + s.o. \\
&= (A_1)_k + (A_2)_k + s.o., \qquad (41)
\end{aligned}$$

where $(A_1)_k$ and $(A_2)_k$ are the k-th elements of A_1 and A_2, respectively, with

$$\begin{aligned}
A_1 &= (nh)^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \nabla_\alpha \ell(\alpha_+^*, g_+(X_i); Y_i) K_i I_i \\ \nabla_g \ell(\alpha_+^*, g_+(X_i); Y_i) K_i I_i \otimes Z_i \end{pmatrix} \\
A_2 &= [(p+1)!]^{-1} (nh)^{-1/2} \sum_{k'=1}^{d_g} \sum_{i=1}^{n} \begin{pmatrix} \nabla_{\alpha g_{k'}} \ell(\alpha_+^*, g_+(X_i); Y_i) g_{+,k'}^{(p+1)}(X_i - c)^{p+1} K_i I_i \\ \nabla_{g g_{k'}} \ell(\alpha_+^*, g_+(X_i); Y_i) g_{+,k'}^{(p+1)}(X_i - c)^{p+1} K_i I_i \otimes Z_i \end{pmatrix}.
\end{aligned}$$

Taking expectation,

$$\begin{aligned}
&\mathbb{E}(A_2)_k \\
&\overset{\text{LIE}}{=} [(p+1)!]^{-1} (nh)^{-1/2} \sum_{k'=1}^{d_g} \sum_{i=1}^{n} \mathbb{E}\{\mathbb{E}[\nabla_{\alpha_k g_{k'}} \ell(\alpha_+^*, g_+(X_i); Y_i) | X_i] g_{+,k'}^{(p+1)}(X_i - c)^{p+1} K_i I_i\} \\
&\overset{u=\frac{X_i - c}{h}}{=} n(nh)^{-1/2} h h^{p+1} \sum_{k'=1}^{d_g} \int_0^1 \mathbb{E}[\nabla_{\alpha_k g_{k'}} \ell(\alpha_+^*, g(c+uh); Y_i) | c+uh] \cdot \\
&\quad \cdot g_{+,k'}^{(p+1)} u^{p+1} K(u) f(c+uh) \mathrm{d}u/(p+1)! + s.o. \\
&\overset{(37)}{=} -n(nh)^{-1/2} h^{p+2} \sum_{k'=1}^{d_g} (\mathcal{I}_{+,\alpha g})_{kk'} g_{+,k'}^{(p+1)}(c) f(c) \int_0^1 u^{p+1} K(u) \mathrm{d}u/(p+1)! + s.o. \\
&= -\overline{B}_{\gamma,+,k} + s.o., \qquad (42)
\end{aligned}$$

where $\overline{B}_{\gamma,+,k}$ is the k-th element of $\overline{B}_{\gamma,+}$. Similar derivation applies to g-parts of A_2. Thus

$$\mathbb{E}A_2 = -\overline{B}_{\gamma,+} + s.o.. \tag{43}$$

By LLN, $A_2 + \overline{B}_{\gamma,+} = o_p(1)$. Lemma 1 then follows from

$$A_1 \xrightarrow{d} \mathcal{N}(0, \Gamma_+). \tag{44}$$

Noting that

$$\mathbb{E}(A_1)_k \overset{\text{LIE}}{=} (nh)^{-1/2} \sum_{i=1}^{n} \mathbb{E}\{\underbrace{\mathbb{E}[\nabla_{\alpha_k} \ell(\alpha_+^*, g_+(X_i); Y_i)|X_i]}_{=0, \text{ by (36)}} K_i I_i\} = 0,$$

and similarly for others part of A_1, we have $\mathbb{E}A_1 = 0$. So

$$\text{Var}[(A_1)_k] = (nh)^{-1/2} \sum_{i=1}^{n} \mathbb{E}[\nabla_{\alpha_k} \ell(\alpha_+^*, g_+(X_i); Y_i)]^2 K_i^2 I_i \to (\Gamma_+)_{kk:1 \le k \le d_\alpha},$$

and similarly for others part of A_1, we have $\text{Var}(A_1) = \Gamma_+$. Then (44) follows from the central limit theorem. The proof of Lemma 1 is complete.

References

Almond, D, J. Doyle Jr., A. Kowalski, and H. Williams (2010): "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns," *Quarterly Journal of Economics*, 125, 591–634.

Arai, A., and H. Ichimura (2015): "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," arXiv:1407.7697v2.

Armstrong, T., and M. Kolesár (2015): "A Simple Adjustment for Bandwidth Snooping," arXiv:1412.0267v3, *Review of Economic Studies*, forthcoming.

