PA

Ecological Regression with Partial Identification

Wenxin Jiang[1,2], Gary King[3], Allen Schmaltz[3] and Martin A. Tanner[2]

[1] Institute of Finance (Adjunct), Shandong University, Jinan, Shandong, China. Email: wjiang@northwestern.edu
[2] Department of Statistics, Northwestern University, Evanston, IL, USA. Email: mat132@northwestern.edu
[3] Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. Email: king@harvard.edu, schmaltz@fas.harvard.edu

Abstract

Ecological inference (EI) is the process of learning about individual behavior from aggregate data. We relax assumptions by allowing for "linear contextual effects," which previous works have regarded as plausible but avoided due to nonidentification, a problem we sidestep by deriving bounds instead of point estimates. In this way, we offer a conceptual framework to improve on the Duncan–Davis bound, derived more than 65 years ago. To study the effectiveness of our approach, we collect and analyze 8,430 2×2 EI datasets with known ground truth from several sources—thus bringing considerably more data to bear on the problem than the existing dozen or so datasets available in the literature for evaluating EI estimators. For the 88% of real data sets in our collection that fit a proposed rule, our approach reduces the width of the Duncan–Davis bound, on average, by about 44%, while still capturing the true district-level parameter about 99% of the time. The remaining 12% revert to the Duncan–Davis bound.

Keywords: asymptotics, bounds, confidence intervals, contextual models, ecological inference, linear regression, partial identification

1 Introduction

Ecological inference (EI) is the task of reconstructing individual behavior from aggregate data or, more specifically, making inferences about a conditional probability distribution when only its marginal distributions are known. As a simple example, suppose in each precinct in the United States, we observe from election results the proportion of people who turn out T_i and from census data the proportion of people who are African American ("black"), X_i. Our goal then is to estimate the cells of the vote/no vote $\times$ black/nonblack cross-tabulation at the district level—with values including the percent of blacks who turn out and the percent of nonblacks who turn out—even though the secret ballot makes it impossible to calculate these cell values directly. EI has numerous applications in many fields, with more complex cases having more than two categories for one or more of the variables, but the same basic issues apply (King 1997, Section 1.1).

The early literature on EI introduced separate deterministic and statistical approaches for estimating the cell values. Duncan and Davis' (1953) *deterministic* approach (hereafter "DD") is to bound the cell entries with no assumptions other than the veracity of the data. For an extreme example, if everyone in a precinct is black, then the percent of black people who turn out to vote is known exactly. Although sometimes useful, as DD bounds are guaranteed to capture the true values, they are often wide and thus not sufficiently informative. In contrast, Goodman's (1953) *statistical* approach ignores information in the deterministic bounds; assumes independence among X_i, the precinct-level cell entries, and the number of people in each precinct (which together we refer to as the "standard EI assumptions"); and can then generate an unbiased estimate of the average cell values from a regression of T_i on X_i and $1 - X_i$ (with no constant term). Unlike DD, Goodman's approach provides sharp point estimates that are consistent under

Authors' note: We thank the editor and anonymous reviewers for their helpful comments. This work was partially supported by the Taishan Scholar Construction Project to W.J. and by the Institute for Quantitative Social Science.

***Political Analysis* (2019)**

DOI: 10.1017/pan.2019.19

Corresponding author
Allen Schmaltz

Edited by
Jeff Gill

these assumptions, but it usually results in highly model dependent inferences, often far outside of the DD bounds and even the unit interval.

These two streams of research merged when King (1997) developed the first model that included information from both precinct-level DD bounds (varying over i) and cross-precinct statistical information. King's Bayesian model makes standard EI independence assumptions a priori but incorporates precinct-level bounds information so that parameters and estimates can be a posteriori dependent, which also guarantees that all estimates (at the aggregate and precinct level) are always within their bounds and these bounds, in turn, are used to improve the statistical estimates. King (1997, Chapter 9) also proposed a "contextual effects" extension to weaken the standard EI independence assumptions in which cell entries are parametric functions of X_i, with precinct-level bounds sufficient for (weak) identification. A rich methodological literature has built on these developments with numerous applications appearing across many fields and disciplines, each of which now includes both statistical and deterministic information (see many examples in King, Rosen, and Tanner 2004).

In this paper, we return to the contextual effect approach, allowing the race-specific probability of voting to depend linearly on X_i, with the slope coefficient representing the linear contextual effect. Although this approach addresses the most consequential violation of the standard EI assumptions, it has well-known identification challenges (e.g., Owen and Grofman 1997; Chambers and Steel 2001; Wakefield 2004). Although we include precinct-level information, much uncertainty remains about the precise values of the contextual effect parameters. We show, however, that this problem fits easily in the framework of "interval data regression," where we regress the varying precinct-level DD bounds on the race proportion in each precinct. Although interval data regression does not fully identify the regression coefficients, it can provide identification regions or bounds (see, e.g., Chernozhukov, Hong, and Tamer 2007; Liao and Jiang 2010). We apply this technique to bound the nonidentified regression parameter in the linear contextual model and then use that information to improve the DD bounds of the quantities of interest.

Like DD, our approach also has no adjustable parameters, which makes it easy to use and robust to claims of hacking: the researcher simply inputs a (sensible) set of ecological data and the method returns accurate bounds on the quantities of interest, usually much more informative than given by DD. However, the bound is no longer model-free as is DD. This leads to two issues in using this method. First, the new bound depends on a linear contextual effect assumption. Violations of the assumptions can cause the bound on the quantity of interest to miss the true district voting proportion, or to even be empty. Second, even if the assumptions hold, the implied regression bound is only derived in the limit of large p (the number of precincts) and can still miss the true district-level voting proportion by an amount on the order of $1/\sqrt{p}$.

To address the second concern, we increase the implied regression bound by a multiple of the standard errors on both sides (similar to forming a confidence interval) before intersecting with the DD bound. To address the first concern, we select only datasets where the implied regression bound has a nonempty intersection with the DD bound. These two ideas together turn out to produce highly accurate estimates for the 8,430 datasets that we have constructed from census and other data sources, where ground truth is known. For most of the datasets, the resulting bounds become much shorter than the DD bound, yet still contain the true district-level proportion. We have made our datasets publicly available via the Harvard Dataverse (Jiang *et al.* 2019) and will add to them over time as a useful resource for researchers in applying or improving EI.

Of course, our datasets may not be representative of every dataset that researchers choose to analyze in the future, and the performance measures may thus differ for different collections of datasets. Also, even the linear contextual effect assumption and our new estimator together do not always overcome the intractable inferential problem posed by information sometimes lost via aggregation, such as due to the secret ballot in the context of vote choice. For example, the proposed interval may be too wide to be informative when both aggregate variables are near 0.5 and have little variation across precincts. In some other datasets, the proposed method produces bounds substantially tighter than DD. Limitations of the proposed method are described in Section 7.1.

We begin by defining the linear contextual model in Section 2 and explain why some of the regression coefficients are not identified. We describe how to bound the unidentified regression coefficients in Section 3 and how to bound the district-level voting proportions in Section 4. In Section 5, we introduce confidence intervals for the bounds to account for finite sample variation. In Section 6, we provide extensive analytic, simulated, and real data examples. In Section 7, we discuss the generality and limitation of the proposed model, offer comparisons with fully identified models based on assumptions, and offer suggestions for future research. Technical details appear in the Supplementary Appendix.

2 The Linear Contextual Effects Model

We now describe our data and quantity of interest (Section 2.1), introduce the nonidentified linear contextual effects model (Section 2.2), give a simple example (Section 2.3), and reveal the conflicting assumptions in the literature that have been suggested for how to achieve identification (Section 2.4).

2.1 Data and Quantities of Interest

We begin with the EI "accounting identity" (i.e., true by definition) for precinct i ($i = 1, \ldots, p$):

$$T_i = X_i \beta_i^b + (1 - X_i)\beta_i^w. \tag{1}$$

Following our running example, T_i is the proportion of people in precinct i turning out to vote, X_i is the proportion of people in the precinct who are "black" (defined as nonwhite), β_i^b is the proportion of black people who turn out to vote, and β_i^w is the proportion of white people who turn out to vote.

Although we would like to know β_i^b and β_i^w for every precinct $i = 1, \ldots, p$, the quantities of interest for this paper will be limited to the district-level proportion of blacks and whites who vote, respectively:

$$B \equiv \sum_{i=1}^{p} N_i X_i \beta_i^b \Big/ \sum_{i=1}^{p} N_i X_i \quad W \equiv \sum_{i=1}^{p} N_i (1 - X_i)\beta_i^w \Big/ \sum_{i=1}^{p} N_i (1 - X_i), \tag{2}$$

where N_i is the total number of people in precinct i.

These quantities are related to each other, after conditioning on T_i, by the accounting identity at the district level:

$$W \sum_{i=1}^{p} N_i (1 - X_i) + B \sum_{i=1}^{p} N_i X_i = \sum_{i=1}^{p} N_i T_i,$$

so one can be derived from the other. Therefore, we focus only on the inference about B from here on.

2.2 The (Nonidentified) Model

We now allow "contextual effects", which in this example means that the race-specific turnout proportions (β_i^b and β_i^w) are allowed to depend on the "context" (e.g., the black proportion X_i). The only essential assumption we make in this paper is as follows:

ASSUMPTION 1 (Linear contextual effects). The random vector $(\beta_i^b, \beta_i^w, X_i, N_i)$ is independent and identically distributed over i, for $i = 1, \dots, p$, and satisfies

$$E(\beta_i^w | X_i, N_i) = w_0 + w_1 X_i \tag{3}$$

and

$$E(\beta_i^b | X_i, N_i) = b_0 + b_1 X_i, \tag{4}$$

where w_0, w_1, b_0, b_1 are nonrandom real parameters.

The current form of the assumption implies that $E(\beta_i^{b,w}|X_i, N_i) = E(\beta_i^{b,w}|X_i)$ (where $\beta_i^{b,w}$ means that this expression holds for β_i^b and also β_i^w), and, therefore, N_i can be omitted. Supplementary Appendix C explains how the effect of N_i can be included in the regression. In the main text, however, we adopt this simpler form of the model omitting N_i since this retains the essential feature of partial identification and it is more common in the literature (see, e.g., Achen and Shively 1995; King 1997, Section 3.2; Altman, Gill, and McDonald 2004).

