

NBER WORKING PAPER SERIES

ESTIMATING RACIAL DISPARITIES WHEN RACE IS NOT OBSERVED

Cory McCartan
Robin Fisher
Jacob Goldin
Daniel E. Ho
Kosuke Imai

Working Paper 32373
<http://www.nber.org/papers/w32373>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2024

We thank Bruce Willsie, CEO of L2, Inc., for providing us with the geocoded voter file we use in this paper. We also thank Hiroto Katsumata, Soichiro Yamauchi, and an anonymous reviewer of the Alexander and Diviya Magaro Peer Pre-Review Program for useful feedback. The views expressed in this paper are those of the authors and do not necessarily represent the views of the US Treasury Department or the National Bureau of Economic Research. Any taxpayer data used in this research was kept in a secured IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2024 by Cory McCartan, Robin Fisher, Jacob Goldin, Daniel E. Ho, and Kosuke Imai. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Estimating Racial Disparities When Race is Not Observed

Cory McCartan, Robin Fisher, Jacob Goldin, Daniel E. Ho, and Kosuke Imai

NBER Working Paper No. 32373

April 2024

JEL No. C10,H22

ABSTRACT

The estimation of racial disparities in various fields is often hampered by the lack of individual-level racial information. In many cases, the law prohibits the collection of such information to prevent direct racial discrimination. As a result, analysts have frequently adopted Bayesian Improved Surname Geocoding (BISG) and its variants, which combine individual names and addresses with Census data to predict race. Unfortunately, the residuals of BISG are often correlated with the outcomes of interest, generally attenuating estimates of racial disparities. To correct this bias, we propose an alternative identification strategy under the assumption that surname is conditionally independent of the outcome given (unobserved) race, residence location, and other observed characteristics. We introduce a new class of models, Bayesian Instrumental Regression for Disparity Estimation (BIRDIE), that take BISG probabilities as inputs and produce racial disparity estimates by using surnames as an instrumental variable for race. Our estimation method is scalable, making it possible to analyze large-scale administrative data. We also show how to address potential violations of the key identification assumptions. A validation study based on the North Carolina voter file shows that BISG+BIRDIE reduces error by up to 84% when estimating racial differences in party registration. Finally, we apply the proposed methodology to estimate racial differences in who benefits from the home mortgage interest deduction using individual-level tax data from the U.S. Internal Revenue Service. Open-source software is available which implements the proposed methodology.

Cory McCartan
New York University
cm6753@nyu.edu

Daniel E. Ho
Stanford University
dho@law.stanford.edu

Robin Fisher
U.S. Department of the Treasury
1500 Pennsylvania Ave
Washington, DC 20220
Robin.fisher@treasury.gov

Kosuke Imai
Harvard University
1737 Cambridge Street
Cambridge, MA 02138
imai@harvard.edu

Jacob Goldin
University of Chicago Law School
1111 E 60th Street
Chicago, IL 60637
and NBER
jsgoldin@gmail.com

Code repository is available at <https://corymccartan.com/birdie/>

1 Introduction

The identification and estimation of racial disparities is of paramount importance to researchers, policymakers and organizations in a variety of areas including public health (Van Ryn and Fu, 2003; Williams and Jackson, 2005), employment (Conway and Roberts, 1983; Greene, 1984), voting (Gay, 2001; Hajnal and Trounstein, 2005; Barreto, 2007), criminal justice (Berk et al., 2021; Chouldechova, 2017; Dressel and Farid, 2018), economic policy and taxation (Brown, 2022; Elzayn et al., 2023), housing (Kermani and Wong, 2021), lending (Chen, 2018), and technology and fairness (Alao et al., 2021). Within the U.S. government, efforts to identify and remedy racial disparities have taken on greater urgency with the recent issuance of Executive Order 13985, which in part directs agencies to conduct equity assessments by developing appropriate methodology.

In many of these areas, however, racial information is not available at the individual level. The unavailability of individual racial information makes it impossible for analysts to simply tabulate variables of interest against race to identify disparities among different racial groups. In fact, in some areas, the law explicitly prohibits the collection of racial information even as it demands fair treatment on the basis of race (see, e.g., the U.S. Equal Credit Opportunity Act). This creates a dilemma for organizations who wish to measure possible disparities in order to monitor the fairness of their decision-making or service provision.

To estimate racial disparities without individual-level racial data, some researchers have turned to ecological inference methods (Goodman, 1953; King, 1997; King et al., 2004; Wakefield, 2004; Imai et al., 2008; Greiner and Quinn, 2009). These methods, however, require strong assumptions, which can be difficult to verify and may provide misleading results (Cho and Manski, 2008). Additionally, they all rely on accurate marginal information about race, which may not always be available.

Where the analysis of racial disparities involves large-scale administrative data, many analysts have adopted Bayesian Improved Surname Geocoding (BISG), which generates individual probabilities of belonging to different racial groups using Bayes' rule applied to last names and geographic location (Fiscella and Fremont, 2006; Elliott et al., 2008; Imai and Khanna, 2016). BISG leverages residential racial segregation and the association between self-reported race and surname to produce generally accurate and calibrated predictions of self-reported individual race (Kenny et al., 2021; DeLuca and Curiel, 2022).

Much attention has been given to ways of increasing the accuracy of BISG and related methods for race *prediction* (Voicu, 2018; Zest AI, 2020; Argyle and Barber, 2022; Imai et al., 2022; Decter-Frain, 2022; Greengard and Gelman, 2023). Unfortunately, broadly accurate BISG racial prediction alone does not guarantee the unbiased estimation of racial *disparities*, which is the ultimate goal of most analysts. To estimate disparities, BISG probabilities (or any other racial predictions) must be combined with information on the outcome variable for which the disparities are of interest. But the most common techniques for doing so are known to be biased when race is correlated with the outcome even after controlling on name and location (Chen et al., 2019; Argyle and Barber, 2022). These approaches include *weighting* the outcome variable by the BISG probabilities, and *thresholding* the BISG probabilities to produce point estimates of individual race (e.g., imputing "Black" as the race for an individual with a 61% probability of being Black according to BISG).

In fact, these methods often *underestimate* the true magnitude of racial disparities, which is problematic for policymakers and analysts who aim to identify and reduce these disparities. As formally discussed in Section 2, the standard methods of racial disparity estimation based on BISG predictions implicitly require individuals' race to be conditionally independent of the outcome given their residence location, surnames, and other observable attributes. This key identification assumption, however, is unlikely to hold because race affects many aspects of society even after accounting for residence location, surnames, and other observable attributes. Some researchers have noted the implausibility of this assumption and have proposed methods to address it, but these alternative approaches are more general than the racial disparity setting and require additional data such as a

partially labeled subset (Fong and Tyler, 2021). Others have advocated for partial identification strategies (Kallus et al., 2021; Elzayn et al., 2023).

To address this challenge, in Section 3, we propose an alternative identification strategy. Specifically, we assume that the outcome is conditionally independent of surname given (unobserved) individual’s race, residence location, and other observed attributes. This assumption is a type of exclusion restriction where surname serves as an instrumental variable for unobserved race. It implies that for two individuals who live in the same area, belong to the same racial group, and share the observable attributes, their surnames have no predictive power of the outcome. Somewhat counter-intuitively, the high-dimensionality of surnames aids rather than hinders identification because it provides a large number of instruments. We argue that this new identification assumption is more credible than the commonly invoked assumption unless surname is directly used to determine the outcome of interest (i.e., name-based discrimination).

Leveraging this identification strategy, in Section 3.2 we introduce a new class of models, Bayesian Instrumental Regression for Disparity Estimation (BIRD_iE), that accurately estimates racial disparities using BISG probabilities. Beyond accuracy, BIRD_iE improves on standard methodology in several ways:

- BIRD_iE includes built-in flexibility for researchers to make problem-specific modeling choices (Section 3.2).
- BIRD_iE can be fit with an EM algorithm that can scale to hundreds of thousands or millions of observations (Section 3.3).
- BIRD_iE produces updated BISG probabilities that incorporate the outcome variable and are likely to be more accurate than the BISG probabilities based only on surnames and geolocation (Section 3.4).
- BIRD_iE can be used iteratively to condition on additional variables whose distribution by race is not known *a priori* (Section 3.5). For example, party identification can be estimated by race *and* turnout.

Finally, in Section 3.6 we address potential violations of the key identification assumption, such as the one caused by overly coarse racial categories, by exploiting auxiliary information about the relations between names and more specific ethnic groups. All of the proposed methodology is implemented in a computationally efficient open-source software package, *birdie*, that is made available with the paper. The software and accompanying documentation are available at <https://corymccartan.com/birdie/>.

In Section 4, we validate the proposed methodology using real-world data taken from the voter file in North Carolina, where self-reported individual-race is observed and can be used to construct the ground-truth of racial disparities. BIRD_iE substantially outperforms existing estimators across different error measures and multiple levels of geolocation specificity. For example, the most popular existing BISG-only disparity estimator pegs the gap at Democratic party registration between White and Black voters at 24.1 percentage points, while the actual gap is 54.6 percentage points—more than double. Our preferred BIRD_iE model using the same BISG probabilities yields an estimate of 48.5 percentage points. This represents about a 80% reduction in bias.

In Section 5, we apply BIRD_iE to large-scale administrative tax data from the U.S. Internal Revenue Service (IRS), which does not collect individual taxpayers’ racial information. We produce the novel estimates of the distribution by race in who claims the home mortgage interest deduction—a question that has been largely hampered by a lack of administrative tax data with taxpayers’ racial information. The results show that there exists a substantial degree of racial disparity with many fewer Black and Hispanic filers claiming the HMID than White and Asian filers. We find that the racial gaps in homeownership rates alone cannot explain this disparity. Section 6 gives concluding remarks.

2 Bias of the Standard Methodology

In this section, we review the assumptions of the standard BISG-based methodology for estimating racial disparities when individual race is not observed. We show that the racial disparity estimates based on the standard methodology are biased unless the outcome variable is independent of race given surname, residence location, and other observed covariates. We argue that this assumption is likely to be violated given the significant role race plays in many aspects of our society.

2.1 Setup and BISG Procedure

Suppose that we have an i.i.d. sample of N individuals from a population of infinite size. For each individual $i = 1, 2, \dots, N$, we define a tuple $(Y_i, R_i, G_i, X_i, S_i)$, where $Y_i \in \mathcal{Y}$ is the outcome of interest for individual i , $R_i \in \mathcal{R}$ is the (unobserved) race of the individual, $G_i \in \mathcal{G}$ is the (geo)location of the individual's residence, $X_i \in \mathcal{X}$ are other observed characteristics, and $S_i \in \mathcal{S}$ is the individual's surname. When we are not referring to a particular individual, we will drop the subscripts for notational simplicity. Note that individual race is unobservable but all other variables are assumed to be observed. The availability of particular (or any) X is not required for either the standard or proposed methodology.

We assume throughout that these variables are discrete, taking a finite set of values, i.e., $|\mathcal{Y}|$, $|\mathcal{R}|$, $|\mathcal{G}|$, $|\mathcal{X}|$, and $|\mathcal{S}|$ are constants. Note that typically \mathcal{S} is high-dimensional as there exist a large number of unique surnames. In practice, residence location G is also discrete, since joint information about location, race, and other variables is generally only available down to the Census block level. For the sake of simplicity, we assume that the outcome variable Y is also discrete, though it is possible to extend the standard and proposed methodologies to continuous outcome variables.

Typically, BISG relies on data from the decennial Census or the American Community Survey (ACS), which provide information on the joint distribution of R and G (and any other covariates X , such as gender or age). It then combines this information with data from the Census Bureau's surname tables (U.S. Census Bureau, 2014), which provide information on the joint distribution of R and S . We summarize this set of information from the Census by two conditional probabilities, $\mathbf{q}_{GX|R}$ and $\mathbf{q}_{S|R}$, and one marginal probability, \mathbf{q}_R .

The BISG estimator of the probability that individual i belongs to race $r \in \mathcal{R}$ can then be written as (Fiscella and Fremont, 2006; Elliott et al., 2008)

$$\hat{P}_{ir} := \frac{q_{G_i X_i | r} q_{S_i | r} q_r}{\sum_{r' \in \mathcal{R}} q_{G_i X_i | r'} q_{S_i | r'} q_{r'}}, \quad (1)$$

where, for example, $q_{G_i X_i | r}$ indicates the estimated conditional probability of residence location G_i and covariates X_i given race r , taken from the Census table $\mathbf{q}_{GX|R}$.

The BISG estimator relies on two key assumptions. The first is that the Census tables reflect the true population distributions of R , S , G , and X .