Barreca, A., M. Guldi, J. Lindo, and G. Waddell (2011): "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *Quarterly Journal of Economics*, 126, 2117-2123.

Barreca, A., J. Lindo, and G. Waddell (2016): "Heaping-Induced Bias in Regression-Discontinuity Designs," *Economic Inquiry*, 54, 268–293.

Bartalotti, O., and Q. Brummet (2017): "Regression Discontinuity Designs with Clustered Data," *Advances in Econometrics*, Vol. 38, 383-420.

Berger, J., and D. Pope (2011): "Can Losing Lead to Winning?" *Management Science,* 57, 817–827.

Bertanha, M. (2017): "Regression Discontinuity Design with Many Thresholds," Working paper, University of Notre Dame.

Caliendo, M, K. Tatsiramos, and A. Uhlendorff (2013): "Benefit Duration, Unemployment Duration and Job Match Quality: A Regression-Discontinuity Approach," *Journal of Applied Econometrics*, 28, 604–627.

Calonico, S., M. Cattaneo, and M. Farrell (2017): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, forthcoming.

Calonico, S., M. Cattaneo, M. Farrell, and R. Titiunik (2016): "Regression Discontinuity Designs Using Covariates," Working paper, University of Michigan.

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

Calonico, S., M. Cattaneo, and R. Titiunik (2015): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110, 1753-1769.

Card, D., D.S. Lee, Z. Pei, and A. Weber (2015): "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2453-2483.

Cattaneo, M.D., M. Jansson, and X. Ma (2017): "Simple Local Polynomial Density Estimators," Working paper, University of Michigan.

Cattaneo, M.D., and J.C. Escanciano (2017): "Introduction: Regression Discontinuity Designs," in Regression Discontinuity Designs: Theory and Applications, *Advances in Econometrics*, volume 38.

Clark, D., and P. Martorell (2014): "The Signaling Value of a High School Diploma," *Journal of Political Economy*, 122, 282-318.

Clark, D., and H. Royer (2013): "The Effect of Education on Adult Mortality and Health: Evidence from Britain," *American Economic Review*, 103, 2087–2120.

Cohodes, S., and J. Goodman (2014): "Merit Aid, College Quality, and College Completion: Massachusetts' Adams Scholarship as an In-Kind Subsidy," *American Economic Journal: Applied Economics*, 6, 251–285.

Dahl, G., K. Loken, and M. Mogstad (2014): "Peer Effects in Program Participation," *American Economic Review*, 104, 2049–2074.

Dobbie, W. and P. Skiba (2013): "Information Asymmetries in Consumer Credit Markets: Evidence from Payday Lending," *American Economic Journal: Applied Economics*, 5, 256–282.

Dong, Y. (2014): "Jump or Kink? Identification of Binary Treatment Regression Discontinuity Design without the Discontinuity," working paper, UC-Irvine.

Dong, Y. (2015): "Regression Discontinuity Applications with Rounding Errors in the Running Variable," *Journal of Applied Econometrics*, 30, 422-446.

Fan, J., M. Farmen, and I. Gijbels (1998): "Local Maximum Likelihood Estimation and Inference," Journal of the Royal Statistical Society (Series B), 60, 591-608.

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

Fan, J., N. Heckman, and M. Wand (1995): "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions," *Journal of the American Statistical Association*, 90, 141-150.

Feir, D., T. Lemieux, and V. Marmer (2016): "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 34, 185-196.

Frandsen, B., M. Frolich, and B. Melly (2012): "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics*, 168, 382-395.

Gerard, F., and M. Rokkanen, and C. Rothe (2016): "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable, with an Application to Unemployment Insurance in Brazil," Working paper, Columbia University.

Haggag, K., and G. Paci (2014): "Default Tips," *American Economic Journal: Applied Economics*, 6, 1-19.

Hall, P., and J. Racine (2015): "Infinite Order Cross-validated Local Polynomial Regression," *Journal of Econometrics*, 185, 510–525.

Hall, P., R.C.L. Wolff, and Q. Yao (1999): "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Han, A., and J.A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, 5, 1-28.

Hansen, B. (2015): "Punishment and Deterrence: Evidence from Drunk Driving," American Economic Review, 105, 1581–1617.

Henderson, D., S. Kumbhakar, Q. Li, and C. Parmeter (2015): "Smooth Coefficient Estimation of a Seemingly Unrelated Regression," *Journal of Econometrics*, 189, 148-162.