Under these assumptions, $(\beta_i^b, \beta_i^w, X_i)$ from each precinct is a vector of random variables sampled from an underlying probability distribution. The conditional expectations $E(\beta_i^{b,w}|X_i)$ are taken over the conditional distribution of $\beta_i^{b,w}$, given X_i, which allows for $\beta_i^{b,w}$ to still be random even after fixing the values of X_i. For example, precincts with similar X_i's (e.g., around 0.5) can still have very different race-specific voting proportions, β_i^b or β_i^w.

Under the assumptions regarding $E(\beta_i^{b,w}|X_i)$, the accounting identity (1) now implies a quadratic regression:

$$E(T_i|X_i) = w_0 + (b_0 - w_0 + w_1)X_i + (b_1 - w_1)X_i^2 \tag{5}$$

$$= w_0 + c_1 X_i + d_1 X_i^2, \tag{6}$$

where

$$c_1 = b_0 - w_0 + w_1 \quad \text{and} \quad d_1 = b_1 - w_1 \tag{7}$$

are the coefficients of X_i and X_i^2, respectively. It then follows that the three parameters (w_0, c_1, d_1) are identifiable (if the X_i's can take three or more distinct values) and can be estimated by (possibly weighted) least squares regression.

The four regression parameters in the linear contextual effects model are related to the three regression parameters in the quadratic regression of T_i vs X_i via

$$(w_0, w_1, b_0, b_1) = (w_0, w_1, c_1 + w_0 - w_1, d_1 + w_1), \tag{8}$$

which are partially identifiable up to one free parameter: (w_0, c_1, d_1) are identified, but w_1 is not.[1]

2.3 A Simple Example

The nonidentifiability of this model described here is well-known, for example, when $b_1 = w_1$, and the resulting T_i, X_i relation is linear (see Freedman *et al.* 1991; King 1997, Section 3.2).

Figure 1 offers a slightly different example to illustrate the nonidentification problem, which we also use in several places below (see Sections 3.1 and 6.1.1). In this example, we observe voter

1 Although we focus on bounding w_1 in this paper, we could have also chosen $b_1 = w_1 + d_1$ as the nonidentified parameter instead. The results are equivalent due to the accounting identity (1). However, in that case, a composite parameter $b_0 + b_1$ (instead of simply w_0) is identifiable, and the notation would be more complex with no additional benefit.

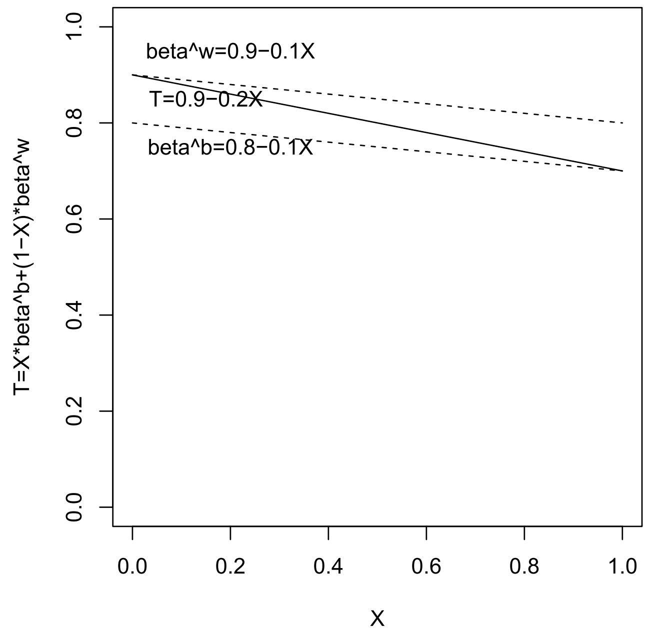

Figure 1. A simple example of nonidentification. The solid line is the total observed voter turnout T_i. The underlying race-specific proportions $\beta_i^{b,w}$ can either both follow the same solid line $0.9 - 0.2x$ (with slope $w_1 = -0.2$) or separately follow the two dashed lines $0.9 - 0.1x$ and $0.8 - 0.1x$ (with slope $w_1 = -0.1$). The slope parameter w_1 (for β_i^w) is therefore not identified. (Another possible value $w_1 = 0$ corresponds to a scenario of constant $(\beta_i^b, \beta_i^w) = (0.7, 0.9)$.)

turnout T_i declining linearly as the black percentage of the precinct, X_i, increases (see the solid black line). However, the reason for this relationship here is not necessarily determined solely from these two marginal variables. It could be that individual black citizens have lower voter turnout than white citizens and so the increasing percentage of black citizens leads to overall turnout declines. Alternatively, it could instead be that the white citizens who live in precincts with many black citizens tend to vote less, for cultural or economic reasons. In the figure, we convey the "contextual dependence" that would not normally be observable, which is how β_i^w and β_i^b vary over precincts; in this case, both declining as X increases. This context may reflect a situation in which precincts with more black citizens happen to be from poorer, inner city areas (as a result, for example, of structural discrimination), with more residential mobility and hence lower turnout.

2.4 Conflicting Advice on Identifying Assumptions

The key problem, then, is that the linear contextual effects model has four parameters, but the derived quadratic regression of the observed T_i versus X_i can only identify three of them. Existing works have addressed this issue, providing at times conflicting advice. In particular, scholars have suggested treating the nonidentified parameter w_1 by setting $w_1 = \max\{-d_1, 0\}$ (Achen and Shively 1995; Altman, Gill, and McDonald 2004), $w_1 = 0$ (Wakefield 2004, Section 1.2), and $w_1 = -d_1/2$ (Wakefield 2004, Section 1.2). Of course, the advice in each case is given with warnings and is appropriate in some circumstances but are neither universally appropriate nor come with decision rules to help researchers decide when to use each one. Ultimately, each of these assumptions is arbitrary, meaning that the results using it give answers that are highly model dependent. In real applications, these assumptions can make a major substantive difference in empirical results.

The approach we introduce below differs in an important respect from this literature. Instead of arbitrarily picking a value for w_1 and hoping it applies across datasets or to the one before us, we derive a prior-insensitive bound for w_1 under the current linear contextual effects model, using the

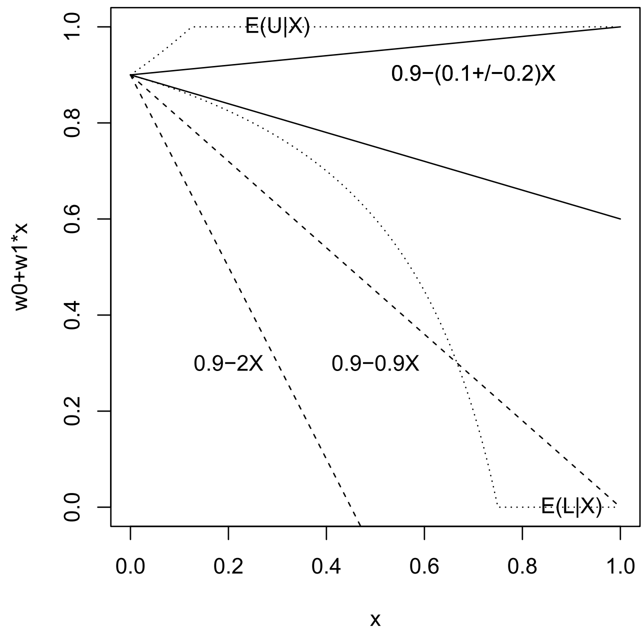

Figure 2. Intuition for bounding w_1. The dotted curves are the expectations of the DD bounds for β_i^w for the simple example of Section 2.3. The solid lines $0.9 - (0.1 \pm 0.2)x$ are obtained by forcing linear contextual effects $E(\beta_i^w|X_i = x) = w_0 + w_1 x$ to lie between the dotted curves. The dashed lines are examples exceeding the expectation of the DD upper bound (see Section 3.1).

expectations of the DD bounds conditional on the X_i's. Our approach is conditional on the linearity of the contextual effects model but should be relatively robust to many types of deviations from linearity.

3 Contextual Model Parameter Bounds

We now offer intuition, followed by more formal theory, for how to bound the nonidentified parameter of the contextual model parameter. In Section 4, we show how to use this result to bound the district-level quantity of interest.

3.1 Intuition

Denote the DD bounds for the unobserved β_i^w as $L_i \le \beta_i^w \le U_i$, where $L_i \equiv \max\{0, (T_i - X_i)/(1 - X_i)\}$ and $U_i \equiv \min\{1, T_i/(1 - X_i)\}$. Under the linear contextual model $E(\beta_i^w|X_i) = w_0 + w_1 X_i$, the observable DD bounds $L_i \le \beta_i^w \le U_i$ form a problem of interval data regression, regressing $[L_i, U_i]$ against X_i. It is well-known (see, e.g., Chernozhukov, Hong, and Tamer 2007; Liao and Jiang 2010) that although interval data regression cannot fully identify the regression coefficients, it can provide their identification regions or bounds. We use this perspective to derive a bound for the nonidentified regression coefficient w_1.

Taking expectations under the linear contextual model gives the corresponding bound in the conditional expectation, $E(L_i|X_i) \le E(\beta_i^w|X_i) \le E(U_i|X_i)$, or

$$E(L_i|X_i) \le w_0 + w_1 X_i \le E(U_i|X_i). \tag{9}$$

These bounds are identifiable from observable quantities. Forcing this bound in the entire domain of X_i leads to a bound for w_1.

Consider the simple example from Section 2.3, where $T_i = 0.9 - 0.2X_i$.