Assumption ACC (Data accuracy). *For all i ,*

$$\begin{aligned} \mathbb{P}(R_i = r) &= q_r \\ \mathbb{P}(S_i = s \mid R_i = r) &= q_{s|r} \\ \mathbb{P}(G_i = g, X_i = x \mid R_i = r) &= q_{gx|r} \end{aligned}$$

Despite the best efforts of the Census Bureau, Assumption ACC may never hold exactly in practice. The decennial census has intrinsic error, including undercounting minority racial groups (U.S. Census Bureau,

2022; Anderson and Fienberg, 1999; Strmic-Pawl et al., 2018), as well as error introduced by privacy-preserving mechanisms (Abowd et al., 2022; Kenny et al., 2023a). And because of births, deaths, and moves, census data are often out-of-date from the moment of publication. These errors have led further extensions of the BISG estimator to account for some of this measurement error (Imai et al., 2022), which can help with accuracy for smaller racial groups. The plausibility of Assumption ACC is stretched even further when the study population is a subset of the whole U.S. population, and so is not covered by national census data. In these cases, analysts should set \mathbf{q}_R to the known or estimated marginal racial distribution in the study population, rather than the national racial distribution, when this margin is known. It may be more plausible then to assume that the conditional distributions $\mathbb{P}(S | R)$ and $\mathbb{P}(G, X | R)$ match the census distributions, even if $\mathbb{P}(R)$ does not (Rosenman et al., 2023). Greengard and Gelman (2023) take this approach a step further by raking BISG probabilities to all known margins, further improving calibration.

The second assumption required by BISG is the following conditional independence relation between an individual’s surname and residence location (as well as other characteristics) given their unobserved race.

Assumption CI-SG (Conditional independence of name and other proxy variables). *For all i ,*

$$S_i \perp \{G_i, X_i\} | R_i.$$

Assumption CI-SG implies, for example, that once we know an individual is White, knowing their surname is Smith tells us nothing about their residence location and other observed characteristics. Although this assumption appears to be reasonable, the lack of granularity in the coding of race may lead to its violation. For example, people with Chinese, Indian, Filipino, Vietnamese, Korean, or Japanese are all coded as one racial group “Asian” in the census. These groups, however, have varying sets of surnames and have different demographic and geographic distributions. For instance, unlike the Smith example, knowing that an Asian individual’s surname is Gupta makes it more likely that they have a higher income and live in the Eastern U.S (Budiman et al., 2019).

Even though Assumptions CI-SG and ACC may not hold exactly, researchers find that BISG produces accurate and generally well-calibrated estimates in practice (Imai and Khanna, 2016; Zhang, 2018; Kenny et al., 2021; DeLuca and Curiel, 2022). We observe this pattern as well in the validation study in Section 4.

Under Assumption CI-SG, by Bayes’ Rule,

$$\mathbb{P}(R_i = r | G_i, X_i, S_i) \propto \mathbb{P}(G_i, X_i | R_i = r) \mathbb{P}(S_i | R_i = r) \mathbb{P}(R_i = r).$$

This justifies the estimator given in Equation (1), and provides us with the following immediate result.

Proposition 2.1 (Accuracy of BISG). *Under Assumptions CI-SG and ACC, the BISG estimator produces correct probabilities. That is, we have $\hat{P}_{ir} = \mathbb{P}(R_i = r | G_i, X_i, S_i)$.*

New methods continue to be developed that improve the calibration of BISG probabilities, including some machine learning methods based on labeled data (Zest AI, 2020; Imai et al., 2022; Argyle and Barber, 2022; Decter-Frain, 2022; Greengard and Gelman, 2023; Cheng et al., 2023). Fundamentally, these approaches focus on building a more accurate model for $R | G, X, S$ at the individual level.

2.2 Bias of Racial Disparity Estimates based on BISG Probabilities

To estimate racial disparities, BISG probabilities (or other racial predictions) must be combined with the outcome variable. There are several common ways researchers do this.

The most frequent is perhaps the *thresholding* or *classification* estimator, which deterministically assigns individuals to a predicted racial category based on the BISG estimates \hat{P}_i (either the largest \hat{P}_{ir} or the one which

exceeds a predetermined threshold). Estimates of $\mathbb{P}(Y = y \mid R = r)$ are then obtained by tabulating the data by these assigned categories.

Another common approach, which attempts to capture the uncertainty inherent in race prediction, is the following *weighting* estimator:

$$\hat{\mu}_{Y|R}^{(wtd)}(y \mid r) = \frac{\sum_{i=1}^N \mathbf{1}\{Y_i = y\} \hat{P}_{ir}}{\sum_{i=1}^N \hat{P}_{ir}}.$$

Unfortunately, accurate and calibrated estimates of individual race predictions alone, unless they are perfect, are not sufficient for unbiased estimation of racial disparities using these standard methodologies.

This should come as no surprise for the thresholding estimator, since it does not take into account prediction uncertainty in the BISG probabilities. This is akin to ignoring measurement errors in the covariates of a regression, something which has long been known to lead to biased coefficient estimates. Unlike the classical errors-in-variables setting, however, the bias of the thresholding estimator is not consistently in the same direction, making it hard to reason about (Chen et al., 2019).

But, the weighting estimator is also biased because the prediction error of race probabilities may be correlated with the outcome variable of interest. Fortunately, unlike the threshold estimator, it is easier to understand the nature of this bias. Chen et al. (2019) show that the asymptotic bias of the weighting estimator is controlled by the residual correlation of Y and R after adjusting for G , X , and S . We reproduce this result here.

Theorem 2.2 (Theorem 3.1 of Chen et al. 2019). *If race is binary (so $\mathcal{R} = \{0, 1\}$), then as $N \rightarrow \infty$,*

$$\hat{\mu}_{Y|R}^{(wtd)}(y \mid r) - \mathbb{P}(Y = y \mid R = r) \xrightarrow{a.s.} -\frac{\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = r\} \mid G, X, S)]}{\mathbb{P}(R = r)}.$$

This result implies that when the BISG residuals $\mathbf{1}\{R = r\} - \mathbb{P}(R = r \mid G, X, S)$ are correlated with the outcome, estimates will be biased, even with infinite data. In fact, the weighting estimator will often underestimate the magnitude of a disparity, as the following corollary shows. For instance, in measuring disparities in loan approval (Y), if Blacks are less likely to be approved for loans across all locations and surnames than Whites, then the weighting estimator would understate the resulting overall White-Black disparity in loan approval rates.

Corollary 2.2.1 (Underestimation of racial disparity). *Let $y \in \mathcal{Y}$. If race is binary (so $\mathcal{R} = \{0, 1\}$), and $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$, then*

$$\hat{\mu}_{Y|R}^{(wtd)}(y \mid 1) - \hat{\mu}_{Y|R}^{(wtd)}(y \mid 0) < \mathbb{P}(Y = y \mid R = 1) - \mathbb{P}(Y = y \mid R = 0).$$

Conversely, Theorem 2.2 implies that conditional independence between individual's race and outcome given their surname, residence location, and other characteristics is sufficient to eliminate the asymptotic bias of the weighting estimator.

Assumption CI-YR (Conditional independence of outcome and race). *For all i ,*

$$Y_i \perp\!\!\!\perp R_i \mid G_i, X_i, S_i.$$

Figure 1(a) shows a causal directed acyclic graph (DAG) that satisfies Assumption CI-YR as well as Assumption CI-SG. The dashed node border for R represents the fact that race is unobserved. The causal structure in Figure 1(a) implies the conditional independence relation $Y \perp\!\!\!\perp R \mid G, X, S$, because all paths from R to Y are blocked by G , X , or S . The key causal assumption of this DAG is that the effect of race R on the outcome Y must

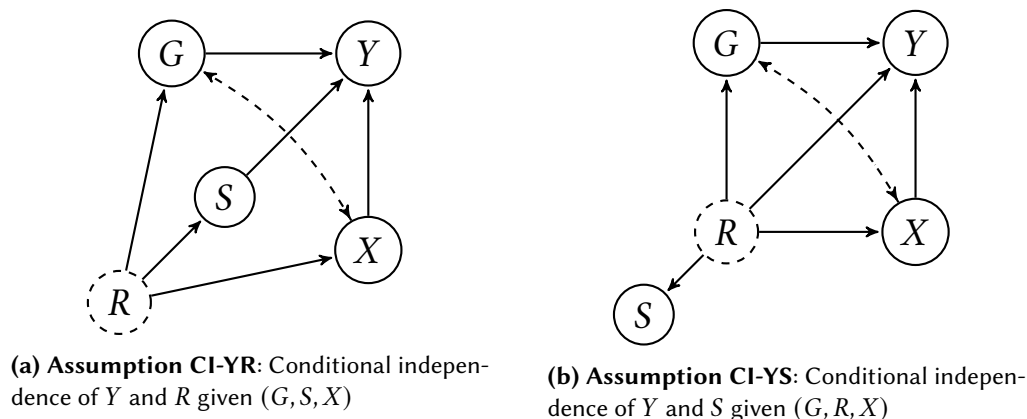


Figure 1: Possible causal structures for which each of the labeled assumptions is satisfied, represented as a directed acyclic graph (DAG) where G is residence location, R is race, S is surname, X is observed covariates, and Y is the outcome. Race R is unobserved, which is signified by a dashed node boundary. Both DAGs also satisfy Assumption CI-SG: Conditional independence of S and (G, X) given R .

be entirely mediated by surname S , residence location G , and other observed characteristics X . This type of exclusion restriction may not be credible in many practical settings because race can affect the outcome through so many factors, biasing the weighting estimator.

But in other settings, Assumption CI-YR may be more plausible. For example, for a manager reviewing job applications on paper, race is typically unobserved, but the manager may be influenced by racial or gender cues in a candidate’s name or address (Park et al., 2009; Åslund and Skans, 2012). So long as we observe all information used by the manager and use it in the BISG estimation, the weighting estimator would give an asymptotically unbiased estimate. Similarly, in evaluating the fairness of algorithmic decision-making, as long as all the information used by the algorithm is incorporated into BISG, Assumption CI-YR would be appropriate. Outside of these cases, however, the weighting estimator is likely to be biased.

Finally, while the discussion in this section has been focused on the BISG methodology, the qualitative results and necessary assumptions carry over to other approaches which produce probabilistic predictions of individual race (Zest AI, 2020; Argyle and Barber, 2022; Imai et al., 2022; Decter-Frain, 2022; Greengard and Gelman, 2023). Just like standard BISG, all of these methods are based on individual names, geographic location (and sometimes other geographic attributes, like ACS or decennial census statistics), and possibly additional individual covariates. Just as with BISG, well-calibrated probabilities are not generally sufficient to produce unbiased estimates of racial disparities using the standard weighting or thresholding estimators.

3 The Proposed Methodology

In this section, we propose an alternative identification strategy for racial disparities that allow race to directly affect the outcome of interest. We show that racial disparity is identifiable if surname is conditionally independent of the outcome given race, residence geolocation, and other observed information, and the aforementioned assumptions required by BISG hold. We develop a class of statistical models, called Bayesian Instrumental Regression for Disparity Estimation (BIRDIE), that estimate racial disparity under this identification condition by using surnames as a high-dimensional instrumental variable for race. These models take as inputs the BISG probabilities, and so can be easily applied on top of existing analysis pipelines, including those with alternative probabilistic race prediction methodologies. We also discuss computational strategies that can be used to apply BIRDIE models to large data sets, and an extension of the methodology to include an additional explanatory

variable that was not used at the BISG stage. Finally, we show how to address the potential violations of the key identification assumptions, such as those caused by name-based discrimination.

3.1 New Identification Strategy

To reduce the potential bias of the weighting estimator, we propose an alternative identification assumption that may be applicable when Assumption CI-YR is not credible. Specifically, we assume that surname, rather than race, satisfies the exclusion restriction conditional on (unobserved) race, residence location, and other observed characteristics.

Assumption CI-YS (Conditional independence of outcome and name). *For all i ,*

$$Y_i \perp\!\!\!\perp S_i \mid R_i, G_i, X_i.$$

Figure 1(b) shows one possible causal DAG that meets this assumption as well as Assumption CI-SG. In this DAG, race can have a direct effect on the outcome Y as well as on residence location G and other observed characteristics X , while all paths from S to Y are blocked by G , X , or R .

This causal structure is plausible in many real-world settings because, unlike Assumption CI-YR, Assumption CI-YS allows race to directly affect the outcome. For an outcome like party registration, Assumption CI-YS would mean that among White voters in a particular geographic region, voters named Smith would *a priori* be no more or less likely to identify with one party than voters named Thomas. In contrast, Assumption CI-YR would mean that among voters named Smith in a particular geographic region, White voters would be *a priori* be no more or less likely to identify with one party than Black voters. In this case, Assumption CI-YR is likely to be violated, while Assumption CI-YS is plausible. It is important to note a key trade-off between the two assumptions. While Assumption CI-YS rules out the possibility that surname directly affects the outcome (e.g., name-based discrimination), such a direct effect is allowed under Assumption CI-YR. Section 3.6 revisits this important issue.