Imbens, G.W., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G.W., and T. Lemieux (2008): "Regression Discontinuity Designs: a Guide to Practice," *Journal of Econometrics*, 142, 615-635.

Koch, S., and J.S. Racine (2016): "Health Care Facility Choice and User Fee Abolition: Regression Discontinuity in a Multinomial Choice Setting," Journal of the Royal Statistical Society (Series A) 179, Part 4, pp. 927–950.

Kolesár, M., and C. Rothe (2017): "Inference in Regression Discontinuity Designs with a Discrete Running Variable." Working paper, Princeton University.

Landais, C. (2015): "Assessing the Welfare Effects of Unemployment Benefits Using the Regression Kink Design," *American Economic Journal: Economic Policy,* 7, 243-78.

Lee, D. S., and D. Card (2008): "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 142, 655–674.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Lindo, J., N. Sanders, and P. Oreopoulos (2010): "Ability, Gender, and Performance Standards: Evidence from Academic Probation," *American Economic Journal: Applied Economics*, 2, 95-117.

Malamud, O., and C. Pop-Eleches (2010): "General Education versus Vocational Training: Evidence from an Economy in Transition," *Review of Economics and Statistics*, 92, 43–60.

McCrary, J. (2008), "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698-714.

McCullagh, P., and J.A. Nelder (1989): *Generalized Linear Models. 2nd Edition.* Chapman and Hall: New York.

Otsu, T., K.-L. Xu, and Y. Matsushita (2013): "Estimation and Inference of Discontinuity in Density," *Journal of Business and Economic Statistics*, 31, 507-524.

Otsu, T., K.-L. Xu, and Y. Matsushita (2015): "Empirical Likelihood for Regression Discontinuity Design," *Journal of Econometrics*, 186, 94-112.

Owen, A. (2001): *Empirical Likelihood.* Chapman and Hall/CRC.

Porter, J., and P. Yu (2015): "Regression Discontinuity Designs with Unknown Discontinuity Points: Testing and Estimation", *Journal of Econometrics*, 189, 132-147.

Shigeoka, H. (2014): "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection," *American Economic Review*, 104, 2152–2184.

Sueyoshi, G. (1995): "A Class of Binary Response Models for Grouped Duration Data," *Journal of Applied Econometrics,* 10, 411-431.

Xu, K.-L. (2016a): "Inference for Local Regression in the Presence of Nuisance Parameters," Working Paper, Indiana University.

Xu, K.-L. (2016b): "Estimation and Inference of Regression Discontinuity Design with Ordered or Discrete Duration Outcomes," Working Paper, Indiana University.

Xu, K.-L., and P.C.B. Phillips (2011): "Tilted Nonparametric Estimation of Volatility Functions with Empirical Applications," *Journal of Business and Economic Statistics* 29, 518-528.

Table 2: Mortality ATEs (for two categories) of the very-low-birth-weight status in California and Texas, 2000-2001. [CI means confidence interval.]

	California	Texas
n	40750	34489
$\widehat{h}$	0.237	0.336
Panel A: τ_1 (One-day mortality)		
ATEs ($\widehat{\tau}$)	-0.0025	-0.0064
90% CI	$[-0.0132, 0.0065]$	$[-0.0181, -0.0009]$
Robust 90% CI	$[-0.0150, 0.0083]$	$[-0.0194, 0.0004]$
t-test P-value	0.576	0.068
Robust t-test P-value	0.636	0.114
Panel B: τ_2 (One-day-to-one-year mortality)		
ATEs ($\widehat{\tau}$)	0.0156	0.0091
90% CI	$[0.0080, 0.0319]$	$[0.0017, 0.0234]$
Robust 90% CI	$[0.0054, 0.0345]$	$[0.0001, 0.0252]$
t-test P-value	0.006	0.056
Robust t-test P-value	0.023	0.099
Panel C: Joint for τ_1 and τ_2		
Wald-test (Significance) P-value	0.020	0.032
Robust Wald-test (Significance) P-value	0.070	0.079
Wald-test (Equivalence) P-value	0.015	0.009
Robust Wald-test (Equivalence) P-value	0.042	0.025

Table 3: Mortality ATEs (for two categories) of the very-low-birth-weight status in California and Texas, 2000-2001, using the *donut* RD estimates (dropping observations with birth weights from 1.485 kg to 1.515 kg). [CI means confidence interval.]