In Figure 2, we illustrate the intuition of how to bound the slope parameter w_1 in the linear contextual model $E(\beta_i^w|X_i = x) = w_0 + w_1 x$ for all $x \in (0,1)$. The intercept parameter is identifiable as $w_0 = E(T_i|X_i = 0) = 0.9$. The slope parameter w_1 is nonidentified, but only partially so. There are hidden constraints: if the line $w_0 + w_1 x = 0.9 - 2x$, then the probability $E(\beta_i^w|X_i = x) = w_0 + w_1 x$ can be negative; so w_1 cannot be as low as -2 ($w_1 \geq -2$). Even if we choose $w_0 + w_1 x = 0.9 - 0.9x$ so that $E(\beta^w|X_i = x) = w_0 + w_1 x$ falls between $[0,1]$ for all $x \in (0,1)$, the line $0.9 - 0.9x$ can still penetrate a large portion of the dotted curve of the expectation of the DD lower bound $E(L_i|X_i = x)$. As such, w_1 also cannot be as low as -0.9 ($w_1 \geq -0.9$). In fact, to force $w_0 + w_1 x$ to fall between the dotted curves $[E(L_i|X_i = x), E(U_i|X_i = x)]$ for all $x \in (0,1)$, we need to have $w_1 \in [-0.3, 0.1] = -0.1 \pm 0.2$, restricted to a small interval in this example. (Incidentally, this bound includes all the three possibilities $w_1 = -0.2$, $w_1 = -0.1$, and $w_1 = 0$, as described in the original example in Figure 1.)

Intuitively, this is how we exploit the expectation of the DD bound (9) to bound the nonidentified contextual effect parameter w_1. More formally, we have the following theoretical results.

3.2 Theory

The following proposition provides a necessary and sufficient condition for this bound in terms of the only nonidentified parameter w_1.

PROPOSITION 1. *Assume a linear contextual effect $E(\beta_i^w|X_i) = w_0 + w_1 X_i$ for all $X_i \in A$ where $A \subset (0,1)$. Then*

$$E(L_i|X_i) \leq E(\beta_i^w|X_i) \leq E(U_i|X_i),$$

for all $X_i \in A$, if and only if the nonidentifiable parameter w_1 satisfies

$$\sup_{X_i \in A}[(E(L_i|X_i) - w_0)/X_i] \leq w_1 \leq \inf_{X_i \in A}[(E(U_i|X_i) - w_0)/X_i].$$

PROOF. If $\sup_{X \in A}[(E(L|X) - w_0)/X] \leq w_1 \leq \inf_{X \in A}[(E(U|X) - w_0)/X]$ holds, then for all $X \in A$, $[(E(L|X) - w_0)/X] \leq w_1 \leq [(E(U|X) - w_0)/X]$. This implies $E(L|X) \leq w_0 + w_1 X \leq E(U|X)$ for all $X \in A \subset (0,1)$.

For the converse: $E(L|X) \leq w_0 + w_1 X \leq E(U|X)$ for all $X \in A \subset (0,1)$ implies $[(E(L|X) - w_0)/X] \leq w_1 \leq [(E(U|X) - w_0)/X]$ holds for all $X \in A$. Now, we take $\inf_{X \in A}$ for both sides of $w_1 \leq [(E(U|X) - w_0)/X]$ and take $\sup_{X \in A}$ for both sides of $[(E(L|X) - w_0)/X] \leq w_1$. □

The above proposition then gives the tightest bound possible on w_1. The upper bound and the lower bound are both constructed out of identifiable quantities. The functions $E(L_i|X_i)$ and $E(U_i|X_i)$ may be estimated by lowess smoothing. If for some reason we would like to avoid such nonparametric estimation (e.g., it may not perform well at boundary values of X_i), we can relax the bounds somewhat and incorporate results from a parametric regression $E(T_i|X_i) = w_0 + c_1 X_i + d_1 X_i^2$.

PROPOSITION 2. *For all $X_i \in [l,u] \subset (0,1)$ where $l < u$, assume linear contextual effect $E(\beta_i^w|X_i) = w_0 + w_1 X_i$ and a quadratic regression $E[T_i|X_i] = w_0 + c_1 X_i + d_1 X_i^2$. Then we have*

$$wl \leq w_1 \leq wu,$$

where $wl = \max_{x \in \{l,u\}} \max\{-w_0/x, (w_0 + c_1 + d_1 - 1)/(1-x) - d_1\}$ and $wu = \min_{x \in \{l,u\}} \min\{(1 - w_0)/x, (w_0 + c_1 + d_1)/(1-x) - d_1\}$.

Before we prove this proposition, we give some intuition on why the proposed bounds make sense. The bounds are obtained by forcing $w_0 + w_1 x$ and $b_0 + b_1 x$ (for all $x \in [l, u]$) to be within $[0, 1]$ since these linear combinations model the $[0, 1]$-valued probabilities (via $E(\beta^{w,b}|X = x)$). For example, one of the proposed lower bounds of w_1 is of the form $-w_0/x$; this makes $w_1 \geq -w_0/x$ and therefore $w_0 + w_1 x \geq 0$. One of the upper bounds of w_1 is $(1 - w_0)/x$; this makes $w_1 \leq (1 - w_0)/x$ and therefore $w_0 + w_1 x \leq 1$. Similarly, the other two functions in the bounds are related to the bounds of $b_0 + b_1 x$ by using (8). Now we prove the proposition rigorously.

Proof. For the bounds in Proposition 1, note that $E(U_i|X_i) = E[\min\{1, T_i/(1 - X_i)\}|X_i] \leq \min\{1, E[T_i|X_i]/(1 - X_i)\}$ due to Jensen's inequality, and similarly $E(L_i|X_i) \geq \max\{0, (E[T_i|X_i] - X_i)/(1 - X_i)\}$. Now apply a quadratic regression $E[T_i|X_i] = w_0 + c_1 X_i + d_1 X_i^2$. Then from Proposition 1, we have

$$\sup_{X_i \in A} \max\{-w_0/X_i, (w_0 + c_1 - 1 + d_1 X_i)/(1 - X_i)\} \\ \leq w_1 \leq \inf_{X_i \in A} \min\{(1 - w_0)/X_i, (w_0 + c_1 + d_1 X_i)/(1 - X_i)\}.$$

Simplifying these bounds for $A = [l, u]$ with the boundary points leads to the proof. □

To use Proposition 2, we need to supply the interval $[l, u]$ where we believe the assumptions hold. One could simply use the data range $l = \min X_i$ and $u = \max X_i$ of the dataset. However, there may be reasons to either reduce this range (e.g., if there are outliers) or increase this range (if there is a belief that the pattern could be reliably extrapolated to some extent beyond the data range). If we attempt to check the assumptions when there is no knowledge regarding the ground truth β_i^w, we could still use the (T_i, X_i) data to fit a quadratic curve on $(0, 1)$ and superimpose it on the scatterplot, using it to rule out unreasonable choices of a range $[l, u]$. For example, when quadratic regression is based on a scatterplot limited in a small domain of $X_i \in [0.5, 0.6]$ and extrapolating the fitted quadratic curve to $x \in [0.1, 0.9]$ leads to $E(T|X = x) = w_0 + c_1 x + d_1 x^2$ breaking the "ceiling" of 1 or the "floor" of 0, then it is obvious that the range $[l, u] = [0.1, 0.9]$ is too wide.

On the other hand, the bigger the set $A = [l, u]$ is for X_i, the tighter the bounds will be in these propositions. Suppose we consider a special case $A \to (0, 1)$. In other words, we assume that the previous quadratic regression model holds for all X_i in the whole range of $(0, 1)$. Then relaxing the bounds of Proposition 2 and taking $l \to 0, u \to 1$, we immediately have the following:

Proposition 3. *For all $X_i \in (0, 1)$, assume linear contextual effect $E(\beta_i^w|X_i) = w_0 + w_1 X_i$ and a quadratic regression $E[T_i|X_i] = w_0 + c_1 X_i + d_1 X_i^2$. Then we have*

$$wl \leq w_1 \leq wu,$$

where $wl = \max\{-w_0, c_1 + w_0 - 1\}$ and $wu = \min\{1 - w_0, c_1 + w_0\}$.

4 District Cell Value Bounds

Section 3 derives bounds for the contextual effect parameter w_1. Our ultimate quantity of interest is the district-level, race-specific vote proportions—the unobserved cell values of the cross-tabulation, B and W. In this section, we derive these bounds, given the bounds on w_1.

4.1 Estimating District-Level Parameters

We first analyze the precinct-level parameter β_i^b. Denote residuals as $e_i^b = \beta_i^b - E(\beta_i^b|X_i)$, $e_i^w = \beta_i^w - E(\beta_i^w|X_i)$. Note that

$$\begin{aligned}
\beta_i^b &= E(\beta_i^b|X_i) + e_i^b && (10)\\
&\overset{\text{by (8)}}{=} w_0 + (c_1 - w_1) + (w_1 + d_1)X_i + e_i^b && (11)\\
&= [w_0 + c_1 + d_1 X_i] + w_1(X_i - 1) + e_i^b && (12)\\
&\equiv b_i(w_1, \theta) + e_i^b, && (13)
\end{aligned}$$

where $\theta \equiv (w_0, c_1, d_1)^T$.

The *district-level parameter* is given as

$$\begin{aligned}
B &\equiv \sum_{i=1}^{p} N_i X_i \beta_i^b / \sum_{i=1}^{p} N_i X_i && (14)\\
&= \frac{\sum_{i=1}^{p} N_i X_i b_i(w_1, \theta)}{\sum_{i=1}^{p} N_i X_i} + \frac{\sum_{i=1}^{p} N_i X_i e_i^b}{\sum_{i=1}^{p} N_i X_i}. && (15)
\end{aligned}$$

By the Law of Large Numbers, for large p, we can ignore the second term of the expansion of (14) with the mean zero residuals e_i^b, when estimating B. We can then form a *point estimate* of B using the first term:

$$\begin{aligned}
B(w_1, \theta) &\equiv \frac{\sum_{i=1}^{p} N_i X_i b_i(w_1, \theta)}{\sum_{i=1}^{p} N_i X_i}\\
&= \frac{\sum_{i=1}^{p} N_i X_i b_i(0, \theta)}{\sum_{i=1}^{p} N_i X_i} - w_1 \frac{\sum_{i=1}^{p} N_i X_i (1 - X_i)}{\sum_{i=1}^{p} N_i X_i}, && (16)
\end{aligned}$$

where

$$b_i(w_1, \theta) \equiv [w_0 + c_1 + d_1 X_i] + w_1(X_i - 1). \quad (17)$$

4.2 Sensitivity of District Cell Value Estimate

The point estimate $B(w_1, \theta)$ will vary with w_1 due to (16). The sensitivity on w_1 can be measured by

$$\frac{\partial B(w_1, \theta)}{\partial w_1} = -r \equiv -\frac{\sum_{i=1}^{p} N_i X_i (1 - X_i)}{\sum_{i=1}^{p} N_i X_i}, \quad (18)$$

which is typically nonzero (unless $X_i \in \{0, 1\}$ for all nonempty precincts). Therefore, the bounds we derived earlier for w_1 will be very useful here for limiting the scope of the influence by w_1.