The appropriateness of each assumption depends on a specific application. In the above hiring example, if the manager reviews applicants anonymously, there will be no name-based discrimination and the assumption is likely to be satisfied. The assumption may be violated in other contexts, however. For example, in studying turnout, if campaigns use the surnames of individual voters to decide whether to mobilize them, Assumption CI-YS will be violated. Yet another possible violation of the assumption is the existence of unobserved confounder that affects both outcome and surname. The country of origin for an immigrant may represent such a confounder: where surnames are informative of country of origin, even within racial groups (as is often the case for Asian individuals), variations in outcomes by country of origin will likely violate CI-YS. This reflects the limitations of the relatively coarse racial classifications used in BISG, as discussed above. Even in these cases, however, conditioning on (unobserved) race is likely to substantially reduce the magnitude of association between outcome and surnames. In Section 3.6, we show how to address this potential violation of Assumption CI-YS.

We briefly note that Assumptions CI-YR and CI-YS are not necessarily mutually exclusive. For example, neither race or surname could have a direct causal effect on the outcome. In these cases, both the weighting estimator and our approach proposed below could give reasonable answers, though the data requirements of each may differ.

The following theorem formally shows that it is possible to nonparametrically identify racial disparities under Assumption CI-YS. The proof of the theorem and all other results in this paper are deferred to the appendix.

Theorem 3.1 (Nonparametric Identification). *For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries*

$b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then, under Assumption CI-YS, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, G = g, X = x)$ are identified if and only if both \mathbf{P} and the augmented matrix $(\mathbf{P} \ \mathbf{b})$ have rank $|\mathcal{R}|$.

The essence of this identification result is the following simple observation. Under Assumption CI-YS, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\mathbb{P}(Y = y \mid G = g, X = x, S = s) = \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x) \mathbb{P}(R = r \mid G = g, X = x, S = s). \quad (2)$$

The leftmost term is estimable from the data and corresponds to the vector \mathbf{b} in Theorem 3.1, while the rightmost term is the BISG estimand and corresponds to the matrix \mathbf{P} in Theorem 3.1. Lastly, the remaining term in the middle can be solved for, since Equation (2) holds across all combinations of Y , G , X , and S , leading to a large system of linear equations. Specifically, we have $(|\mathcal{Y}|-1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{S}|$ equations with $(|\mathcal{Y}|-1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{R}|$ unknowns. Since $|\mathcal{R}| \ll |\mathcal{S}|$, we can identify these unknowns as long as the linear system has sufficient rank. Our result is closely related to causal identification based on proxy variables in the presence of unmeasured confounding (Kuroki and Pearl, 2014; Miao et al., 2018; Knox et al., 2022). Here, we use surname as a proxy variable for (unobserved) race to identify racial disparities.

Together with Proposition 2.1, Theorem 3.1 implies that racial disparities can be identified under Assumptions CI-SG, ACC, and CI-YS. The identifying equation (2) shows that $\mathbb{P}(Y = y \mid G = g, X = x, S = s)$ is linear in the BISG estimands $\mathbb{P}(R = r \mid G = g, X = x, S = s)$. Thus, it is natural to consider the following least-squares estimator of $\mathbb{P}(Y = y \mid R, G = g, X = x)$ under this alternative identification strategy,

$$\hat{\mu}_{Y|RGX}^{(\text{ols})}(y \mid \cdot, g, x) = (\hat{\mathbf{P}}_{I(xg)}^\top \hat{\mathbf{P}}_{I(xg)})^{-1} \hat{\mathbf{P}}_{I(xg)} \mathbf{1}\{\mathbf{Y}_{I(xg)} = y\},$$

where as above $\hat{\mathbf{P}}$ is the matrix of BISG probabilities, and $I(xg)$ is the set of individuals i with $X_i = x$ and $G_i = g$. Here and throughout the paper, a dot will indicate a vector constructed over that index, so $\hat{\mu}_{Y|RGX}^{(\text{ols})}(y \mid \cdot, g, x)$ is a vector of conditional probabilities for a particular outcome level y across all racial groups in \mathcal{R} . By post-stratifying this estimator across the (G, X) cells, we arrive at an estimator of $\mathbb{P}(Y = y \mid R)$,

$$\hat{\mu}_{Y|R}^{(\text{p-ols})}(y \mid r) = \sum_{x \in \mathcal{X}, g \in \mathcal{G}} (\hat{\mathbf{P}}_{I(xg)}^\top \hat{\mathbf{P}}_{I(xg)})^{-1} \hat{\mathbf{P}}_{I(xg)} \mathbf{1}\{\mathbf{Y}_{I(xg)} = y\} q_{gx|r},$$

since $q_{gx|r} = \mathbb{P}(G = g, X = x \mid R = r)$ under Assumption ACC. This estimator is unbiased, as the following theorem shows.

Theorem 3.2 (Unbiasedness of OLS Estimator). *If Assumptions CI-SG, ACC, and CI-YS hold, and the identification conditions in Theorem 3.1 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,*

$$\mathbb{E}[\hat{\mu}_{Y|R}^{(\text{p-ols})}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

It is worth comparing this OLS estimator with the weighting estimator $\hat{\mu}_{y|r}^{(\text{wtd})}$. The next theorem shows that within the (G, X) cells these two estimators are guaranteed to disagree, unless either the BISG probabilities perfectly discriminate or the weighting estimator is constant across races. Unfortunately, these two conditions are almost never met in practice. This underscores the importance of selecting the appropriate assumption (Assumption CI-YR or CI-YS) for a particular analysis, since they imply different estimators with different results.

Theorem 3.3 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators). *For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\hat{\mu}_{Y|R}^{(wtd)}(y | \cdot) = \hat{\mu}_{Y|R}^{(ols)}(y | \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j | G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k | G_i, X_i, S_i) = 0$ and vice versa) or $\hat{\mu}_{Y|R}^{(wtd)}(y | j) = \hat{\mu}_{Y|R}^{(wtd)}(y | k)$.*

Despite potential advantages over the weighting estimator, the OLS estimator is not well-suited to estimation in practice, since it ignores the fact that the unknown parameters are probabilities and thus constrained to be nonnegative and sum to 1. As a result, in any particular sample, the estimator can produce impossible or contradictory estimates. This is particularly problematic because a large number of unique surnames make both \mathbf{P} and \mathbf{b} high-dimensional. To address this challenge, and to open the door to more flexible modeling, we next propose a Bayesian modeling approach that is based on our identification strategy and yet satisfies necessary constraints.

3.2 Bayesian Instrumental Regression for Disparity Estimation

The BIRDIE approach combines a user-specified complete-data outcome model $\pi(Y | R, G, X, \Theta)$, parametrized by Θ , with the BISG model in order to estimate the distribution $Y | R$ that is of interest. In this regard, it mirrors the two-stage instrumental variables (IV) regression: a first stage (BISG) that estimates the relationship between instrument (surname) and variable of interest (race), and a second stage (BIRDIE) that uses the first-stage estimates to produce valid estimates of the quantity of interest. Unlike two-stage IV, however, the BIRDIE approach is based on a coherent joint distribution of data, unknown parameters, and race. We exploit this fact and develop the general BIRDIE modeling approach below.

Specifically, the BIRDIE posterior is obtained by applying Assumptions CI-SG, ACC, and CI-YS to the joint distribution $\pi(Y, R, G, X, S, \Theta)$:

$$\begin{aligned} \pi(\Theta, \mathbf{R} | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) &\propto \pi(\Theta, \mathbf{R}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \\ &\propto \pi(\Theta) \prod_{i=1}^N \pi(Y_i | R_i, G_i, X_i, \Theta) \pi(R_i | G_i, X_i, S_i) \\ &= \pi(\Theta) \prod_{i=1}^N \pi(Y_i | R_i, G_i, X_i, \Theta) \hat{P}_{iR_i}, \end{aligned} \quad (3)$$

where as above \hat{P}_i are the BISG probability estimates for individual i , which depend on Census data represented in $\mathbf{q}_{GX|R}$, $\mathbf{q}_{S|R}$, and \mathbf{q}_R , but not on the outcome-model parameters Θ . As a result, these “first-stage” BISG estimates can be plugged directly into the BIRDIE posterior computation without any loss of Bayesian coherency. This remains true if a more complex model is used in place of the BISG probabilities (such as those of [Zest AI, 2020](#); [Imai et al., 2022](#); [Argyle and Barber, 2022](#); and [Decter-Frain, 2022](#)), so long as the parameters of the first-stage model are *a priori* independent of the parameters of the BIRDIE model.

To apply this general BIRDIE model to a particular analysis requires choosing a complete-data outcome model, given by the likelihood $\pi(Y_i | R_i, G_i, X_i, \Theta)$ and prior $\pi(\Theta)$. Since Y is discrete, a categorical regression model is appropriate for $\pi(Y_i | R_i, G_i, X_i, \Theta)$. The parametrization of the categorical regression will depend on the analyst’s goals, computational resources, and prior beliefs about the structure of the problem. We present here several reasonable alternatives that trade off modeling flexibility and computational efficiency.

Complete-pooling model. The simplest possible model is one in which the relationship between Y and R does not vary with G or X . This model is parametrized by $\Theta = \{\theta_r\}_{r \in \mathcal{R}}$, which describe the distribution of Y

within every level of R :

$$Y_i | R_i, G_i, X_i, \Theta \sim \text{Categorical}_{\mathcal{Y}}(\theta_{R_i})$$

$$\theta_r \stackrel{iid}{\sim} \text{Dirichlet}(\alpha),$$

where $\text{Categorical}_{\mathcal{Y}}$ denotes a discrete (categorical) distribution on the set \mathcal{Y} . With known \mathbf{R} , the posterior of θ_r is conjugate, a fact which will make computation under the EM scheme described in Section 3.3 below extremely efficient. Of course, this efficiency comes at the cost of a restrictive model that allows for no role of G and X . If in reality $\mathbb{P}(Y | R)$ does vary along these dimensions, it is possible that the posterior of θ_r will not accurately estimate $\mathbb{P}(Y | R)$. In any case, if the analyst is interested in subgroup or small-area estimates of $\mathbb{P}(Y | R)$, the complete-pooling model will be of little use.

Saturated (no-pooling) model. At the other end of the spectrum from the complete-pooling model is a *saturated* or no-pooling model, which estimates a different distribution of $Y | R$ within every level of G and X :

$$Y_i | R_i, G_i, X_i, \Theta \sim \text{Categorical}_{\mathcal{Y}}(\theta_{R_i G_i X_i})$$

$$\theta_{r g x} \stackrel{iid}{\sim} \text{Dirichlet}(\alpha).$$

This model is closest to the OLS estimator, though it ensures that all probability estimates lie in $[0, 1]$. As with the complete-pooling model, the posterior of $\theta_{r g x}$ is conjugate to its prior, and so computation can be made efficient. Additionally, this model allows for any arbitrary relationship between Y , R , G , and X . Since it is fully nonparametric, the posterior will converge to the true $\mathbb{P}(Y | R, G, X)$ with enough data in each (G, X) cell. However, in practice, the model can suffer from the curse of dimensionality: the number of (G, X) cells may be relatively large compared to the amount of available data, or even exceed it, especially since G can be quite large, covering many blocks or ZIP codes. In these cases, the prior will dominate the data in each cell, which could have a large biasing effect even on overall inferences about $\mathbb{P}(Y | R)$.

General mixed-effects model. As a compromise between the complete-pooling and no-pooling model, a partial pooling approach based on a multinomial mixed-effects model can be used. Properly specified, the mixed-effects model maintains the flexibility of the saturated model while avoiding its high bias and variance in finite samples.

$$Y_i | R_i, G_i, X_i, \Theta \sim \text{Categorical}_{\mathcal{Y}}(g^{-1}(\mu_{r g x}))$$

$$\mu_{r g x y} = \mathbf{W} \beta_{r y} + \mathbf{Z} \mathbf{u}_{r y}$$

$$\mathbf{u}_{r y} | \phi_{r y} \sim \mathcal{N}\left(0, \Sigma(\phi_{r y})\right)$$

$$\beta_{r y} \stackrel{iid}{\sim} f_r^{(\beta)}, \quad \phi_{r y} \stackrel{iid}{\sim} f_r^{(\phi)},$$

where g^{-1} is a softmax or other link function, \mathbf{W} and \mathbf{Z} are matrices of fixed and random effects, respectively, ϕ is a vector of random-effect parameters, and $f^{(\beta)}$ and $f^{(\phi)}$ are some priors for the super-scripted parameters. Some fixed or random effects could be shared across combinations of R and Y , though this could complicate computation. In practice, we recommend including X and especially G in the model as random effects, with hierarchical structure as appropriate. Such a structure partially pools estimates of $\mathbb{P}(Y | R, X, G)$ towards an overall estimate of $\mathbb{P}(Y | R)$, allowing the model to share information between geographic areas. This should prove especially useful in cases where some areas have few or no observations for certain racial groups.