	California	Texas
n	40269	34033
$\widehat{h}$	0.282	0.326
Panel A: τ_1 (One-day mortality)		
ATE ($\widehat{\tau}$)	0.0027	−0.0045
90% CI	$[-0.0078, 0.0129]$	$[-0.0146, 0.0056]$
Robust 90% CI	$[-0.0111, 0.0163]$	$[-0.0166, 0.0075]$
t-test P-value	0.680	0.462
Robust t-test P-value	0.755	0.536
Panel B: τ_2 (One-day-to-one-year mortality)		
ATE ($\widehat{\tau}$)	0.0127	0.0119
90% CI	$[0.0053, 0.0293]$	$[0.0036, 0.0298]$
Robust 90% CI	$[0.0011, 0.0334]$	$[0.0009, 0.0326]$
t-test P-value	0.017	0.035
Robust t-test P-value	0.078	0.082
Panel C: Joint for τ_1 and τ_2		
Wald-test P-value	0.054	0.087
Robust Wald-test P-value	0.199	0.187

Table 4: Mortality ATE (for one combined category) of the very-low-birth-weight status in California and Texas, 2000-2001. [CI means confidence interval.]

	California	Texas
n	40750	34489
$\widehat{h}$	0.310	0.512
$\widehat{\tau}$	0.0101	0.0034
90% CI	[0.0007, 0.0263]	[−0.0036, 0.0174]
Robust 90% CI	[−0.0070, 0.0340]	[−0.0076, 0.0214]
t-test P-value	0.083	0.277
Robust t-test P-value	0.278	0.434

Table 5: Individual mortality ATE (for each of two categories) (estimated by *marginal* local MNL approach) of the very-low-birth-weight status in California and Texas, 2000-2001. [CI means confidence interval.]

	California	Texas
n	40750	34489
Panel A: τ_1 (One-day mortality)		
$\widehat{h}$	0.403	0.327
ATEs ($\widehat{\tau}$)	-0.0021	-0.0064
90% CI	$[-0.0117, 0.0032]$	$[-0.0179, -0.0004]$
Robust 90% CI	$[-0.0148, 0.0063]$	$[-0.0192, 0.0009]$
t-test P-value	0.345	0.084
Robust t-test P-value	0.507	0.132
Panel B: τ_2 (One-day-to-one-year mortality)		
$\widehat{h}$	0.208	0.336
ATEs ($\widehat{\tau}$)	0.0140	0.0091
90% CI	$[0.0044, 0.0305]$	$[0.0019, 0.0236]$
Robust 90% CI	$[0.0018, 0.0332]$	$[0.0010, 0.0245]$
t-test P-value	0.028	0.053
Robust t-test P-value	0.067	0.074

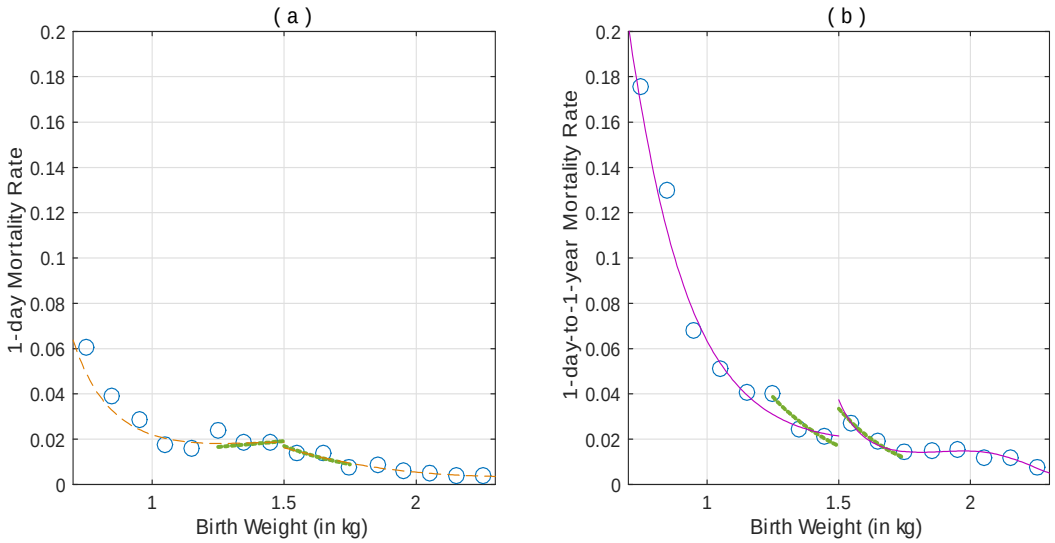

Figure 1: Infant mortality rates in California (two categories), 2000-2001. (a) First-day death; (b) Death between one day and one year. [Curves are cubic multinomial (over the whole range) and linear multinomial (within the $\widehat{h}$-window) fits. Circles are average outcomes within bins of the birth weight (with the bin size 0.1).]