For any possible value of the partially identified w_1, the district-level parameter $B = \sum_i N_i X_i \beta_i^b / \sum_i N_i X_i$ is estimated by the point estimator $B(w_1, \theta)$ following (16). We will now use the bounds on w_1 to bound this district-level parameter estimate $B(w_1, \theta)$, and estimate its θ parameter by regression.

4.3 Bounding District Cell Value

Due to Proposition 2 or Proposition 3, we know that $w_1 \in [wl, wu]$, where $wu = wu(\theta)$ and $wl = wl(\theta)$ depend on θ. Then

$$B(w_1, \theta) \in [Bl, Bu] \equiv [B(wu(\theta), \theta), B(wl(\theta), \theta)]. \quad (19)$$

The parameters $\theta = (w_0, c_1, d_1)^T$ can be estimated from a least squares regression

$$\hat{\theta} = (\hat{w}_0, \hat{c}_1, \hat{d}_1)^T \leftarrow \min_{w_0, c_1, d_1} \frac{\sum_{i=1}^{p} \rho_i [T_i - (w_0 + c_1 X_i + d_1 X_i^2)]^2}{\sum_{i=1}^{p} \rho_i}, \quad (20)$$

possibly weighted by some choice ρ_i.

Replacing θ in (19) by $\hat{\theta}$, we obtain the estimated bounds for the district parameter B. Since this is implied from a regression model of linear contextual effects, one may call this a *"regression bound,"* which will be our proposed interval estimate for B.

DEFINITION 1 (Regression bound).[2] *A regression bound for the district parameter* $B = \sum_i N_i X_i \beta_i^b / \sum_i N_i X_i$ *is of the form*

$$[\hat{B}l, \hat{B}u] \equiv [B(wu(\hat{\theta}), \hat{\theta}), B(wl(\hat{\theta}), \hat{\theta})], \tag{21}$$

where the functional form of the point estimate $B(w_1, \theta)$ *follows (16),* $wu = wu(\theta)$ *and* $wl = wl(\theta)$ *are the bounds of the* w_1 *parameter according to Proposition 2 or Proposition 3, and* $\hat{\theta}$ *estimates the regression coefficients* θ *from (20).*

5 Confidence Intervals

The previous regression bound $[\hat{B}l, \hat{B}u]$ for B does not take into account sampling variation. It assumes, for example, that the quadratic regression coefficients $\hat{w}_0, \hat{c}_1, \hat{d}_1$ are the true coefficients, while in reality they are estimated from p precincts and are subject to sampling error. Due to sampling error, it may be possible that according to the sample estimates, $B \notin [\hat{B}l, \hat{B}u]$, even if the model assumptions for linear contextual effects are valid, when we should automatically have $B \in [\hat{B}l, \hat{B}u]$ in the large p limit. (See Supplementary Appendix B.) To solve this problem, we will provide asymptotic conservative confidence intervals for B in this section, where $\hat{B}l$ will be reduced (and $\hat{B}u$ will be increased) by a typical size of the sampling variation.

Since $[\hat{B}l, \hat{B}u] \equiv [B(wu(\hat{\theta}), \hat{\theta}), B(wl(\hat{\theta}), \hat{\theta})]$ depends on the functional forms of $wl(\cdot)$ and $wu(\cdot)$, we first need to analyze in detail these functional forms.

In Propositions 2 and 3, the bounds wl and wu are functions of the quadratic regression coefficients $\theta = (w_0, c_1, d_1)^T$. The lower bounds can be expressed in the form

$$wl(\theta) = \max_{j=1}^{J}\{gl_j^0 + gl_j^T \theta\}, \tag{22}$$

and the upper bounds can be expressed in the form

$$wu(\theta) = \min_{j=1}^{J}\{gu_j^0 + gu_j^T \theta\}. \tag{23}$$

For Proposition 2, $J = 4$,
$gl_1^0 = 0, gl_1^T = (-1/l, 0, 0), gl_2^0 = -1/(1-l), gl_2^T = (1/(1-l), 1/(1-l), 1/(1-l) - 1),$
$gl_3^0 = 0, gl_3^T = (-1/u, 0, 0), gl_4^0 = -1/(1-u), gl_4^T = (1/(1-u), 1/(1-u), 1/(1-u) - 1),$
$gu_1^0 = 1/l, gu_1^T = (-1/l, 0, 0), gu_2^0 = 0, gu_2^T = gl_2^T,$
$gu_3^0 = 1/u, gu_3^T = (-1/u, 0, 0), gu_4^0 = 0, gu_4^T = gl_4^T.$
For Proposition 3, $J = 2$,
$gl_1^0 = 0, gl_1^T = (-1, 0, 0), gl_2^0 = -1, gl_2^T = (1, 1, 0),$
$gu_1^0 = 1, gu_1^T = (-1, 0, 0), gu_2^0 = 0, gu_2^T = (1, 1, 0).$
Using this notation, we have the following result.

PROPOSITION 4. *Let* $B = \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i X_i}$ *be the district parameter of voting proportion for a candidate of interest among all the black people in a district with p precincts. Let DD be the Duncan and Davis (1953) bound for B, following*

2 This bound corresponds to a special case with the choice of $\lambda = 0$ in a technical report by Jiang *et al.* (2018), who allow the residuals of the T_i X_i regression to be incorporated in the bound.

$$\mathrm{DD} = \left[\frac{\sum_{i=1}^{p} N_i \max\{0, T_i - (1 - X_i)\}}{\sum_{i=1}^{p} N_i X_i}, \frac{\sum_{i=1}^{p} N_i \min\{T_i, X_i\}}{\sum_{i=1}^{p} N_i X_i} \right]. \tag{24}$$

As $p \to \infty$, an asymptotic conservative confidence interval for B of the form

$$\mathrm{CI}_x \equiv [\hat{B}L - xSL, \hat{B}U + xSU] \cap \mathrm{DD} \tag{25}$$

has asymptotic coverage probability at least $\Phi(x)$.

Here, we use the following system of notation:

$x > 0$,

$\hat{\theta}^T = (\hat{w}_0, \hat{c}_1, \hat{d}_1)$ *which is estimated by quadratic regression (20), which has robust sandwich asymptotic variance matrix V,*[3]

$r \equiv \frac{\sum_i N_i X_i (1 - X_i)}{\sum_i N_i X_i}$,

$h_0 \equiv 0$,

$h^T \equiv \frac{\sum_i N_i X_i (1, 1, X_i)}{\sum_i N_i X_i}$,

$S_1 = (1/2)\sqrt{\sum_i \left(\frac{N_i X_i}{\sum_i N_i X_i}\right)^2}$,

$\hat{B}L = \max_{j=1}^{J}\{\hat{B}L_j\}$,

$\hat{B}U = \min_{j=1}^{J}\{\hat{B}U_j\}$.

For $j = 1, \ldots, J$, the gl_j's and gu_j's are defined above after (22) and (23),

$\hat{B}L_j = h_0 - rgu_j^0 + (h - rgu_j)^T \hat{\theta}$,

$\hat{B}U_j = h_0 - rgl_j^0 + (h - rgl_j)^T \hat{\theta}$,

$SL_j \equiv S_1 + \sqrt{(h - rgu_j)^T V (h - rgu_j)}$,

$SU_j \equiv S_1 + \sqrt{(h - rgl_j)^T V (h - rgl_j)}$,

$SL = SL_{\hat{j}}$ *where* $\hat{j} \equiv \arg\max_{j=1}^{J}\{\hat{B}L_j\}$,

$SU = SU_{\tilde{j}}$ *where* $\tilde{j} \equiv \arg\min_{j=1}^{J}\{\hat{B}U_j\}$.

For this result to hold, we assume that the linear contextual model holds conditional on both N_i and X_i on the entire support of these random variables and also for all X_i in a range specified in either Proposition 2 or Proposition 3. We assume that the robust variance V is of order $O_p(1/p)$. In addition, we assume the following "tie-breaking" conditions:

(i) *Assume that $N_i X_i (1 - X_i)$ is not almost surely 0.*
(ii) *Assume that the minimizing entry of $wu = \min_{j=1}^{J}\{gu_j^0 + gu_j^T \theta\}$ is unique and not tied with the other entries, and similarly the maximizing entry of $wl = \max_{j=1}^{J}\{gl_j^0 + gl_j^T \theta\}$ is unique and not tied with the other entries.*
(iii) *Assume that $wu(\theta) \neq wl(\theta)$.*

A derivation of this confidence interval CI_x in Proposition 4 is included in Supplementary Appendix A.

Remark 1. The tie-breaking conditions can be checked by examining the data at hand. The condition on N_i and X_i is satisfied if N_i is not almost surely 0 and if X_i does not almost surely take a boundary value (0 or 1) for nonempty precincts with $N_i > 0$. The conditions on θ will hold for almost all true parameters (except on a set with Lebesgue measure 0, where some of the $2J$ points $\{gu_j^0 + gu_j^T \theta, gl_j^0 + gl_j^T \theta, j = 1, \ldots, J\}$ are exactly tied). In the Bayesian sense when θ is regarded as a vector of continuous random variables, these conditions hold with probability one since any ties would force θ to lie on a lower dimensional manifold which has zero Lebesgue measure.