We also recommend including group-level covariates as fixed effects, which will help share information across random effects and significantly improve generalization performance to unseen random effect levels (Buttice and Highton, 2013). For example, if G records counties, analysts could include racial and socioeconomic variables

measured at the county level as predictors. This would help produce more accurate estimates of $\mathbb{P}(Y | R, G)$ to the extent that variation in these probabilities is associated with these racial and socioeconomic variables. Ultimately the structure of this general model will have to be chosen based on the data and the relevant research question.

3.3 Computation

The posterior in Equation (3) contains the high-dimensional discrete nuisance parameter \mathbf{R} , which poses a challenge for computation. We suggest two approaches for handling \mathbf{R} , one suited to small sample sizes, and one suited to large sample sizes. We also discuss uncertainty quantification for conjugate complete-data models.

Small samples: Inference directly on the marginal likelihood. Since \mathbf{R} is discrete, we can marginalize it out as follows:

$$\pi(\Theta | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) = \sum_{\mathbf{r} \in \mathcal{R}^N} \pi(\Theta, \mathbf{r} | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \propto \pi(\Theta) \prod_{i=1}^N \sum_{r \in \mathcal{R}} \pi(Y_i | r, G_i, X_i, \Theta) \hat{P}_{ir}. \quad (4)$$

This decouples the total number of parameters from the sample size. Additionally, Equation (4) has only continuous parameters, and so can be used with any general Bayesian inference procedure such as Markov chain Monte Carlo (MCMC). However, the moderate-to-high dimensionality in practical settings, even after integrating out \mathbf{R} , and the expensive likelihood calculation due to the sum nested within the outer product, makes MCMC algorithms computationally too expensive outside of relatively small datasets.

Large samples: Expectation-Maximization. When the number of individuals exceeds a thousand or so, we propose an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to calculate the maximum *a posteriori* (MAP) estimate of Θ for BIRDiE models. The EM algorithm alternates between an E-step which calculates the expected log posterior density Q , averaging over the missing \mathbf{R} , and an M-step which maximizes Q over values of Θ .

Specifically, given a current parameter estimate $\Theta^{(t)}$, the expected log posterior density can be written as

$$\begin{aligned} Q(\Theta^{(t+1)} | \Theta^{(t)}) &= \mathbb{E}_t[\log \pi(\Theta^{(t+1)}, \mathbf{R} | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})] \\ &= \sum_{\mathbf{r} \in \mathcal{R}^N} \pi(\mathbf{r} | \Theta^{(t)}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \log \pi(\Theta^{(t+1)}, \mathbf{r} | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \\ &= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^N \sum_{r \in \mathcal{R}} \left\{ \left(\log \pi(Y_i | r, G_i, X_i, \Theta^{(t+1)}) + \log \hat{P}_{ir} \right) \right. \\ &\quad \left. \times \pi(R_i = r | \Theta^{(t)}, Y_i, G_i, X_i, S_i) \right\} \\ &= C + \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^N \sum_{r \in \mathcal{R}} \tilde{P}_{ir|Y}^{(t)} \log \pi(Y_i | r, G_i, X_i, \Theta^{(t+1)}), \end{aligned}$$

where C is a constant that can be ignored as it does not depend on $\Theta^{(t+1)}$, and $\tilde{\mathbf{P}}_{|Y}^{(t)} = \pi(\mathbf{R} | \Theta^{(t)}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ are the BISG probabilities updated with Bayes' rule using the outcome \mathbf{Y} :

$$\tilde{P}_{ir|Y}^{(t)} = \frac{\pi(Y_i | r, G_i, X_i, \Theta^{(t)}) \hat{P}_{ir}}{\sum_{r' \in \mathcal{R}} \pi(Y_i | r', G_i, X_i, \Theta^{(t)}) \hat{P}_{ir'}}. \quad (5)$$

At the M-step, $Q(\Theta^{(t+1)} | \Theta^{(t)})$ is straightforward to maximize, since it is just the log complete-data posterior, with likelihood weights given by the $\tilde{P}_{ir|Y}^{(t)}$. Additionally, if Θ can be partitioned into parameters which only

affect individuals in each racial group (as is the case with all the models described in Section 3.2 above), then the maximization can be performed separately on each group of individuals.

A critical advantage of this EM scheme over working directly with the marginal likelihood is that the maximization in the M-step can be performed using sufficient statistics calculated as part of the E-step, rather than on all of the individual entries in the data. Since the M-step is usually the practical (if not also asymptotic) bottleneck in the computation, this is enormously helpful—the problem size scales with $|\mathcal{Y}| \times |\mathcal{X}| \times |\mathcal{G}|$ rather than with N . Specifically, notice that we can rewrite $Q(\Theta^{(t+1)} | \Theta^{(t)})$ (dropping the unnecessary constant) as

$$\begin{aligned} Q(\Theta^{(t+1)} | \Theta^{(t)}) &= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^N \sum_{r \in \mathcal{R}} \tilde{p}_{ir|Y}^{(t)} \log \pi(Y_i | r, G_i, X_i, \Theta^{(t+1)}) \\ &= \log \pi(\Theta^{(t+1)}) + \sum_{r \in \mathcal{R}} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \log \pi(y | r, g, x, \Theta^{(t+1)}) \left(\sum_{i \in \mathcal{I}(y x g)} \tilde{p}_{ir|Y}^{(t)} \right), \end{aligned}$$

where analogously to above $\mathcal{I}(y x g)$ is the set of individuals with $Y_i = y$, $X_i = x$, and $G_i = g$.

For BIRDIE models where the complete-data likelihood is conjugate to the prior, such as the complete- and no-pooling models, these sufficient statistics are used in the M-step anyway, and can be efficiently calculated during the E-step as well. In combination with the acceleration scheme described next, this allows the entire EM algorithm to be run to convergence on data with hundreds of thousands or millions of individuals in a matter of seconds.

While EM algorithms are highly stable, due to their monotonic increasing of the marginal likelihood, they are also often slow to converge (Laird, 1993). To address this, we propose, and include in our open-source software implementation, the use of fixed-point iteration accelerators such as Anderson acceleration or SQUAREM (Varadhan and Roland, 2008). These techniques can substantially reduce the overall computational time without meaningfully affecting the stability of inference.

Uncertainty quantification via blocked Gibbs sampling. While computationally efficient, the EM algorithm does not provide any uncertainty quantification. However, in large samples, sampling and model-based uncertainty dominated by biases caused by even small violations of the underlying assumptions, a problem we discuss below in Section 3.6 and the accompanying appendix. Even so, it is often useful to have some measure of sampling uncertainty. This is possible using blocked Gibbs sampling for some BIRDIE models with conjugate complete-data posteriors, such as the complete- and no-pooling models described above.

Our Gibbs sampling strategy is to alternate sampling from the model-updated BISG probabilities $\pi(\mathbf{R} | \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}, \Theta)$ and sampling from the complete-data posterior $\pi(\Theta | \mathbf{Y}, \mathbf{R}, \mathbf{G}, \mathbf{X}, \mathbf{S})$. The first step involves the same calculations as the E-step in Equation (5). The second step is computationally tractable in medium-to-large samples when the complete-data likelihood is conjugate to the prior, as is the case for the Categorical-Dirichlet pooling models proposed above.

Categorical-Dirichlet models with latent discrete variables, like the pooling BIRDIE models, are often tackled using a collapsed Gibbs sampler where Θ has been marginalized out. We found that approach to be unsuccessful here, however, since the one-by-one updating of each individual’s R_i meant that the sampler was unable to traverse to the correct region of parameter space and got stuck near the BISG initialization. In contrast, the blocked Gibbs sampler is vectorized over individuals and rapidly moves to the mode identified by the EM algorithm.

For non-conjugate BIRDIE models with few parameters, bootstrapping the EM procedure is computationally feasible and can be used to approximate the covariance matrix of the MAP estimate.

3.4 Updated Individual Race Probabilities

The EM algorithm produces the updated individual race probabilities given in Equation (5). One feature of these updated probabilities is that it is appropriate to apply the weighting estimator to them to estimate disparities. This is because the updated probabilities condition on \mathbf{Y} , and so the asymptotic bias term in Theorem 2.2 becomes zero. In fact, the weighting estimate from the updated probabilities is identical to the BIRDiE estimate. While more study is required, for downstream settings where weights are needed, generating these weights with BISG followed by BIRDiE will likely produce more accurate results than simply using BISG weights alone.

3.5 Additional Explanatory Variables

Often, researchers are interested in not just $\mathbb{P}(Y | R)$ but also $\mathbb{P}(Y | W, R)$, for some variable $W \in \mathcal{W}$ which is not part of the BISG predictors (X, G) . For example, a lending firm auditing potential racial disparities in lending decisions would likely be interested both in how the rate of loan approval (Y) varies by race, but also how loan approval varies by race, conditional on a measure of creditworthiness (W). The unconditional disparities reflect realities of systemic racism and inequality, while the conditional disparities measure the fairness of the firm's lending decisions after controlling for these systemic factors. Such estimates could be used to compute various measures of algorithmic fairness, including calibration parity and false positive error rate balance. Another scenario is a policy evaluation study, where researchers are interested in how the impact of policy varies across racial groups. Such an analysis requires incorporating an interaction between race and the treatment variable.

There are two main ways to perform such an analysis with our proposed methodology. The first, and perhaps simplest, is to apply the methodology to the combined variable $\underline{YW} \in \mathcal{Y} \times \mathcal{W}$. This will produce estimates of $\mathbb{P}(Y, W | R)$, from which $\mathbb{P}(Y | W, R)$ can be straightforwardly calculated by appropriate normalization. This approach will work well if $|\mathcal{Y}|$ and $|\mathcal{W}|$ are both small, so that $|\mathcal{Y} \times \mathcal{W}|$ is of manageable size. If one of these variables has many levels, however, directly estimating the distribution of $Y, W | R$ could be less efficient, as it does not account for information about the marginal distributions $Y | R$ and $W | R$.

An alternative approach is to first apply the proposed methodology to estimate $\mathbb{P}(W | R)$. This allows for calculation of model-updated BISG probabilities $\hat{\mathbf{P}}_{|W} = \pi(\mathbf{R} | \hat{\Theta}, \mathbf{W}, \mathbf{G}, \mathbf{X}, \mathbf{S})$, which are also computed as a byproduct of the EM algorithm described above. Then, the methodology can be applied again, using $\hat{\mathbf{P}}_{|W}$ as the input probabilities rather than the original BISG probabilities, to estimate $\mathbb{P}(Y | W, R)$. This approach will likely perform better when W consists of multiple predictors or if either $|\mathcal{Y}|$ or $|\mathcal{W}|$ are large.

Both of these approaches require the following identifying assumption, which generalizes Assumption CI-YS.

Assumption CI-YWS (Conditional independence of outcome, predictor and name). *For all i , $(Y_i, W_i) \perp\!\!\!\perp S_i | R_i, G_i, X_i$, or, equivalently,*

$$W_i \perp\!\!\!\perp S_i | R_i, G_i, X_i \quad \text{and} \quad Y_i \perp\!\!\!\perp S_i | W_i, R_i, G_i, X_i.$$

In the lending example, this would translate to the assumption that a measure of creditworthiness is independent of last name after controlling for race, location, and covariates, and that lending decisions are independent of last names after controlling for creditworthiness, race, location, and covariates.

3.6 Addressing Potential Violations of the Assumptions

BIRDiE crucially relies on Assumption CI-YS for identification. In addition, like the weighting and thresholding estimators, it also relies upon Assumptions CI-SG and ACC, which are required for the BISG race probabilities to be accurate. Unfortunately, these assumptions may not exactly hold in practice, and are also not testable in

observed data. In this section and Appendix C, we develop sensitivity analyses that assess how violations of these assumptions affect the estimates of racial disparities.

First, BIRDIE assumes that conditional on unobserved race and observed covariates, outcomes and surnames are independent. As discussed in Section 3.1, however, association between the outcome and country of origin or racial subgroups may lead to correlation between surnames and outcome even after controlling for race and geography. To address this problem, suppose that a low-dimensional summary statistic of surname, $f : \mathcal{S} \rightarrow \mathbb{R}^d$, $d \ll |\mathcal{S}|$, is available, where f may map each surname to a finer ethnic group within each racial category. For example, Imai is a Japanese name whereas McCartan is a name of Irish origin. If f can classify surnames into finer racial subgroups or countries of origin—even approximately—then it can be used to control for this channel of possible violations of Assumption CI-YS. Formally, we relax Assumption CI-YS as follows.