Table 6: Individual Mortality ATE (for each of two categories) (estimated by *marginal* local linear approach) of the very-low-birth-weight status in California and Texas, 2000-2001. [CI means confidence interval.]

	California	Texas
n	40750	34489
Panel A: τ_1 (One-day mortality)		
$\widehat{h}$	0.292	0.302
ATEs ($\widehat{\tau}$)	-0.0019	-0.0067
90% CI	$[-0.0116, 0.0078]$	$[-0.0169, 0.0036]$
Robust 90% CI	$[-0.0132, 0.0102]$	$[-0.0190, 0.0054]$
t-test P-value	0.744	0.284
Robust t-test P-value	0.830	0.360
Panel B: τ_2 (One-day-to-one-year mortality)		
$\widehat{h}$	0.198	0.170
ATEs ($\widehat{\tau}$)	0.0129	0.0145
90% CI	$[-0.0022, 0.0280]$	$[-0.0014, 0.0303]$
Robust 90% CI	$[-0.0063, 0.0302]$	$[-0.0032, 0.0340]$
t-test P-value	0.160	0.133
Robust t-test P-value	0.281	0.174

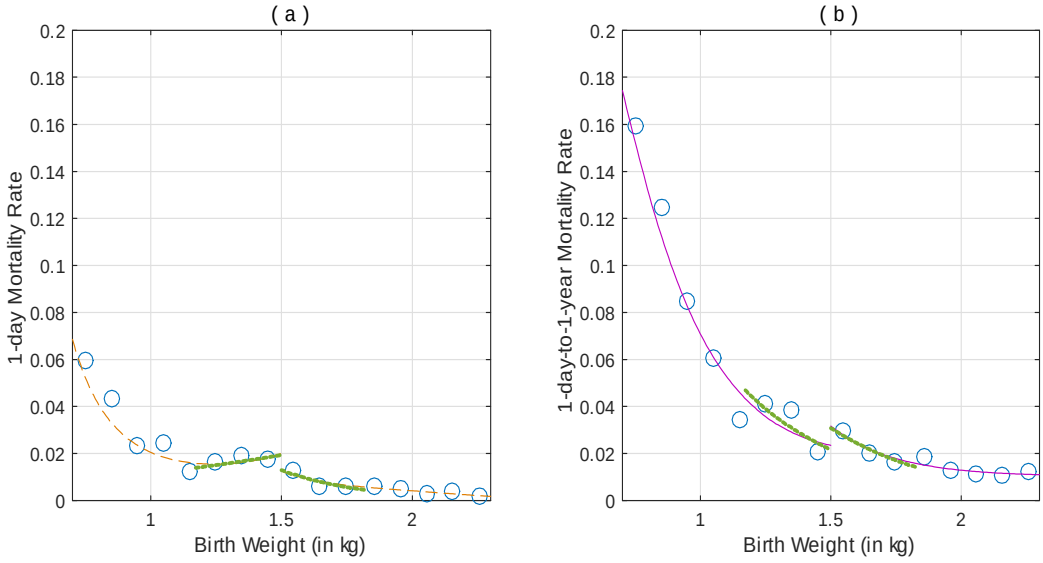

Figure 2: Infant mortality rates in Texas (two categories), 2000-2001. (a) First-day death; (b) Death between one day and one year. [Curves are cubic multinomial (over the whole range) and linear multinomial (within the $\widehat{h}$-window) fits. Circles are average outcomes within bins of the birth weight (with the bin size 0.1).]

Table 7: Simulations: unordered DGPs. [The point estimate and the length of confidence interval are reported as averages over 5,000 replications.]

PANEL A: DGP 1				
	$n = 4000$		$n = 8000$	
	$j = 1$	$j = 2$	$j = 1$	$j = 2$
τ^*	−0.0041	0.0144		
$\widehat{h}$	0.255		0.259	
	(0.098)		(0.092)	
$\widehat{\tau}$	−0.0053	0.0141	−0.0051	0.0138
	(0.0144)	(0.0188)	(0.0089)	(0.0126)
90% CI length	0.0370	0.0492	0.0249	0.0338
90% CI length (rob)	0.0518	0.0682	0.0345	0.0466
90% CI Cov.rate	0.771	0.743	0.772	0.732
90% CI Cov.rate (rob)	0.890	0.860	0.885	0.851
90% Joint CI Cov.rate	0.694		0.690	
90% Joint CI Cov.rate (rob)	0.861		0.852	
PANEL B: DGP 2				
	$n = 4000$		$n = 8000$	
	$j = 1$	$j = 2$	$j = 1$	$j = 2$
τ^*	−0.0085	0.0070		
$\widehat{h}$	0.246		0.266	
	(0.100)		(0.094)	
$\widehat{\tau}$	−0.0083	0.0092	−0.0086	0.0092
	(0.0162)	(0.0177)	(0.0089)	(0.0115)
90% CI length	0.0361	0.0517	0.0229	0.0340
90% CI length (rob)	0.0503	0.0728	0.0325	0.0487
90% CI Cov.rate	0.749	0.808	0.734	0.797
90% CI Cov.rate (rob)	0.869	0.905	0.867	0.894
90% Joint CI Cov.rate	0.721		0.689	
90% Joint CI Cov.rate (rob)	0.870		0.867	