3 See, for example, https://www.stata.com/manuals/p_robust.pdf

REMARK 2. Instead of the analytic method described here, one may consider using the bootstrap to estimate the standard deviation (sd) of the bound estimate $\hat{B}L$ (and similarly for $\hat{B}U$) and replace the SL in the formula of CI_x by $\mathrm{sd}_{\mathrm{boot}}(\hat{B}L)$. However, we suspect that this bootstrap method would not be theoretically valid here. The reason is that we are not interested in how much $\hat{B}L$ varies from its own nonstochastic large sample limit, that is, the typical size of $\hat{B}L - \lim_{p\to\infty}\hat{B}L$. We are really interested in the typical size of $\hat{B}L - B$ instead. However, the district-level parameter $B = \frac{\sum_{i=1}^{p} N_i X_i \beta_i^b}{\sum_{i=1}^{p} N_i X_i}$ is a nonidentified stochastic quantity, and its sampling variations would be ignored by bootstrapping $\hat{B}L$ alone. Nevertheless, in practice, the bootstrap method may still work well heuristically for describing the sampling variation.

6 Illustrations and Applications

We now give theoretical and simulation analyses (Section 6.1) and empirical applications (Section 6.2) of our new bounds.

6.1 Theoretical

We will compare the proposed bound CI_x to the Duncan and Davis (1953) bound DD, as defined in Proposition 4. For any interval A, we will use $|A|$ to denote its length. We will use $x = 0$ and $x = 1$ for illustration.

To measure the success of the proposed method, we examine:

(1) Whether the new interval estimate contains the true district parameter: $B \in \mathrm{CI}_x$.
(2) How narrow the new interval estimate is compared to the DD bound: the width ratio $\mathrm{WR}_x \equiv |\mathrm{CI}_x|/|\mathrm{DD}|$.

In the examples below, we assume $X \sim \mathrm{Unif}[0,1]$ and N_i is constant for all i, unless otherwise stated.

6.1.1 Continuation of Example in Section 2.3

We first return to the simple example of Section 2.3. We observe $T_i = 0.9 - 0.2X_i$. The regression parameters are $(w_0, c_1, d_1) = (0.9, -0.2, 0)$. Here one can apply Proposition 3 to obtain $[wl, wu] = [\max\{-0.9, -0.2+0.9-1\}, \min\{1-0.9, -0.2+0.9\}] = [-0.3, 0.1]$. In this case, in the limit of a large number of precincts (large p), the proposed interval in Section 4 becomes $[BL, BU] = (0.9-0.2) - [E(X(1-X))/EX][0.1, -0.3]$, where $E(X(1-X))/EX = 1/3$ for uniform X. Therefore, $[BL, BU] \approx [0.67, 0.80] \approx 0.73 \pm 0.07$. What about the true district B? In the large p limit, $B = EX\beta^b/EX$, but we pointed out that β^b is unidentified. For example, it could be either $0.9-0.2X$ or $0.8-0.1X$ as shown in Figure 1. In the first case, $B = E(X(0.9-0.2X))/EX = 0.9-0.2(2/3) = 0.77$, and in the second case, $B = E(X(0.8-0.1X))/EX = 0.8-0.1(2/3) = 0.73$. In either case, the true B still falls in the proposed interval $[0.67, 0.80] \approx 0.73 \pm 0.07$. This interval may still seem not particularly tight, but this is necessary due to the intrinsic indeterminacy. For example, in another scenario, constant $(\beta_i^b, \beta_i^w) = (0.7, 0.9)$ can also explain the observed T, X relation, as mentioned in the discussion of Freedman *et al.* (1991) in King (1997, Chapter 3.2). This would lead to $B = 0.70$ being still included and quite close to the lower end of the proposed interval $[0.67, 0.80]$.

The large sample limit of the DD bound is $[E\max\{0, T-1+X\}/EX, E\min\{T, X\}/EX] \approx [0.61, 0.93] \approx 0.77 \pm 0.16$. So the proposed bound $[0.67, 0.80] \approx 0.73 \pm 0.07$ is contained inside the DD bound and the width ratio (in the large p limit) is about $0.07/0.16 \approx 0.44$. So the proposed bound actually becomes less than half as wide in this case, compared to DD.

Different relations between T and X could lead to different width ratios of the proposed method in comparison to the DD bound. We provide several additional examples below.

6.1.2 *Additional Examples*

EXAMPLE 1. $\beta_i^b = T + \tau(1 - X_i) \in [0,1]$ and $\beta_i^w = T - \tau X_i \in [0,1]$, where probability constraints entail $\tau \in \pm \min(T, 1-T)$ and $T \in (0,1)$. Then the plot T_i against X_i is a flat $T_i = T$. Here, one can show by Proposition 3 that $[wl, wu] = \pm \min(T, 1-T)$. In this case, in the limit of large precincts and large number of precincts (large N_i and p), it can be shown analytically that the true parameter $B \approx E\beta_i^b = T + \tau/3 \in \mathrm{CI}_0 \approx T \pm (1/3)\min(T, 1-T) \subset \mathrm{DD} \approx [T^2, 2T - T^2]$. Also, $\mathrm{WR}_0 \equiv |\mathrm{CI}_0|/|\mathrm{DD}| \approx 1/[3\max(T, 1-T)] \in (1/3, 2/3)$. In summary, the proposed bound tightens the DD bound while still containing the true parameter.

EXAMPLE 2. $\beta_i^b = \tau(1 - X_i)$, $\beta_i^w = 1 - \tau X_i$, where $\tau \in [0,1]$. Then the plot T_i against X_i is $T_i = 1 - X_i$. Here, one can show by Proposition 3 that $[wl, wu] = [-1, 0]$. In this case, in the limit of large precincts and large number of precincts (large N_i and p), it can be shown analytically that $B \approx E\beta_i^b = \tau/3 \in \mathrm{CI}_0 \approx [0, 1/3]$, $\mathrm{DD} \approx [0, 1/2]$. Also, $\mathrm{WR}_0 \equiv |\mathrm{CI}_0|/|\mathrm{DD}| \approx 2/3$. In summary, the proposed bound tightens the DD bound while still containing the true parameter.

EXAMPLE 3. $\beta_i^b = 0, \beta_i^w = 1 - X_i$. Then the plot T_i against X_i is $T_i = (1 - X_i)^2$. Here, one can show by Proposition 3 that $[wl, wu] = [-1, -1]$; so w_1 is identified. In this case, in the limit of large precincts and a large number of precincts (large N_i and p), it can be shown analytically that $B = 0$, $\mathrm{CI}_0 \approx [0, 0]$, $\mathrm{DD} \approx [0, 2E\min(X, (1 - X)^2)] \approx [0, 0.3032767]$. Also, $\mathrm{WR}_0 \equiv |\mathrm{CI}_0|/|\mathrm{DD}| \approx 0$.

We now generate $p = 1000$ precincts all with population $N_i = 150$ for this example. For sample estimates based on this finite dataset, we obtain true $B = 0$, $\mathrm{DD} = [0, 0.301843]$.

We apply Proposition 2 for this example with $[l, u] = [\min(X_i), \max(X_i)] = [0.001473298, 0.9988792]$.

We obtain $\hat{B}l = 0.000023$ and $\hat{B}u = 0.000305$ which are very close to $B = 0$, but $\mathrm{CI}_0 = [\hat{B}l, \hat{B}u]$ excludes the true B due to sampling variation. On the other hand, the proposed interval estimate narrowly misses the true B due to sampling variation. The confidence interval CI_x for $x = 1$ is $([-0.018185, 0.018556] \cap \mathrm{DD}) = [0, 0.018556]$, which does contain the true B now and is still very narrow. (Here, intersection with the DD bound improves the lower bound to be 0.) In summary, the regression bound CI_0 can miss the true parameter due to sampling variation. However, after expanding the bound to account for the sampling variation, CI_1 does contain the true parameter B and is still much narrower than the DD bound.

EXAMPLE 4. Consider $p = 1000$ precincts all with population $N_i = 150$. We let $X_i \sim \mathrm{Unif}[0, 0.95]$, $\beta_i^b \approx (N_i X_i)^{-1}\mathrm{Bin}(N_i X_i, 1/(1+\exp(-b0 - b1 \times X_i - (1 - X_i) \times \epsilon_i^b)), \beta_i^w \approx (N_i(1 - X_i))^{-1}\mathrm{Bin}(N_i(1 - X_i), 1/(1 + \exp(-w0 - w1 \times X_i - (1 - X_i)\epsilon_i^w))$ (the approximation $\approx$ here involves operations such as rounding $N_i X_i$ and adding 1 to avoid zero or fractional counts), where $\epsilon_i^{b,w}$'s are *iid* $N(0, s^2)$, $s = 0.5$, $b0 = 2.197225$, $b1 = -1.791759$, $w0 = 2.197225$, $w1 = 0$, $T_i \approx \beta_i^b X_i + \beta_i^w (1 - X_i)$ (the approximation $\approx$ here involves operations such as replacing X_i by a rounded version of $N_i X_i$ divided by N_i). The resulting T_i versus X_i scatterplot is given by Figure 3.

We apply Proposition 2 for this example with $[l, u] = [\min(X_i), \max(X_i)] = [0.001400, 0.948935]$.

In this case, it can be shown that $B = 0.733582 \in \mathrm{CI}_0 = [0.704450, 0.750966] \subset \mathrm{DD} = [0.636268, 0.931647]$. Also, $\mathrm{WR}_0 \equiv |\mathrm{CI}_0|/|\mathrm{DD}| = 0.157478$.

The CI_1 is $[0.682944, 0.772162]$, which is also narrower than the DD interval and contains the true B.