Assumption CI-YSF (Partial conditional independence of outcome and name). *For all i ,*

$$Y_i \perp\!\!\!\perp S_i \mid f(S_i), R_i, G_i, X_i.$$

The next theorem shows that it is still possible to nonparametrically identify racial disparities under Assumption CI-YSF under the identification condition, which is only slightly stronger than for Theorem 3.1.

Theorem 3.4 (Nonparametric Identification Under Assumption CI-YSF). *Let $f : \mathcal{S} \rightarrow \mathbb{R}^d$, $d < |\mathcal{S}|$, with range $f(\mathcal{S})$. For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then under Assumption CI-YSF, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, f(S) = z, G = g, X = x)$ are identified if and only if both \mathbf{P} and the augmented matrix $(\mathbf{P} \ \mathbf{b})$ have rank $|\mathcal{R}|$.*

As long as the dimension d of the surname summary statistic $f(S)$ is much smaller than the (usually large) number of surnames $|\mathcal{S}|$, racial disparities are likely to be identified under Theorem 3.4 when they are already identified under Theorem 3.1. Thus, Assumption CI-YSF and Theorem 3.4 can be used in conjunction with carefully chosen f in order to probe likely failure modes of the more restrictive Assumption CI-YS. If estimates are not much affected by the inclusion of $f(S)$, then researchers can be more confident in the plausibility of Assumption CI-YS.

Second, bias can also arise from violations of the assumptions underlying the BISG methodology (Assumptions CI-SG and ACC). Of course, this is not unique to the proposed methodology: violations of these assumptions will also affect the validity of other disparity estimators such as weighting or thresholding. However, since as discussed above the BISG assumptions may rarely hold exactly in practice, we provide in Appendix C several results characterizing how the model’s estimates are affected by bias in the BISG probabilities.

4 Empirical Validation with the Voter File

To better understand how BIRDIE performs in real-world contexts, we apply it to North Carolina voter registration data. This voter file contains individual-level self-reported race for almost all voters and hence the “ground truth” relationship between outcome and race is known. We compare the performance of BIRDIE models against those of the weighting and thresholding estimators. We also evaluate how the estimation error depends on the level of geographic precision used in the BISG probabilities. Finally, we briefly demonstrate various extensions of the BIRDIE methodology: small-area estimates (as discussed in Section 3.2), improved individual race predictions (Section 3.4), estimation conditional on an additional explanatory variable (Section 3.5), and sensitivity analysis for potential assumption violation (Section 3.6).

4.1 North Carolina Voter File

Like most other Southern states, which have a history of disenfranchising minority voters, the state of North Carolina asks (and previously required) every voter to self-report their race upon registration. This data, along with voters' names, addresses, gender, party registration (if any), and voting history, is part of the voter file that the secretary of state makes publicly available. This feature makes the voter file an ideal validation setting for the proposed methodology. The outcome we examine here, party registration, is the product of many unobservable factors, and is known to differ across racial groups. Since self-reported race is available, inferences about these racial disparities using the estimators discussed here can be compared to the corresponding ground truth.

Estimation of party registration by race is of substantive interest as well, especially in the context of the Voting Rights Act of 1965 (VRA). The relationship between these variables is critical for understanding the impact of policy changes such as redistricting or election rules on compliance with the VRA, and for establishing legal standing to challenge these policies under the VRA. As many states do not ask for self-reported race during voter registration, methods like BIRDIE are important tools for evaluating VRA compliance.

We use a subset of the October 2022 voter file which could be linked to a proprietary voter file provided by L2, Inc., a leading national non-partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters, and consultants for use in campaigns. The L2 file geocoded each address to a Census block, which allows for the finest block-level BISG predictions. We also removed any records without individual race information, since our goal is validation compared to some ground truth, rather than inference about the entire population of registered North Carolina voters. Altogether, 22.1% of records either had missing race information or could not be linked to the L2 file.

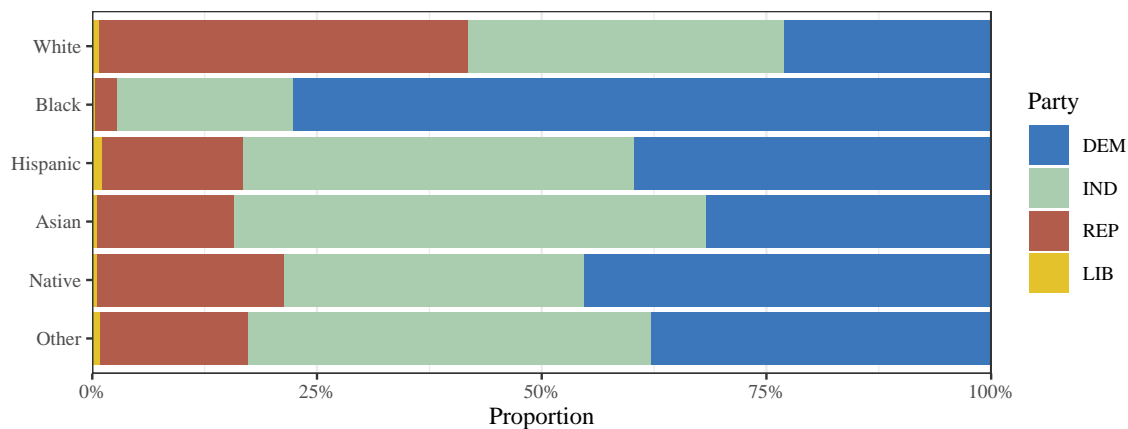


Figure 2: Distribution of party registration by race for a sample of 1,000,000 North Carolina voters. Parties are Libertarian (LIB), Republican (REP), Independent (IND), and Democratic (DEM).

The overall merged voter file contains 5,754,912 voters, 71.1% of which are White, 21.1% of which are Black, and 7.8% of which belong to another race. To reduce computational burden, we further subsampled this file by selecting 1,000,000 records at random without replacement. This sample size is large enough to ensure that sampling error in the estimates is negligibly small.

Figure 2 shows the distribution of party registration by self-reported race in this subsample. White voters disproportionately register as Republicans, while Black voters disproportionately register as Democrats. This serves as the ground-truth in our validation analysis presented below.

4.2 The Model Setup

We first calculate BISG probabilities using 2010 Census data at the census block, tract, ZIP code tabulation area (ZCTA), and county level. Every record in the voter file contains county information, while roughly 13% of records are missing ZIP codes and 27% of records are missing blocks/tracts; when these finer geographic identifiers were missing, we used county-level Census tables in the BISG calculations.

The BISG probabilities are broadly accurate. Using the maximum a posteriori racial category as a prediction, we obtain accuracy of 76.2% for the county probabilities, 78.4% for the ZIP code probabilities, 78.5% for the tract probabilities, and 79.6% for the block probabilities. An alternative measure of the quality of the BISG probabilities is the logarithmic score, a proper scoring rule which rewards precise and calibrated probabilistic estimates (higher values are better). The logarithmic scores for the BISG probabilities are -0.618 for counties, -0.587 for ZIP codes, -0.585 for tracts, and -0.607 for blocks. For comparison, the prior-only logarithmic score (i.e., using no name or geographic information) is -0.867 . The worsened performance for block-level versus tract-level probabilities likely stems from the larger impact of census measurement error at smaller geographies, a problem that could be addressed using newer BISG methods such as those of Imai et al. (2022).

Since the goal of our validation study is to compare BIRDIE estimates with weighting and thresholding estimates, we do not make additional comparisons between BISG probabilities and those generated with alternative racial prediction methods. To the extent competing racial prediction methods improve prediction accuracy, we expect the gap between different disparity estimation methods (weighting, thresholding, BIRDIE) to narrow, consistent with Theorems 2.2 and 3.3. As we have discussed, however, high accuracy of racial prediction alone is neither necessary nor sufficient for accurate estimation of racial disparities. If other racial prediction methods produce increased accuracy at the cost of worsened calibration, accuracy in estimating racial disparities may be poor whether using weighting, thresholding, or BIRDIE estimates.

In our validation, for a given set of BISG probabilities, we estimate the conditional distribution of each outcome variable given race using BIRDIE with both saturated pooling and multinomial mixed-effects models introduced above. We then compare the resulting estimates based on these BIRDIE models against those of the two existing estimators—the weighting estimator as well as a thresholding estimator that deterministically assigns each individual the maximum *a posteriori* racial category. For the saturated BIRDIE model, we use geographic effects matching the geographic level used in the BISG probabilities (e.g., county effects for the county-level BISG probabilities), except for the block-level probabilities. Due to the large number of individual census blocks, we use tract-level effects instead for this particular model. For the mixed-effects BIRDIE model, in addition to these geographic effects we add two geography-level covariates: the White and Black fraction of the population in each individual’s geography of residence. These covariates should help further regularize and share information among the individual geographic effects.

We use noninformative or weakly informative priors for both BIRDIE models. For the saturated model, we use a uniform prior for the Dirichlet hyperparameters. For the mixed model, the prior on the fixed effects β_r for each racial group is a weakly informative Normal with standard deviation $2p_r$, where p_r is the share of the racial group in the sample; this encourages more shrinkage for groups where less data is available. The overall global intercept received a $\mathcal{N}(0, 5^2)$ prior. The prior on the random intercept scale is Inv-Gamma(4, 1.5), designed to support a range of possible heterogeneity of the outcome-race relationship across geographic levels while discouraging the M step from finding a mode at zero; it places 95% of its mass between 0.17 and 1.376. The random intercept correlation matrix (across levels of Y) received an LKJ prior with shape parameter 2. Variations of these prior choices did not noticeably affect the top-line estimates, however, due to the large quantity of data overall. To give an idea of the computational efficiency of the proposed method, the maximum runtime of the saturated of the BIRDIE models fit to estimate party registration was 6.9 seconds (on a modern laptop with 8GB RAM), and the maximum runtime for the mixed model estimates was 22.6 seconds.

4.3 Estimates of Racial Disparity in Party Registration

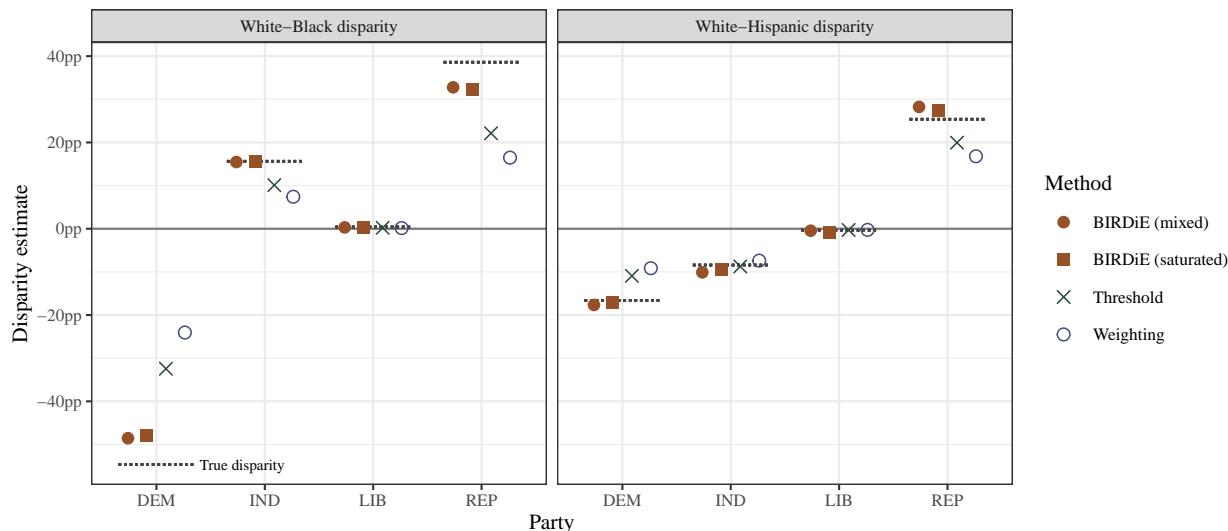


Figure 3: Error in the White-Black and White-Hispanic disparity estimates for party registration, by estimation method. The true disparities are indicated by the dotted lines. All methods used block-level data for this figure; the results for other levels of geographic detail are generally similar. Estimation uncertainty is minimal and hence suppressed from the figure for clarity.