Table 8: Simulations: ordered DGPs. [The point estimate and the length of confidence interval are reported as averages over 5,000 replications.]

PANEL A: DGP 3				
	$n = 4000$		$n = 8000$	
	$j = 1$	$j = 2$	$j = 1$	$j = 2$
τ^*	0.0088	−0.0032		
$\widehat{h}$	0.267		0.273	
	(0.092)		(0.089)	
$\widehat{\tau}$	0.0083	−0.0032	0.0085	−0.0029
	(0.0149)	(0.0193)	(0.0094)	(0.0135)
90% CI length	0.0381	0.0553	0.0260	0.0382
90% CI length (rob)	0.0522	0.0745	0.0364	0.0525
90% CI Cov.rate	0.722	0.791	0.742	0.771
90% CI Cov.rate (rob)	0.843	0.885	0.862	0.869
90% Joint CI Cov.rate	0.682		0.691	
90% Joint CI Cov.rate (rob)	0.846		0.849	
PANEL B: DGP 4				
	$n = 4000$		$n = 8000$	
	$j = 1$	$j = 2$	$j = 1$	$j = 2$
τ^*	0.0014	−0.0063		
$\widehat{h}$	0.272		0.281	
	(0.100)		(0.100)	
$\widehat{\tau}$	0.0015	−0.0047	0.0017	−0.0054
	(0.0133)	(0.0196)	(0.0080)	(0.0136)
90% CI length	0.0338	0.0563	0.0227	0.0387
90% CI length (rob)	0.0468	0.0768	0.0316	0.0534
90% CI Cov.rate	0.744	0.789	0.747	0.771
90% CI Cov.rate (rob)	0.872	0.885	0.873	0.881
90% Joint CI Cov.rate	0.712		0.696	
90% Joint CI Cov.rate (rob)	0.865		0.858	

Table 9: The immediate-dropout effect of being on probation in the first year. [CI means confidence interval.]

	All	Male	Female
n	44362	16981	27381
$\widehat{h}$	0.877	0.445	0.473
$\widehat{\tau}$	−0.0118	−0.0555	−0.0011
90% CI	$[-0.0256, -0.0047]$	$[-0.0894, -0.0365]$	$[-0.0130, 0.0215]$
90% CI (robust)	$[-0.0314, 0.0011]$	$[-0.0898, -0.0361]$	$[-0.0145, 0.0230]$
t-test P-value	0.016	0.000	0.684
Robust t-test P-value	0.124	0.000	0.709