EXAMPLE 5. Consider $p = 1000$ precincts all with population $N_i = 150$. We let $X_i \sim \mathrm{Unif}[0, 0.7]$, $\beta_i^b \approx (N_i X_i)^{-1}\mathrm{Bin}(N_i X_i, 1/(1+\exp(-b0 - b1 \times X_i - (1 - X_i) \times \epsilon_i^b)), \beta_i^w \approx (N_i(1 - X_i))^{-1}\mathrm{Bin}(N_i(1 - X_i), 1/(1 + \exp(-w0 - w1 \times X_i - (1 - X_i)\epsilon_i^w))$ (the approximation $\approx$ here involves operations such as rounding $N_i X_i$ and adding 1 to avoid zero or fractional counts), where $\epsilon_i^{b,w}$'s are *iid* $N(0, s^2)$, $s = 1$, $b0 = 0$, $b1 = 0$, $w0 = 2.197225$, $w1 = 0$, $T_i \approx \beta_i^b X_i + \beta_i^w (1 - X_i)$ (the approximation $\approx$

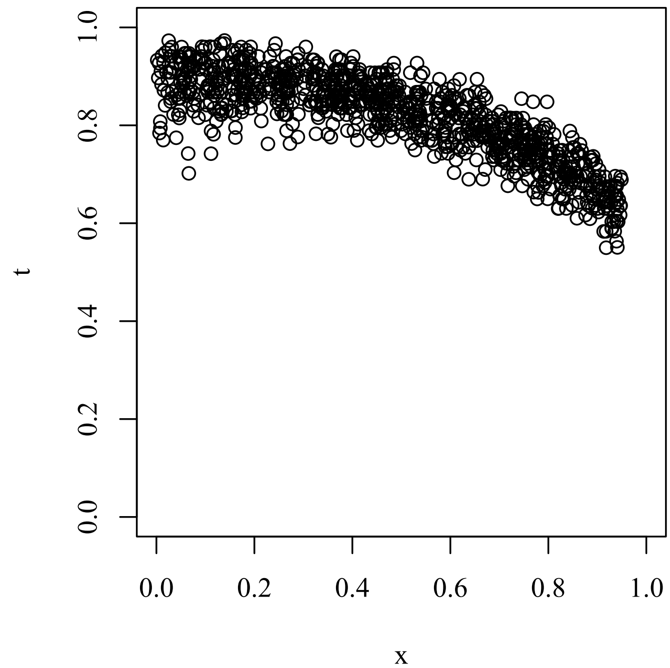

Figure 3. T versus X scatterplot for Example 4.

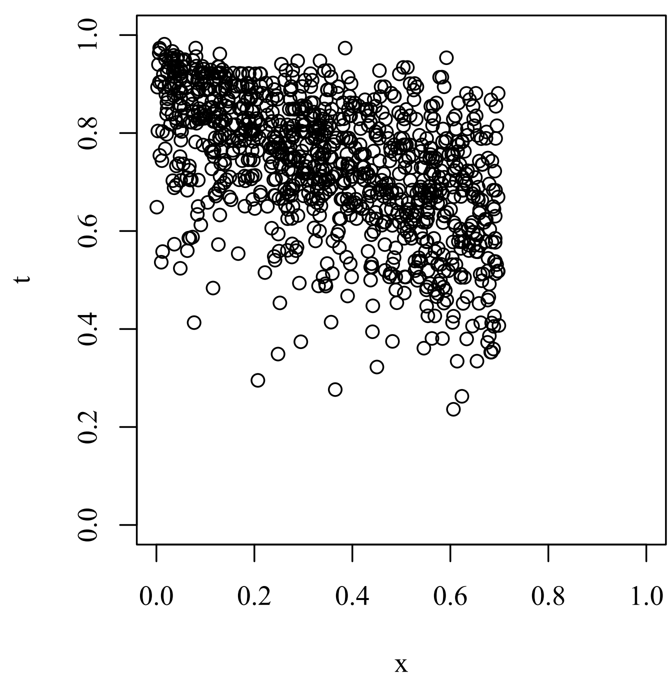

Figure 4. T versus X scatterplot for Example 5.

here involves operations such as replacing X_i by a rounded version of $N_i X_i$ divided by N_i). The resulting T_i versus X_i scatterplot is given by Figure 4.

We apply Proposition 2 for this example with $[l, u] = [\min(X_i), \max(X_i)] = [0.001031, 0.699215]$.

In this case, it can be shown that $B = 0.499342 \in \mathrm{CI}_0 = [0.399895, 0.759834] \subset \mathrm{DD} = [0.340341, 0.961388]$. Also, $\mathrm{WR}_0 \equiv |\mathrm{CI}_0|/|\mathrm{DD}| = 0.579568$.

The CI_1 is [0.368232, 0.799369], which is also narrower than the DD interval and contains the true B.

The CI_1 used in Examples 3–5 will have at least $\Phi(1) \approx 84\%$ coverage probability asymptotically, according to Proposition 4. In repetitions of 1000 simulations, we found that CI_1 is very conservative: $P[B \in CI_1] = 944/1000, 1000/1000, 1000/1000$, respectively, in Examples 3, 4, and 5. The comparison for (mean width of CI_1, mean width of DD bound) is (0.0178, 0.3032), (0.0946, 0.2974), (0.4405, 0.6227), respectively. These results demonstrate that the proposed confidence intervals are considerably more informative about B compared to the DD bounds, as shown in repeated simulations.

It is noted that in Examples 4 and 5, the true models do not follow the linear contextual model or quadratic regression of T_i versus X_i. The $\beta_i^{b,w}$'s follow overdispersed logistic regression model with heteroscedastic normal random effects.

6.2 Empirical

Given that some information is forever lost during the process of aggregating individual-level data, it is important to develop models tuned to the specific types of datasets similar to those used in practice. Unfortunately, for the very reason that EI is a problem in the first place, datasets with true labels in target application areas, such as elections and voting rights litigation, are typically not available. The nature of the learning problem is thus intrinsically much more difficult than a traditional supervised learning problem where labeled examples sampled from the target distribution are abundant. As such, the most recent work on EI has evaluated approaches using a very small number of datasets with ground truth, combined with artificial, simulated data. Here, we dramatically increase the number of datasets with ground truth labels on social data for evaluation of our proposed model as well as to serve as a test bed for future approaches to EI model building. We describe the data we collected followed by our empirical results.

6.2.1 Data

Datasets from previous works (e.g., King 1997; Wakefield 2004; Imai, Lu, and Strauss 2008) include data on voter registration and race in 1968; literacy by race in 1910; and party registration in south-east North Carolina in 2001. We use these data and also collect datasets from the US Centers for Disease Control and Prevention on mortality rates by gender and race (CDC 2017); literacy rates and educational attendance by gender from the 2001 Census of India (Office of the Registrar General & Census Commissioner 2001); and additional datasets from the US Census and American Community Surveys from 1850 to 2016 via the Integrated Public Use Microdata Series (Ruggles *et al.* 2017). From these sources, we created 8,430 datasets (i.e., X, T pairs). Some of these datasets are dependent across time and levels of geographic granularity. For example, for the US Census and American Community Surveys, we have 4 unique X variables and 75 unique T variables analyzed across available years and geographic units (Minor Civil Divisions or counties). In some cases, additional datasets are created by dichotomizing individual-level multicategory variables in different ways. For example, we create binary variables from the number of family members in a household by dichotomizing as one and greater than one family members, and then in a separate dataset as up to two and more than two family members, and so forth.

The datasets contain a total of 44,164,540 geographic units (precincts, counties, etc.), with an average of about 5,239 geographic units per dataset and a median of 478, ranging from 145 to 41,783. Our replication data are publicly available via Harvard Dataverse (Jiang *et al.* 2019). We discuss limitations of evaluating EI methods with these data in Section 7.1.

Table 1. Effectiveness in terms of the nominal coverage probability, $\Phi(x)$; proportion of intervals that capture the true district value, $p(B \in \mathrm{CI}_x)$; and the width ratio among those selected. The last two columns repeat the previous two among only the 88% of the datasets that do not revert to DD bounds.

x	$\Phi(x)$	$p(B \in \mathbf{CI}_x)$	$E[\mathbf{WR}_x]$	$p(B \in \mathbf{CI}_x \mid \mathbf{selected})$	$E[\mathbf{WR}_x \mid \mathbf{selected}]$
0.00	0.5000	0.9410	0.4474	0.9330	0.3721
0.25	0.5987	0.9833	0.5534	0.9810	0.4926
0.50	0.6915	0.9891	0.6145	0.9876	0.5619
0.75	0.7734	0.9928	0.6610	0.9918	0.6148
1.00	0.8413	0.9951	0.6995	0.9945	0.6586
1.25	0.8944	0.9967	0.7320	0.9962	0.6955
1.50	0.9332	0.9985	0.7605	0.9982	0.7279
1.75	0.9599	0.9991	0.7859	0.9989	0.7568
2.00	0.9772	0.9995	0.8085	0.9995	0.7824

6.2.2 *Analysis*

Our goal is scientifically appropriate EI (including, when appropriate, conclusions such as "we don't know anything") even in the presence of (a) assumptions that are violated and (b) data where most or all of the relevant individual-level information has been aggregated away. The specific method we evaluate here has no adjustable parameters and works on all input data. It begins with the easy-to-apply bounds of Proposition 2 based on a quadratic regression (leaving the nonparametric regression approach in Proposition 1 to future work because it involves tuning the smoothing parameter and is harder to derive the confidence intervals). The method then uses CI_x if $\hat{B}l \le \hat{B}u$ and DD covers part of CI_0, and otherwise reverts to the DD bounds. (This simple heuristic eliminates cases when the bounds flip, which can occur in practice when assumptions are violated; see Supplementary Appendix B, Remark 2.)

Table 1 displays the effectiveness of our methodology for all datasets in our collection, given differing confidence levels, $\Phi(x)$ (in the second column). We observe (in the third column) that our proposed bounds consistently capture the true value more often than the nominal coverage intervals, meaning that our bounds are highly accurate but also conservative. For example, at the 96% confidence interval (second to last row), our bounds capture the truth 99.91% of the time rather than 96%. The improvement relative to DD appears as the ratio of the length of our new confidence interval to the length of the original DD bound width, as reported in the fourth column. This number is always less than 1.0, often substantially so.

By inferring from these data, we recommend that, in practice, researchers use our bounds while setting $x = 0.5$ (see the third row of numbers for $\mathrm{CI}_{0.5}$), which is a reasonable trade-off between the capture probability and the width ratio for the observed datasets. It captures the truth in 98.9% of our 8,430 datasets and narrows the bounds by 38.5% relative to the 65-year-old DD bounds. We also analyze 92 of the 8,430 datasets where the bounds do not capture the truth, by constructing Figure 5. The figure gives a histogram of the size of the misses, the vast majority of which are very small, almost all less than 0.05.

For completeness, we also repeat the calculation for columns 3 and 4 among only those datasets where our method does not revert to DD bounds. These results appear in the final two columns. Because our method reverts to the DD bounds in only 12% of our datasets, narrowing the bounds in the remaining 88%, the last two columns do not differ much from columns 3 and 4.