We first examine the relative accuracy of the proposed methods in estimating the disparity between White and Black, and White and Hispanic voters, in party registration. For example, the true difference in Democratic registration between Black and White voters in the sample is 54.6 percentage points, meaning Black voters register Democratic at a much higher rate. However, the standard weighting approach produces an estimate of only 24.1 percentage points for this disparity—less than half the true value. This is consistent with Corollary 2.2.1, which states that the weighting estimator tends to underestimate the magnitude of racial disparity. The thresholding estimator, while slightly better, also misses the mark, with an estimate of 32.5 percentage points. In contrast, the saturated BIRDiE model produces an estimate of 47.9 percentage points, and the mixed BIRDiE model estimates 48.5. These estimates are only slightly lower than the ground truth.

Figure 3 compares the empirical performance of the BIRDiE models against that of the weighting and thresholding estimators across all of these possible disparity measurements, using the block-level BISG predictions. For White-Black (left plot) and White-Hispanic (right plot) disparities in party registration, the BIRDiE models (solid circles and squares) substantially outperform the two commonly used estimators (open circles and crosses). For two major parties, both the weighting and thresholding estimators exhibit a substantial amount of estimation error, for example, exceeding 20 percentage points for the White-Black disparity for the Democratic party. In contrast, the two BIRDiE models yield a much smaller estimation error that ranges within several percentage points for all racial disparity estimates. The saturated and mixed-effects BIRDiE models perform similarly with no discernible difference.

For a more comprehensive look at the error in the estimated partisanship-by-race distributions, we turn to the total variation (TV) distance, which is calculated as

$$d_{TV}(\hat{\mu}_{Y|R}, \mu_{Y|R}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \sum_{r \in \mathcal{R}} |\hat{\mu}_{Y,R}(y, r) - \mu_{Y,R}(y, r)|,$$

where $\mu_{Y,R}$ is the joint distribution of Y and R . The TV distance is an upper bound on the error in *any* probability calculated from the estimated joint distribution, and as such is useful general-purpose measure of estimation

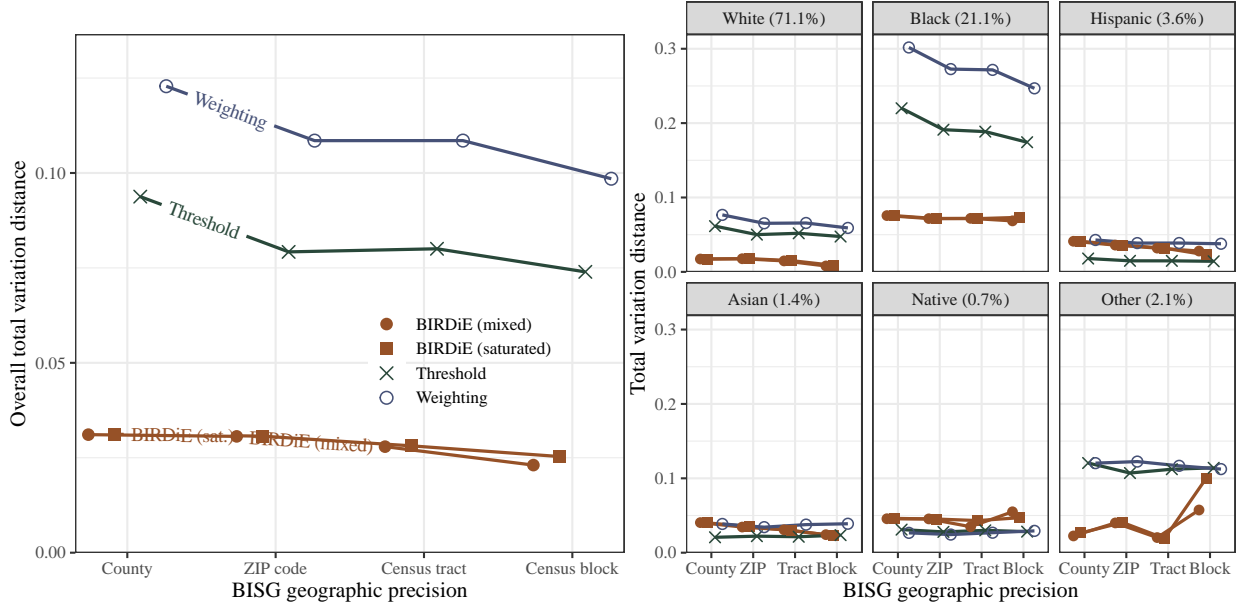


Figure 4: Total variation distance between the estimated and actual distribution of party registration, by estimation method and level of geographic detail used in the BISG predictions. The left plot shows the overall total variation distance, while the right plot decomposes it by racial group.

error. The left plot of Figure 4 shows the TV distance for each estimator, not just for the block-level BISG estimates used in Figure 3 but also across the range of geographic levels used in the BISG calculation (x-axis).

We find that both BIRDiE models substantially outperform every alternative method at every geographic level. In general, the estimates based on the BIRDiE models exhibit a total variation distance whose magnitude is about one third and one fourth of that for the thresholding and weighting estimators, respectively. As before, the saturated and mixed-effects BIRDiE models perform similarly. According to this measure, we find that finer geographic data provide only minor improvements in accuracy for the BIRDiE or conventional estimates. While possibly counterintuitive, this finding underscores the fact that calibrated BISG probabilities, rather than highly precise probabilities, are all that is needed for accurate disparity estimation.

Of course, both calibrated and precise probabilities are to be preferred to imprecise but calibrated probabilities. In practice, however, there may be a tradeoff between the two. For example, including first names in the BISG predictions may increase their precision. But, first names may lead to worse calibration, since BISG methods which use surnames make an somewhat unrealistic conditional independence assumption, and data on first names by race come from non-census sources (Tzioumis, 2018; Rosenman et al., 2023). Additionally, unlike surnames, first names (which are usually chosen by parents) can be more correlated with socioeconomic variables, leading to violations of Assumptions CI-YS.

Lastly, we measure the TV distance for each conditional distribution by race:

$$d_{TV}^{(r)}(\hat{\mu}_{Y|r}, \mu_{Y|r}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} |\hat{\mu}_{Y,r}(y, r) - \mu_{Y,r}(y, r)|.$$

The right plot of Figure 4 shows these within-race TV distances, to illuminate how the estimators perform on each subgroup. In general, the BIRDiE models are more accurate than the weighting and thresholding estimators for the White and Black racial groups which together make up 96% of the sample. All estimators perform roughly equally well for Hispanic, Asian, and Native voters (though the thresholding estimator performs well

for Hispanic voters), exhibiting relatively small estimation error. The weighting and thresholding estimators perform particularly poorly for Black voters and for the “Other” racial group.

4.4 Small-area Estimation

An advantage of the BIRD*i*E methodology is its explicit modeling of $\mathbb{P}(Y | R, G, X)$, which produces not only estimates of the marginal $\mathbb{P}(Y | R)$ but also subgroup estimates of how these conditional distributions vary across covariates and geographic areas. This section examines the accuracy of the saturated and mixed-effects BIRD*i*E models in recovering small-area relationships between party registration and race, compared to standard methodology that simply applies the weighting and thresholding estimators within each geographic area. We study accuracy at the county, ZIP code, and tract level, using the BISG probabilities and BIRD*i*E models that were applied to each. Since in fitting the BIRD*i*E model to block-level BISG probabilities we used tract-level random intercepts, we do not present block-level estimates.

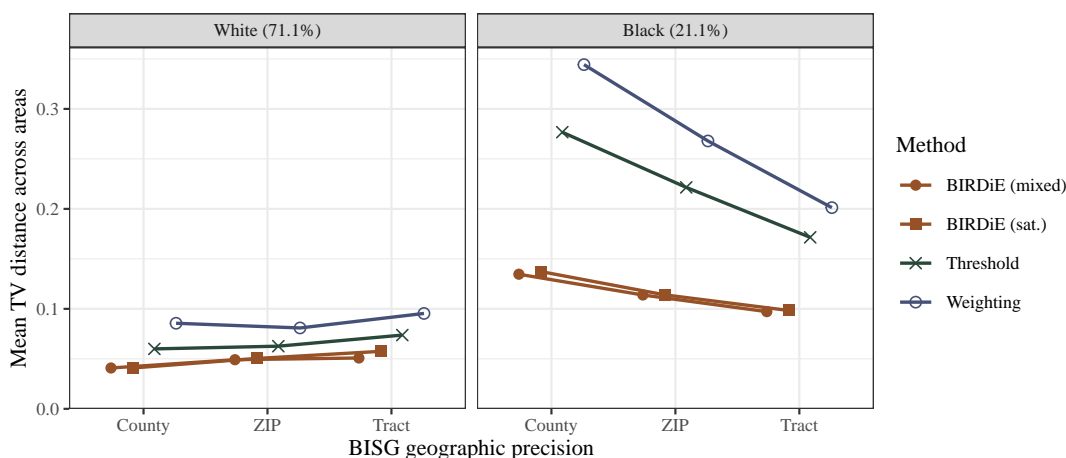


Figure 5: Accuracy of small-area estimates by race, as measured by the average total variation distance.

We evaluate the small-area estimates by calculating the mean total variation distance between the estimated and true conditional distributions of party registration by race (averaging across geographic areas). Figure 5 summarizes our results, which qualitatively track the patterns found overall in Figure 4. The two BIRD*i*E models exhibit substantially lower error than the weighting and thresholding estimators. Across all methods, the error is lower for White voters, who make up the bulk of the sample. Somewhat surprisingly, the amount of error does not appear to vary much for the BIRD*i*E models across different levels of geography—tract-level estimates are roughly as accurate as county-level estimates, on average.

Between the BIRD*i*E models, the mixed-effects model slightly outperforms the saturated model. This reflects the value in partially pooling estimates through the random effect structure.

4.5 Improved Individual Race Probabilities

As discussed in Section 3.4, we can use the conditional distribution $\mathbb{P}(Y | R, X, G)$ estimated with a BIRD*i*E model to create model-updated BISG probabilities $\hat{\mathbf{P}}_{|Y} = \pi(\mathbf{R} | \hat{\Theta}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ by applying Bayes’ rule. These updated probabilities may be more accurate than the original BISG probabilities.

For example, using the estimates produced by the mixed BIRD*i*E model applied to party registration with block-level BISG estimates, the MAP prediction accuracy increases from 79.6% with the input probabilities to 82.9% with the updated probabilities. These increases are significantly larger than the differences in accuracy between BISG probabilities using different levels of geographic precision.

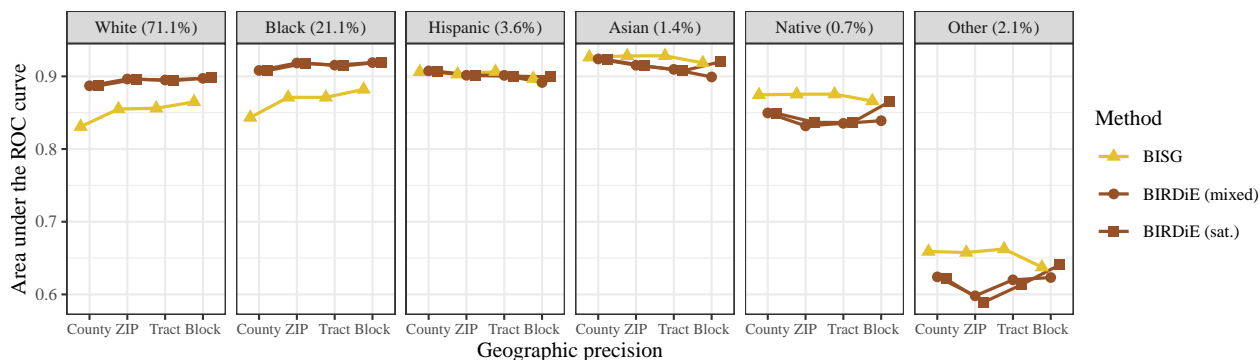


Figure 6: Race probability predictive accuracy, as measured by the area under the receiver operating characteristic (ROC) curve, for the input BISG probabilities as well as the BIRDIE-updated probabilities, by race and level of geographic precision. Larger values indicate more precise predictions.

The improvements are reflected in other accuracy measures as well. Figure 6 shows the accuracy of the predictions by race, as measured by the area under the receiver operating characteristic (ROC) curve. The updated probabilities are noticeably more accurate than the input probabilities for White and Black voters, about as accurate for Hispanic and Asian voters, and slightly less accurate for Native and “Other” voters for some geographic levels.

4.6 Estimates Conditional on an Additional Variable

The North Carolina voter file also provides an opportunity to demonstrate the methodology described in Section 3.5 to produce estimates conditional on another predictor variable that is not used in the BISG probabilities. We will estimate party registration rates by race among voters and nonvoters in the 2020 election. Following the discussion in Section 3.5, we will compute these estimates two ways: (1) by estimating party registration and 2020 turnout jointly by race, and (2) by first estimating 2020 turnout by race, then estimating party registration by race and 2020 turnout.