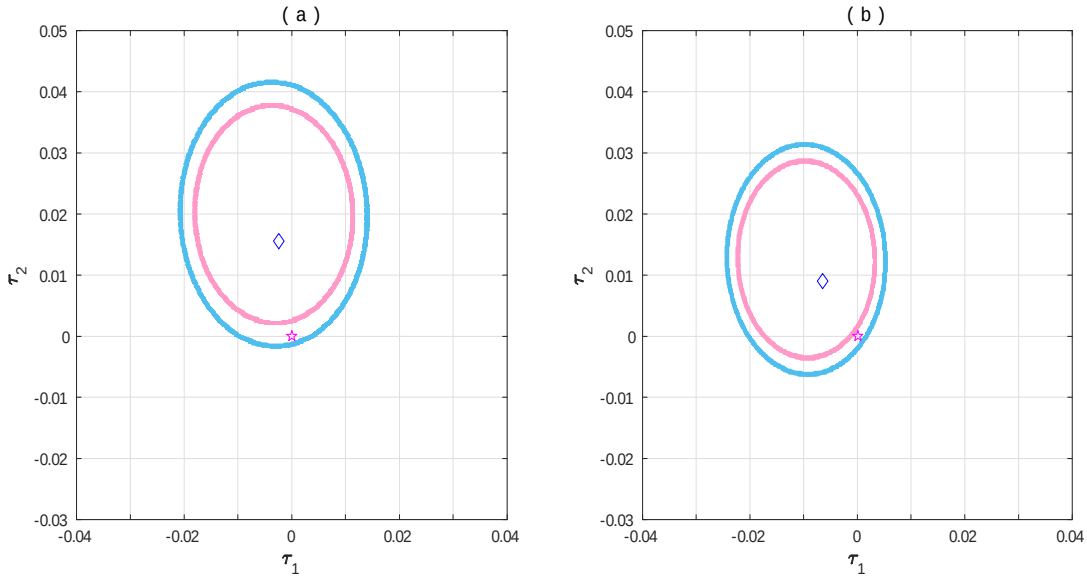

Figure 3: The 95% Confidence Regions: (a) California (b) Texas. [Notes: The outer ellipse is the robust confidence region. The inner ellipse is the standard (non-robust) confidence region. The diamond ◊ denotes the point estimate $(\widehat{\tau}_1, \widehat{\tau}_2)$, and the star denotes the origin (0,0).]

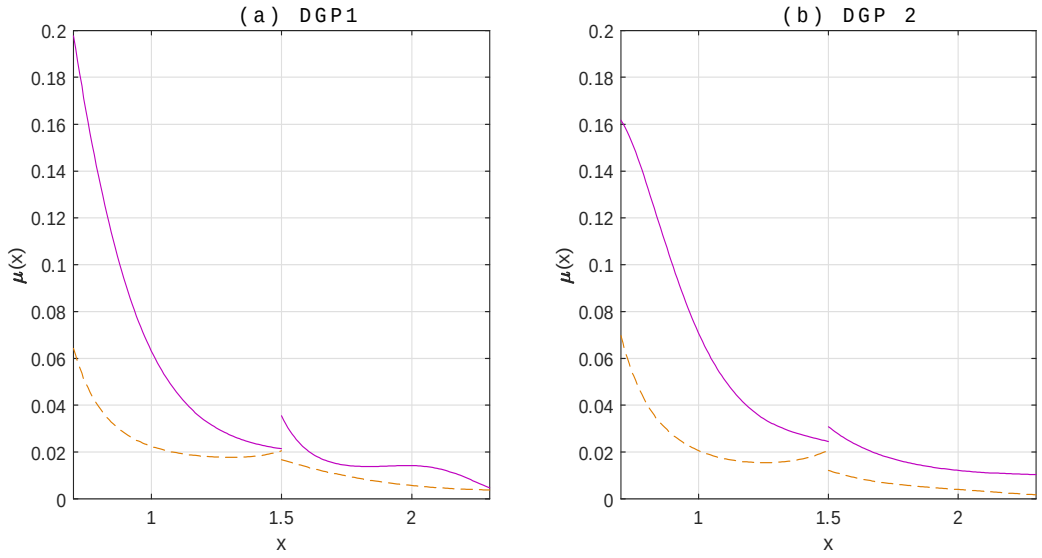

Figure 4: DGPs 1-2 in simulations. The figures show $\mu_{+,1}(x)$ and $\mu_{-,1}(x)$ (dash lines), and $\mu_{+,2}(x)$ and $\mu_{-,2}(x)$ (solid lines). The function form: $\mu(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \sin x$.