Finally, we summarize these results in Figure 6 by plotting the width of our proposed bounds (horizontally) by the width of the DD bounds (vertically). Each dot is one of our 8,430 datasets. The green dots on the diagonal represent the 12% of datasets for which our method automatically returned the DD bound. For all others, the bounds are narrower and thus more informative, which is reflected in the figure by being above the diagonal line. Among these, the few red dots are those

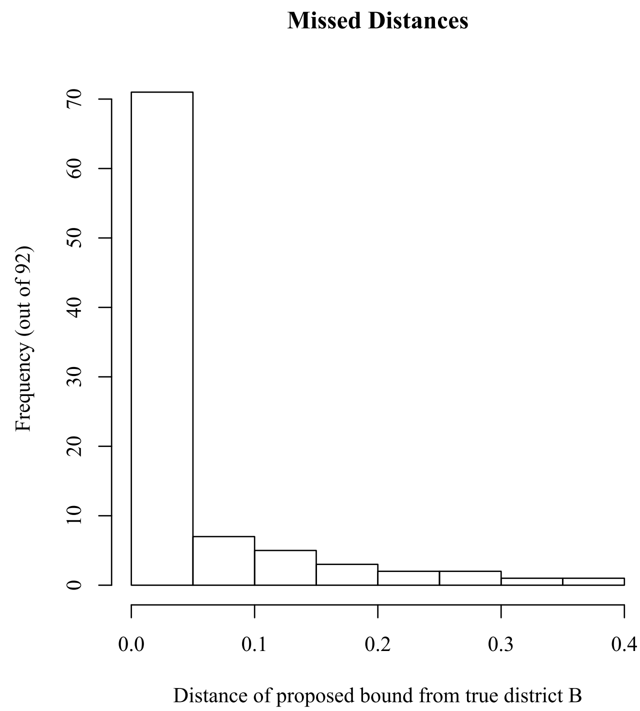

Figure 5. Histogram for how far away the true B falls outside of $\mathrm{CI}_{0.5}$ for the 92 datasets (out of the total 8,430) for which the bounds were applied, but the true district-level B value was not captured.

where the truth is not captured and the mass of black dots is where the truth is captured. The figure reveals that our approach is able to improve the most over the DD bounds when the DD bounds have widths farther from the 0 or 1 extremes of the unit interval.

Overall, the results of analyzing more than 8,000 datasets with known truth suggest that our approach generates considerably more information than the bounds proposed by, and used routinely in the literature since, Duncan and Davis (1953) with very little cost.

7 Concluding Remarks

We now discuss limitations (Section 7.1), comparisons with fully parametric identified models (Section 7.2), and suggestions for future research (Section 7.3).

7.1 Limitations

Our work adds a single essential assumption, requiring contextual effects, if any, to be linear (Assumption 1). This is far more general than the traditional approaches, which assume zero contextual effects (or effects that have zero correlation with X) and which are regularly falsified by real data with knowledge of the ground truth. Much of the problem with Goodman's regression giving answers outside of the known DD bounds is precisely because of this implicit zero contextual effect assumption that we generalize and thus avoid.

Yet, when Assumption 1 fails, the bounds produced by our method may not capture the truth. The key questions in practice are how often such problems occur and how can one know about such violations. Fortunately, we have found in Section 6.2 that violations bad enough to violate our bounds are rare in our collection of datasets. However, in the difficult field of EI, we must constantly be aware that it is always theoretically possible to violate assumptions without any signal in observable data. Consider the following example:

Example 6. Suppose $X_i \sim \mathrm{Unif}[0,1]$ and N_i is independent of X_i and $\beta_i^{b,w}$. Suppose we have quadratic contextual effects $\beta_i^b = T + b_2(X_i^2 - 1)$, $\beta^w = T + b_2(X_i^2 + X_i)$, where to ensure

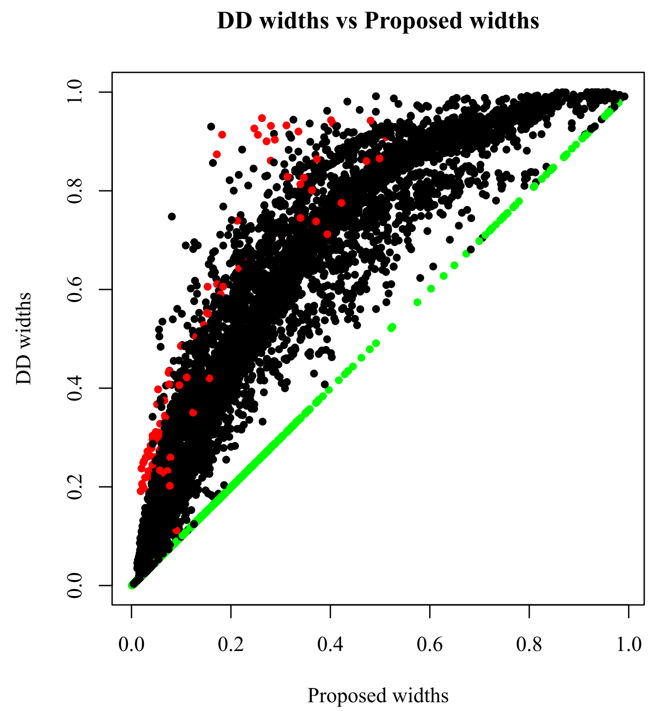

Figure 6. Widths of the proposed bounds relative to the DD bounds. Green points represent datasets in which the bounds reverted to those of the DD bounds, and red points indicate the 92 datasets (out of the total 8,430) for which the bounds were applied, but the true district-level B value was not captured.

these are probabilities valued in $[0, 1]$ for all possible X, we restrict $T \in (0, 1)$ and $b_2 \in [\max\{-T/2, -(1-T)\}, \min\{T, (1-T)/2\}]$. Then $T_i = \beta_i^b X_i + \beta_i^w (1 - X_i) = T$. The observed data (X_i, T_i) would be the same as our Example 1 earlier ($T_i = T$). We have already found that the large sample limit of our proposed bound is $\mathrm{CI}_0 = T \pm (1/3)\min\{T, 1-T\}$. The large sample limit of true B is now $E(N_i X_i \beta_i^b)/E(N_i X_i) = T - b_2/2$. It is then possible that for large enough b_2, $B \notin \mathrm{CI}_0$ (e.g., when $b_2 = T = 1/3$). The same holds in the large sample limit for CI_x with any $x > 0$ since the sampling variation that differentiates between CI_x and CI_0 disappears in the large sample limit.

If all datasets were generated from this model (e.g., with $b_2 = T = 1/3$), then the asymptotic coverage probability of any CI_x would be 0 and we would not be able to avoid such data sets without knowledge of the ground truth. Fortunately, this kind of "nondetectable violation" happens quite rarely, at least in our data. For example, the nondetectable violation in Example 6 is caused by the quadratic effects in β_i^b and β_i^w canceling each other exactly by chance. In addition, our interval estimates are robust in the sense that even a small amount of violation of the assumptions do not matter. For example, the quadratic effect b_2 does not have to be exactly 0 for CI_0 to capture B. This is in contrast to traditional point estimates and their confidence intervals, which will miss the true parameter due to any bias when the sample size p is sufficiently large, since the width of the confidence interval typically shrinks at the rate of $1/\sqrt{p}$.

From the thousands of real datasets on which we evaluated the approach, we found that most practically important violations can be easily detected if CI_0 is empty (i.e., the regression bound either flips or does not intersect with the DD bound at all). Supplementary Appendix B examines this analytically for the limit of large p (see Remark 2). The logic there is to prove that if the assumptions hold, then CI_0 should not be empty. Therefore, if CI_0 is found to be empty, then something must be wrong about the assumptions.

As shown in Section 6.2, in most cases in our real data, we have nonempty CI_0. When applying the CI_x for $x > 0$ on the selected datasets with nonempty CI_0, we found that our conservative confidence interval CI_x tends to capture the true district parameter B more often than the stated level of confidence $\Phi(x)$, while tightening the DD bound. For example, $\mathrm{CI}_{0.5}$ has nominal coverage probability about 70%, but it actually captures B more than 90% of the selected datasets (see Table 1).

Although these datasets dramatically increase researchers' ability to evaluate methods of EI empirically, the data may not be representative of every dataset that researchers choose to analyze in the future. The performance measures may thus differ for different collections of data sets. In all likelihood, the width ratio may be less sensitive to new data than widths themselves, and the capture probability may be less sensitive than the size of the misses or the probability that the proposed method reverts to the DD bound. Although our data are not randomly selected from the set of all possible datasets (which would not necessarily be useful anyway), these datasets make it possible to move at least some decisions about which model is appropriate from a theoretical or normative choice to a more sound empirical basis. Researchers should evaluate the application of our methodology or any other to their own data based, in part, on how statistically similar their datasets are to those in our collection.

For example, our method appears to be less useful for the 189 datasets in our collection with X defined by gender. Although it is well-known that gender data are difficult to handle for any EI methods, it also poses challenges to our regression approach since it is well-known that regression coefficients cannot be determined very well if the range of X is small. This happens to be the case for gender (having $X_i \approx 0.5$ for almost all i) for the obvious reason that men and women tend to live together. The narrow range for X, which implies that most information has been aggregated away, also makes it easy for quadratic regression of T, X to be distorted by outliers or influential points since an outlier (say a precinct with a prison composed mostly of males with zero voters) would be far from the mass of other data points and an unreliable basis on which to make inferences about the rest of the data. Possibly for these reasons, our proposed bound for gender datasets tends to revert more often than for other data to the DD bound and, when not reverting, it tends to either fail to tighten the DD bound much or miss the true parameter more often. We have studied this problem but have not found a way to automatically identify problematic datasets, without excluding too many false positives for which our proposed method works better. Our technical report suggests a second heuristic, that is, to revert to the DD bound also if its width exceeds 0.7 (Jiang *et al.* 2018). The rationale is that this represents a dataset where there is intrinsic lack of information, and the proposed regression method should not be expected to work reliably. Adding this second heuristic indeed is successful in reducing the misses of the true parameters for the 189 datasets with gender variables: from about 8% (16/189) down to about 2% (3/189). However, the percentage of all 8,430 datasets where the proposed method does not revert to the wider DD bound also deteriorated (from about 88% to about 60%). We leave the study of EI in the context of low information data to future study.