We will use a multinomial mixed-effects BIRDIE model applied to the block-level BISG probabilities, with random intercepts by tracts, for all the estimation. Fitting this model to a combined 2020 turnout/party variable (i.e., one with eight levels: no/DEM, yes/DEM, and so on) produces estimates of the joint distribution of party registration and turnout in 2020 by race. Normalizing these probabilities within turnout and race groups produces estimates of party registration by race and turnout. The total variation distance between these estimates and the true distribution is 0.051, which indicates close agreement. The accuracy of the most-likely race predictions from the BISG probabilities updated with both party and turnout is about the same as the party-alone accuracy, indicating that turnout is perhaps less correlated with race after controlling for location and party registration.

Figure 7 presents the errors in the estimates for this method (solid square), the two-step method (solid circle), and for the weighting (open circle) and thresholding (cross) estimators. The TV distance between these two-stage estimates and the true distribution is 0.052, very similar to the error in the joint-estimation approach. The two-step approach produces similar estimates to the joint-estimation approach, though the latter performs better in the “Other” category. As discussed in Section 3.5, while both approaches produce highly accurate estimates in this example, we would expect the two-stage approach to be superior when one or both of the variables has more levels.

In contrast to the BIRDIE models, the weighting and thresholding estimates of party registration by 2020 turnout and race include large errors, especially for Black voters with all the errors exceeding 10 percentage

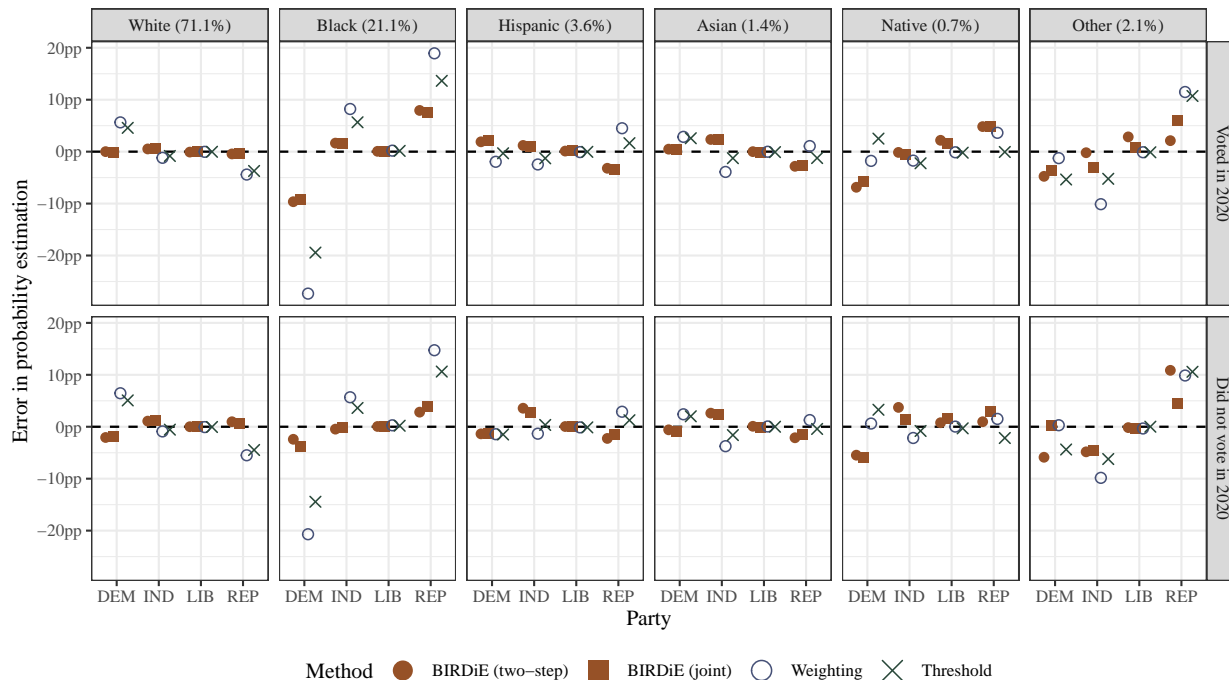


Figure 7: Error in the conditional expectation estimates for party registration by turnout and race, by estimation method. All methods used block-level data for this figure. Estimation uncertainty is minimal and hence suppressed from the figure for clarity.

points. The TV distance for the weighting estimator is 0.196, and the distance for the thresholding estimator is 0.147—around 3–4 times higher than for the estimates based on the BIRDiE models.

4.7 Sensitivity Analysis

Finally, we examine the sensitivity of our party registration estimates to potential violations of the key identifying Assumption CI-YS, following the method outlined in Section 3.6 that is based on a low-dimensional summary statistic of surnames. We use a publicly available sample of 5% of the individual records for the 1930 Census (Ruggles et al., 2021), which contains individual names, individual and parental birthplace, and detailed race, ethnicity, and tribal codes. Since many regions of Asia, particularly Vietnam, experienced little emigration to the United States before 1930, we further supplement this data with around 3,000 Asian surnames classified into six regional subgroups: Chinese, Filipino, Indian, Japanese, Korean, Vietnamese, NHPI, and Other (Guage et al., 2023).

Using these subgroups and the 1930 birthplace and racial data, we can classify most surnames in the voter file into nine groups (see Appendix D for a brief description of the groupings and the most common 50 surnames for each group). While somewhat arbitrary, these groups are chosen to combine countries of origin which had significant immigration to the United States during similar periods.

We first evaluate the plausibility of Assumption CI-YS by examining the correlation between the residuals of the BIRDiE model fit and indicator variables for each of the nine surname groups. Under Assumption CI-YS, this residual correlation should be zero everywhere. As Figure 8 shows, however, for many groups and party labels, the correlation is small but deviates from zero more than would be expected given only sampling variation. Here, we use the residuals from the county-level saturated model specification, but the results are not sensitive to this choice.

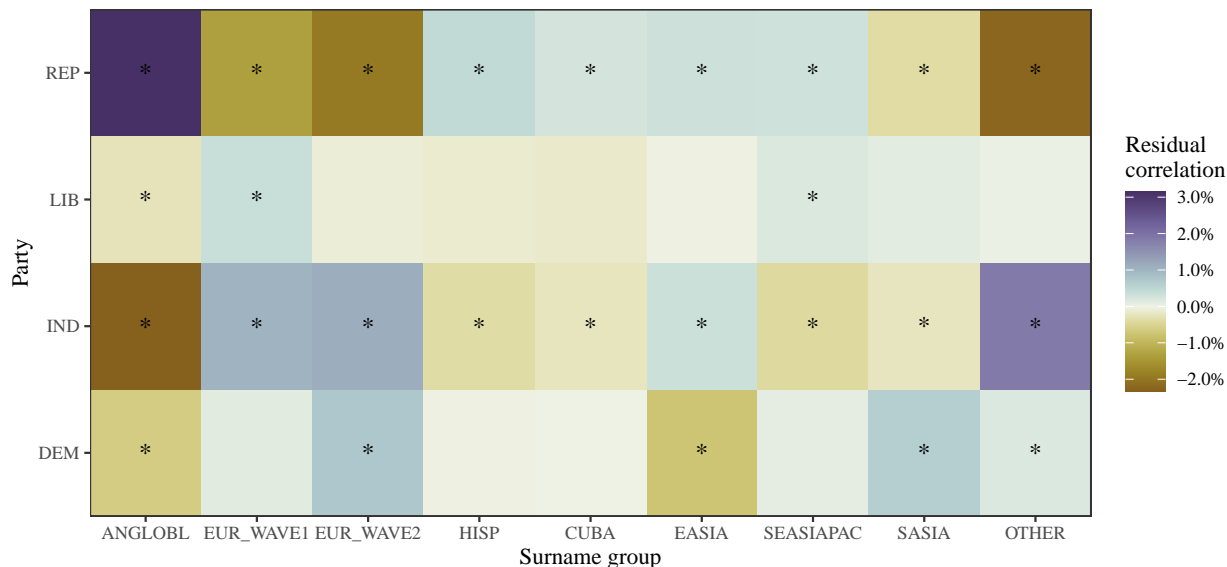


Figure 8: Residual correlation between party registration and nine surname groups, after controlling for race and location. Correlations whose 90% confidence intervals exclude zero are marked with an asterisk. See Appendix D for details on the surname groups.

Notably, voters with names in the Anglosphere and Black surname group, which includes surnames that are relatively more common among many-generation residents of the U.S., such as Smith, Williams, and Brown, are significantly less likely to register as Democrats and independents, and more likely to identify as Republicans, even after controlling for race and geography. Meanwhile, voters with names in the First and Second wave European immigration surname groups, which include surnames more common among 19th and 20th century immigrants from Europe, display the opposite pattern. Differences among surname groups designed to correlate with membership in various Asian subgroups are also visible. As might be expected, the relatively many significant correlations are indicative more of the large number of observations in the data, rather than large residual correlations themselves—all of the correlations are quite small in magnitude, with most on the order of 0.01 or so. Thus, we expect our party-by-race estimates to be little affected by the inclusion of the surname group indicators.

Indeed, re-fitting the county-level saturated model with an additional surname group covariate produces nearly identical results, albeit at a moderately higher computational cost given the increased number of $(G, X, f(S))$ cells. This re-fitting requires Assumption CI-YSF, which relaxes Assumption CI-YS. We find that the average party registration rate estimate changes only by 0.010, with the largest change being 0.036, for the rate of Republican registration among Other voters. The accuracy of the updated BISG probabilities is likewise virtually unchanged. All in all, this analysis provides confidence that violations of Assumption CI-YS for the North Carolina voter file are likely minor and would have minimal effects on quantitative and qualitative conclusions.

5 Application to Tax Data

We further demonstrate the use of the proposed methodology by applying it to study racial differences in the home mortgage interest deduction (HMID) claims, using individual-level tax data from the U.S. Internal Revenue Service (IRS). This data includes the universe of income tax returns filed by U.S. taxpayers. The IRS does not collect, and the data does not include, information on taxpayer race or ethnicity.

5.1 Home Mortgage Interest Deduction (HMID)

The HMID is designed to incentivize home ownership; homeowners with a mortgage qualify for an itemized deduction based on the amount of mortgage interest they pay during the year. The deduction is only available for taxpayers who itemize their deductions. Following increases to the standard deduction and other tax code changes as part of the “Tax Cuts and Jobs Act” of 2017 (P.L. 115-97), roughly 90% of taxpayers take the standard deduction and do not itemize. These taxpayers are unable to take advantage of the HMID.

The Treasury Department estimates the HMID costs the government about \$25 billion in foregone revenue in 2019. By budgetary cost, the deduction is the largest in the income tax code ([Congressional Research Service, 2017](#)). Because it is only available to homeowners, the HMID may disproportionately benefit taxpayers of racial groups that have a high homeownership rate. Prominent legal scholars have criticized the possible disproportionate benefits, with [Brown \(2022, p. 94\)](#) referring to the subsidy from the HMID as “little more than the twenty-first-century version of redlining” and concluding it “must be repealed.” On the other hand, it is also possible that if Black homeowners faced higher mortgage rates, they could in principle benefit more than would be expected based on homeownership rates alone. Lack of administrative data on HMID claims by race made it difficult to quantify racial disparities that potentially exist.

The Treasury’s internal Office of Tax Analysis has recently used an extension of the standard BISG model to estimate the usage the HMID and other deductions by race from individual-level data ([Cronin et al., 2023](#)). External researchers have also studied HMID usage by analyzing survey data or data on proxies like home ownership ([Sullivan et al., 2017](#)). Both types of analyses have found that White taxpayers benefit disproportionately from the HMID, though the magnitude of the disparity is unclear, especially given the methodological challenges identified in this paper. Here, we hope to use the BIRD*i*E methodology to provide a more precise answer to the question of which groups are using the HMID and how much they benefit from it.

5.2 Estimation Procedure

We use a random 10% sample of individual tax returns (Form 1040s) filed with the IRS for tax year 2019, a total of 17,145,898 observations. To calculate individual race probabilities for every observation, we use the ZIP code tabulation area (ZCTA) corresponding to the geocoded address listed on the return, plus the last name of the primary filer using a standard BISG model. This means that conclusions about racial groups here refer to the race of the primary filer, and not the race of other household members. For the roughly 3.4 million records for which geocoding was not successful, only last names were used.

The outcome variable is the amount of the HMID claimed by the filer, discretized into 11 levels: one for a deduction of \$0, capturing roughly 90% of the sample, and ten levels corresponding to the deciles of the HMID among those taking the deduction. Given the size of the data, our outcome model is the relatively simple no-pooling model for HMID level by geography and racial group. We coarsen the geography variable used for modeling to the Public Use Microdata Area (PUMA) level.