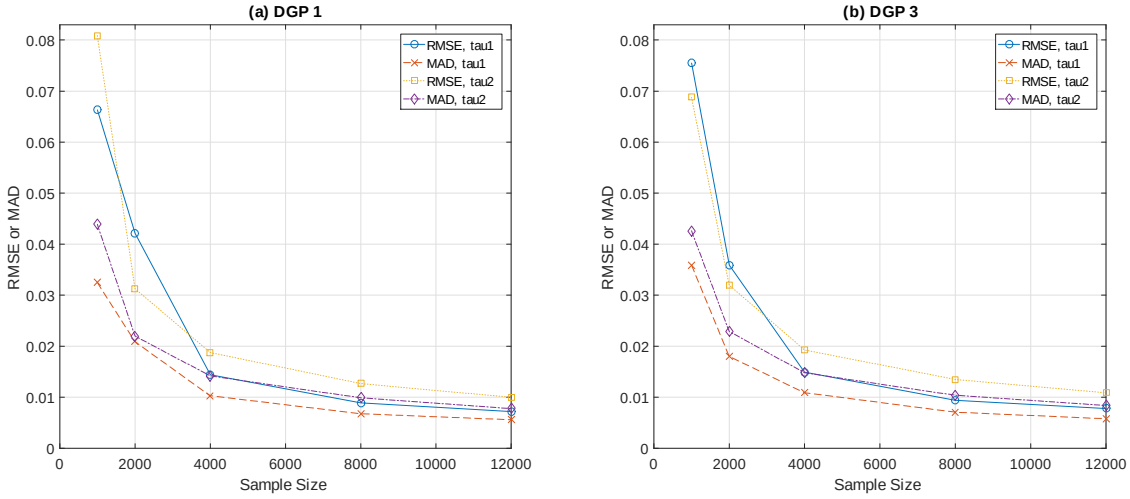

Figure 5: RMSEs (square root of mean square errors) and MADs (mean absolute deviations) of the point estimators $\widehat{\tau}_1$ and $\widehat{\tau}_2$ for a variety of sample sizes in simulations.

Table 10: The graduation-rate effect of being on probation in the first year. [CI means confidence interval.]

Panel A: Data-determined bandwidth			
n	18983		
$\widehat{h}$	0.385		
Wald-test P-value	0.193		
Rrobust Wald-test P-value	0.231		
	Within 4 Years	In the 5th Year	In the 6th Year
$\widehat{\tau}$	0.0482	−0.0099	−0.0052
90% CI	[0.0134, 0.1073]	[−0.0731, 0.0203]	[−0.0316, 0.0329]
90% CI (robust)	[0.0111, 0.1096]	[−0.0754, 0.0227]	[−0.0333, 0.0346]
t-test P-value	0.034	0.353	0.973
Robust t-test P-value	0.043	0.376	0.974
Panel B: Fixed bandwidth			
n	18983		
$\widehat{h}_{\text{LSO}}$	0.600		
Wald-test P-value	0.258		
Robust Wald-test P-value	0.438		
	Within 4 Years	In the 5th Year	In the 6th Year
$\widehat{\tau}$	0.0134	0.0196	−0.0108
90% CI	[0.0062, 0.0812]	[−0.0614, 0.0146]	[−0.0226, 0.0283]
90% CI (robust)	[−0.0021, 0.0895]	[−0.0697, 0.0230]	[−0.0283, 0.0341]
t-test P-value	0.055	0.311	0.852
Robust t-test P-value	0.116	0.406	0.879

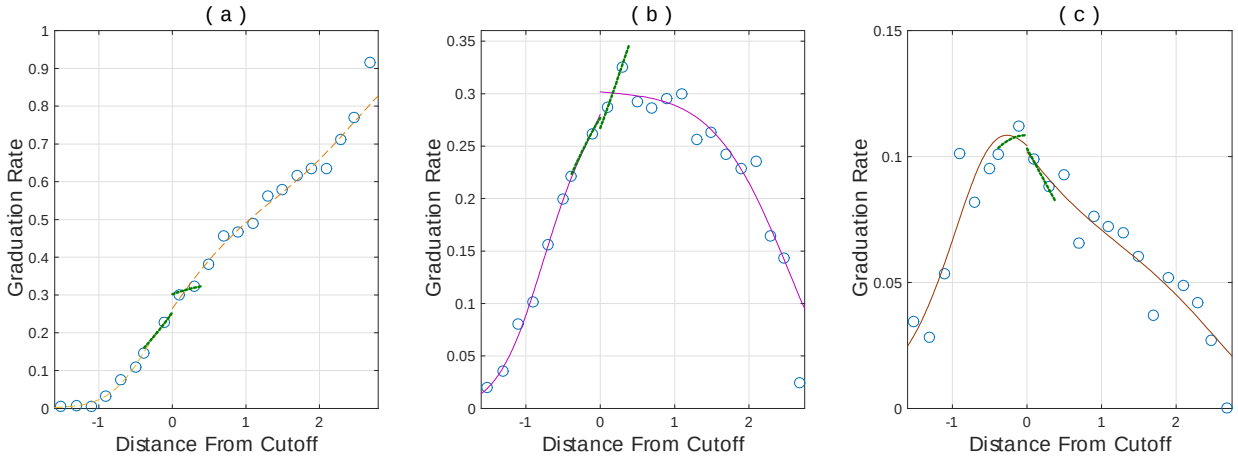

Figure 6: The graduation-rate effect of being placed on probation. (a) Graduated within 4 years; (b) Graduated in the 5th year; (c) Graduated in the 6th year. [Curves are cubic multinomial (over the whole range) and linear multinomial (within the $\widehat{h}$-window) fits. Circles are average outcomes within bins of the first-year GPA (with the bin size 0.2).]