More generally, our datasets may have T, X distributions different from others. Future researchers may wish to derive more general characterizations of what types of datasets are likely to have accurate intervals with narrow widths. For now, we can suggest one preliminary result about this important subject that works well with the data we have analyzed. For example, for intervals derived from Proposition 2, if the relation between T and X is determined by a quadratic regression fit $t(x) = w_0 + c_1 x + d_1 x^2$ that is linear (where $d_1 = 0$) and if $t(0)$, $t(1)$ are both in range $(0, 1)$, then the large p limit for the width of the proposed interval $D_1 = (1 - \chi)(1 - \tau)/(1 - \delta)$, where $\tau = |t(0.5) - 0.5|/0.5$, $\chi = ENX^2/ENX$, and δ uses information about a symmetric range $[l, u] = [\delta, 1 - \delta] \subset (0, 1)$ where we assume the contextual model in Proposition 2. This implies that in order to have narrow proposed intervals for large p, we hope to have $[l, u]$ close to $(0, 1)$,

$t(0.5) = E(T|X = 0.5)$ to be far away from 0.5, and make ENX^2/ENX large. The latter happens when nonzero values of X tend to be close to 1 for all precincts with $N > 0$.

For more than half of our datasets, $ET > 0.8$, and this may be favorable for generating narrow proposed intervals since ET is similar to the influential factor $t(0.5)$. For datasets for which ET tends to be closer to 0.5, the interval width could be too wide to be informative on B. However, even in such cases, our intervals still tend to be comparatively shorter than DD. It is also noted that when the interval is wide, there is a reason for it to be so, since each location of the wide interval could be the true value of B in a reasonable scenario. We argue it is better to expose this intrinsic indeterminacy, rather than producing shorter intervals that are sensitive to further assumptions and can miss the truth badly when such assumption fails. Moreover, as noted above, the width of the proposed interval depends on several factors. Even if ET is close to 0.5, it is still possible for the interval to be narrow if the X distribution is favorable. For example, if most precincts have a high proportion black X, while a smaller number of other precincts are predominantly white, then the width of the proposed interval can still be very narrow even if the T is located around 0.5.

7.2 Comparison with Identifiable Methods

An alternative approach is to assume nonlinear contextual effects such as $E(\beta_i^b|X_i) = 1/(1 + e^{-b_0-b_1X_i})$ and $E(\beta_i^w|X_i) = 1/(1+e^{-w_0-w_1X_i})$. At first sight, this seems to avoid the nonidentifiability problem in the model $E(T_i|X_i) = X_i/(1 + e^{-b_0-b_1X_i}) + (1 - X_i)/(1 + e^{-w_0-w_1X_i})$. However, the limitations of such an approach are well-known: "Unfortunately, assuming nonlinearity theoretically removes the nonidentifiability but in practice is totally dependent on the form chosen, and parameter estimates will in general be highly unstable" (Wakefield 2004, Section 1.3).

In contrast, our approach is to directly confront the nonidentifiability problem by modeling only the linear contextual effects. Our linearity assumption may be wrong, but a linear relationship between two bounded variables is normally a reasonably good first approximation. Not always, of course, but at least readers will always fully understand the assumption. This seems to be preferable to point estimation based on a fully parametric approach with model dependence and instability hidden in difficult-to-detect ways.

A similar comment can be made in comparison to any method that is made identifiable only in a way that is sensitive to some assumption. For example, in the "extended" model of King (1997), linear contextual effects (usually with diffuse priors) are placed on the untruncated means of the underlying truncated bivariate normal (TBVN) distribution of the precinct quantities of interest. This model is identifiable due to the truncated normal distributional assumption on the precinct quantities and has the advantage over our approach of providing sharp point estimates and precinct-level estimates. In contrast, our proposed approach makes no distributional assumption and, at the cost of only providing bounds and no precinct-level estimates, should be relatively robust. We are also able to offer explicit formulas that reveal the scope of the indeterminacy that remains regardless of whether the precinct quantities truly follow a truncated normal distribution. The proposed method is also computationally much faster than the extended TBVN model, which is based on a fully Bayesian model with approximation via Monte Carlo simulation.

In general, for any identifiable model sensitive to the modeling or prior distribution assumptions, the resulting credible interval or confidence interval will be narrower (by an order $1/\sqrt{p}$) than ours (of order 1). This means that if the assumptions that lead to identifiability of the full model are correct, it will capture the true parameter more precisely. However, this model may have poorer coverage properties and may not capture the truth when the assumptions are wrong.

We also note that our approach may provide some useful insight for improving models that are identifiable. For example, in our experiments with the TBVN model, our heuristic for selecting datasets also seems to help improve the success of the TBVN credible intervals too and seems to be a general indicator of information in individual behavior being destroyed during the aggregation process. Dilating the credible intervals by the estimate plus or minus an order-$\sqrt{p}$ multiplier of the standard error can greatly improve the coverage probability. This could help any confidence interval or credible interval that has width of order $1/\sqrt{p}$ to battle intrinsic indeterminacy since the large multiplier effectively dilates the interval to be order 1 and can become more robust against violation of the assumptions. Another approach that fits under our conceptual framework would be to add a constant offset such as ± 0.1 to all the district estimates of Goodman's regression that fall inside the DD bounds, and to use their intersections, which often works well too.

7.3 Suggestions for Future Research

When data have influential observations or seem to belong to several different clusters, we found that a divide and conquer strategy may be helpful. One could divide the data into several parts and apply either the proposed bound or the DD bound to each part, depending on the observed pattern in the particular part. The proposed bound could be applied to any part of the data that displays a common pattern (e.g., those of linear or quadratic regression). For parts of the data that are outliers or that otherwise lack a clear pattern for linear or quadratic regression, one could apply the DD bound. The bounds would then be combined by weighting the number of relevant people in each part of the data to obtain a single bound. In initial experiments of such an approach, we segmented the data visually and found that this strategy can sometimes rectify the misses or nonselection of the current method. We leave automating the process of segmentation to future work.

We have thus far only considered one variable X_i for the contextual effect. One may also consider adding other covariates to the contextual effect models, modeling both β_i^b and β_i^w. We also only focus on inference for the district-level parameter; it would be important to obtain useful bounds for the precinct-level parameters β_i^b and β_i^w (that would parallel the precinct-level estimates in King (1997)), probably by modeling the distribution of the residuals $(\beta_i^b - E(\beta_i^b|X_i)), (\beta_i^w, E(\beta_i^w|X_i))$, or at least the second moments such as $var((\beta_i^b, \beta_i^w)^T|X_i)$. (The residuals average out in the district-level estimates, so we could still get useful bounds for the district-level parameter in the current paper, even without modeling the residuals.) Finally, it would be useful to extend the ideas in this paper to the case of more general $R \times C$ tables (perhaps generalizing Cho and Manski 2008).

Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2019.19.

References

Achen, C. H., and W. P. Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Altman, M., J. Gill, and M. McDonald. 2004. *A Comparison of the Numerical Properties of EI Methods*, edited by G. King, O. Rosen, and M. A. Tanner, 383–409. New York: Cambridge University Press.

Centers for Disease Control and Prevention (CDC), National Center for Health Statistics. 2017. Underlying Cause of Death 1999–2016 on CDC WONDER Online Database. Data are from the Multiple Cause of Death Files, 1999–2016, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/ucd-icd10.html (retrieved in 2017).

Chambers, R. L., and D. G. Steel. 2001. "Simple Methods for Ecological Inference in 2 × 2 Tables." *Journal of Royal Statistical Society Series A* 164(Part 1):175–192.

Chernozhukov, V., H. Hong, and E. Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica* 75:1243–1284.

Cho, W. K. T., and C. F. Manski. 2008. "Cross Level/Ecological Inference." In *Oxford Handbook of Political Methodology*, edited by J. Box-Steffensmeier, H. Brady, and D. Collier, 530–569. Oxford, UK: Oxford University Press.

Duncan, O. D., and B. Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18:665–666.

Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smyth, and C. G. Everett. 1991. "Ecological Regression and Voting Rights." *Evaluation Review* 15:659–817 (with discussion).

Goodman, L. 1953. "Ecological Regression and the Behavior of Individuals." *American Sociological Review* 18:663–664.

Imai, K., Y. Lu, and A. Strauss. 2008. "Bayesian and Likelihood Inference for 2×2 Ecological Tables: An Incomplete Data Approach." *Political Analysis* 16(1):41–69.

Jiang, W., G. King, A. Schmaltz, and M. A. Tanner. 2019. "Replication Data for: Ecological Regression with Partial Identification," https://doi.org/10.7910/DVN/W7KZVL, Harvard Dataverse, V1.

Jiang, W., G. King, A. Schmaltz, and M. A. Tanner. 2018. Ecological Regression with Partial Identification. Technical Report. https://arxiv.org/abs/1804.05803. Replication Data: https://doi.org/10.7910/DVN/8TB7GO, Harvard Dataverse, V1.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

King, G., O. Rosen, and M. A. Tanner. 2004. *Ecological Inference: New Methodological Strategies*. New York: Cambridge University Press.

Liao, Y., and W. Jiang. 2010. "Bayesian Analysis in Moment Inequality Models." *The Annals of Statistics* 38:275–316.

Office of the Registrar General & Census Commissioner, India. 2001. Census of India 2001. Accessed at https://data.gov.in (retrieved in 2017).

Owen, G., and B. Grofman. 1997. "Estimating the Likelihood of Fallacious Ecological Inference: Linear Ecological Regression in the Presence Context Effects." *Political Geography* 16:675–690.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek. 2017. *Integrated Public Use Microdata Series: Version 7.0 [dataset]*. Minneapolis, MN: University of Minnesota, 2017. https://doi.org/10.18128/D010.V7.0. Accessed at https://usa.ipums.org/usa/ (retrieved in 2018).

Wakefield, J. 2004. "Prior and Likelihood Choices in the Analysis of Ecological Data." In *Ecological Inference: New Methodological Strategies*, edited by G. King, O. Rosen, and M. A. Tanner, 13–50. New York: Cambridge University Press.