Using a larger geography like PUMAs rather than ZCTAs is necessary to ensure a reasonable amount of data in each outcome-race-geography cell of the outcome model. PUMAs partition each state into areas containing roughly 100,000 people. Compared to alternative units of analysis like states or counties, PUMAs are therefore adaptive to population density; a large city might be contained entirely in a single county, but with a dozen or more distinct PUMAs, while the surrounding rural areas might be spread over many counties but only a few PUMAs. Since we expect more geographic heterogeneity in and around cities, this feature of PUMAs lends itself well to our analysis. Where PUMA was not available for the 3.4 million missing geocodes, we used an indicator for state of residence instead, which is available for every record in the sample.

There are 1,961 distinct PUMAs (or states) in the sample, compared to 28,880 ZCTAs. With 10 non-zero outcome levels and 6 racial groups, the typical cell in the outcome model is expected to have around 14 observations.

Smaller racial groups would have fewer observations than this, on average. Therefore, to help regularize the PUMA-level model estimates, we impose a weak empirical Bayes prior based on a global estimate of HMID by race from the simple weighted estimator. The effective data size of the prior is just 0.1 observations per racial group. Thus the prior will have a meaningful impact only in those areas where there are close to zero expected members of a racial group, according to the BISG probabilities.

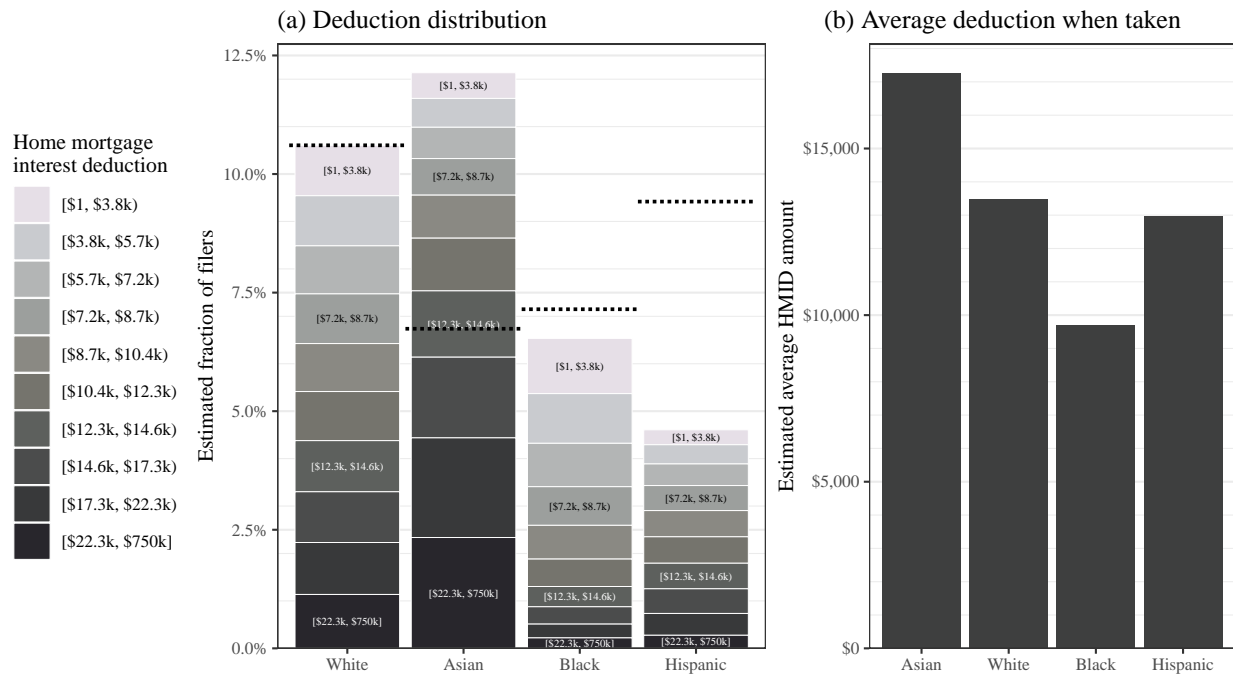


Figure 9: Estimates of usage of the home mortgage interest deduction by race. Panel (a) shows estimates of the average size of the deduction among filers who took the deduction at all. Panel (b) shows estimates of the proportion of filers who took the deduction, further broken into deciles by deduction amount. The dotted lines mark the expected height of each bar if the only disparities were those in mortgage rates between racial groups.

5.3 Findings

Figure 9(a) shows the estimated proportion of filers in each racial group which take the HMID at all, and the distribution of the HMID claim amount among those who do. Racial disparities are immediately apparent: while 10.6% of White filers take the HMID, just 6.5% of Black filers and 4.6% of Hispanic filers do. In contrast, roughly 12% of Asian filers take the deduction.

How much of this disparity is explained by differences in home ownership rates across racial groups? We develop estimates of the fraction of each racial group that has a mortgage for the home they live in based on data from the 2010 decennial census (see Appendix E for details). We then plot as dashed lines in Figure 9(a) estimates of the fraction of each racial group that would claim the HMID if every filer with a mortgage claimed the deduction at the same rate as White filers do. The figure shows that the lower share of Black taxpayers claiming the HMID can be explained by the lower home ownership rate among that group. However, for other groups, disparities remain after controlling for having a mortgage. In particular, Hispanic filers claim the HMID at a 2.5 percentage point lower rate than would be expected based on mortgage rates alone, and Asian filers claim the HMID at a higher rate than their share of the population with mortgages would imply. These results suggest that closing disparities in home ownership may not be sufficient to eliminate disparities in who benefits from the HMID, as evidently other aspects of filers’ situations beyond home ownership, such as eligibility for other itemized deductions, are affecting whether their HMID benefits.

Beyond disparities in the rate that taxpayers claim any HMID, there are also racial differences in the amount of the HMID among claimants. This is apparent in both the distribution across HMID deciles in Figure 9(a) as well as in Figure 9(b), which displays estimates of the mean HMID amount among filers who claim the deduction. These estimates were produced by weighting the observed HMID amounts according to BIRDIE-updated race probabilities, as described in Section 3.4. Compared to White claimants, whose average deduction is \$13,500, HMID amounts for Black claimants are skewed toward the lower deciles, translating to a \$2,100 to \$3,800 lower average HMID amount for these groups. In contrast, Hispanic claimants deduct just \$500 less than White claimants, and Asian claimants deduct \$3,800 more, on average. Asian filers' deductions fall into the highest deciles at more than double the rate of any other group. In fact, a higher fraction of Asian filers take at least a \$17,000 deduction than the fraction of Hispanic filers who take any deduction at all. Table 1 compares these estimates for claimants to the unconditional averages of the HMID benefit amount across racial groups.

Overall, our findings support the claims of researchers such as [Moran and Whitford \(1996\)](#) and [Brown \(2022\)](#) and others that the HMID is disproportionately unavailable to Black and Hispanic taxpayers. Our estimates show that the picture is complicated further by differences between racial groups even accounting for the prevalence of mortgages. In addition, the pattern of disparities in overall HMID claims looks different from the disparities in the amount of the HMID among claimants.

6 Discussion

We have introduced a new identifying assumption and accompanying model, BIRDIE, and clarified other assumptions implicit in approaches to disparity estimation when individual race is not observed. In many real-world applications, we believe that the new model and identification condition are appropriate and will produce significantly improved estimates. However, there is no one-size-fits-all approach for the estimation of racial disparities. For example, the existence of name-based discrimination may violate our identification assumption especially when racial categories, for which data are available, are coarse. Although we provide a sensitivity analysis that partially addresses this concern, careful consideration of the underlying causal and information structure is required to avoid making the incorrect conclusions.

As our empirical demonstration shows, in realistic settings BIRDIE can substantially outperform existing estimators of racial disparities, both in aggregate and for small areas. The BIRDIE methodology also produces improved BISG probabilities, and can be used to estimate disparities conditional on other variables. These additional features should prove helpful in practical settings.

6.1 Recommendations for Practitioners

Given the large amount of missing data inherent to the type of racial disparity estimation studied here, choices about data selection, processing, and modeling can have a significant impact on estimates and substantive conclusions. We collect here several recommendations for practitioners using BISG and BIRDIE methodology in their research.

Table 1: Estimates of the HMID claim rate and the average deduction by racial group.

Average claim	White	Black	Hispanic	Asian
Rate	10.6%	6.5%	4.6%	12%
Among claimants	\$13,500	\$9,700	\$13,000	\$17,300
Unconditionally	\$1,400	\$600	\$600	\$2,100

- **Choose estimation methodology and input data based on the specifics of the research question.** As we have stressed, the causal structure of the research setting determines whether BISG should be used with the weighting or BIRDIE estimators. Research on populations which are very different from the general U.S. population may also benefit from the use of alternative or additional data on race and geography specific to that population. The choice of whether to use state-, county-, ZIP-, tract-, or block-level data likewise depends on aspects of the data under study, and the scale of geographic variation in the outcome variable and quantities of interest.
- **Focus on the calibration, not the predictive accuracy, of BISG probabilities.** Traditionally, BISG-type methods have been evaluated by their predictive accuracy, measured by the mean agreement between the thresholded racial categories and ground truth, or the AUROC of the BISG probabilities. We find that, especially in cases with abundant individual-level data, far more important than maximizing predictive accuracy is ensuring that the BISG probabilities are properly calibrated. Accurate but biased racial prediction will lead to bias in downstream estimates, regardless of which disparity estimator is used. To evaluate probabilistic calibration in validation settings, we recommend visual diagnostics like binned residual plots, as well as numerical summaries like the logarithmic score.
- **Decide whether additional covariates need to be collected.** Additional covariates can make the BISG and BIRDIE assumptions more plausible, and can increase the accuracy of BISG probabilities and the precision of BIRDIE estimates. We recommend prioritizing covariates which are highly predictive of both outcome and race. However, including all available covariates without considering their effect on the various BISG and BIRDIE assumptions is likely to cause problems. In particular, if a covariate's distribution against both race and geography is not known, avoid making unrealistic independence assumptions that are required for the inclusion of the covariate in the BISG model. Rather, follow the approach outlined in Section 3.5. Covariate choice should be driven by substantive considerations, not convenience.
- **Unless computational limits are severe, use the mixed-effects BIRDIE model with group-level covariates.** The mixed-effects model uses the data itself to determine how much to pool disparity estimates across different geographies. This avoids the dual pitfalls of Simpson's paradox (a risk if no geographical information is used) and over-regularization (caused by the prior if no pooling is performed). When using the mixed-effects model, group-level covariates are likely to improve both overall and especially small-area accuracy, and they are easy to collect from public sources. For example, a good default for a ZIP-code level model applied to a political outcome variable would be to include the percentages of the major racial groups in the ZIP code, as well as the ZIP code's income level, population density, and partisanship, in the model.
- **Perform sensitivity analyses.** Both the BISG and BIRDIE models use priors, which should be perturbed to examine the sensitivity of the results to prior selection. The sensitivity of BIRDIE to its key identifying assumption should also be assessed, at minimum using the auxiliary covariate approach demonstrated in Section 4.7.
- **Consider validating estimates with a small-scale survey.** Even with administrative microdata observations in the millions, there is no substitute for a high-quality random sample of individuals for whom race can be observed. Such a sample can be used to validate the various assumptions made by BISG and BIRDIE, as well as providing a sanity check against BIRDIE disparity estimates. Future improvements to BIRDIE could also directly integrate survey data into the workflow.

6.2 Ethical Considerations

As researchers have increasingly applied racial prediction and imputation methods to administrative records, including many publicly available datasets, there has been growing concern about ethical and privacy considerations around performing such predictions. While some scholars do not view racial prediction methods as privacy risks (Bun et al., 2021, “Statistical Inference is Not a Privacy Violation”), we believe it is important for researchers to consider the implications of their use of racial prediction methods (Kenny et al., 2023b).

Recently, Lee and Velez (2023) studied public perception of these ethical considerations through a large factorial survey experiment that asked participants if they viewed a hypothetical study as ethically permissible based on three study factors. They find that studies which focused on accurate estimation of racial disparities were viewed as more ethically permissible than studies that overestimated or underestimated the size of disparities. This tracks with arguments made by many scholars of progressive tax policy (Brown, 2022; Bearer-Friend, 2019).

Compared to previous approaches to racial disparity estimation, which emphasized maximizing accuracy of individual racial predictions to minimize measurement error, BIRD_iE focuses on the accuracy of estimating racial disparities. In fact, as our validation study demonstrated, different sets of individual race predictions of varying individual accuracies (county-based versus block-based BISG), when used with BIRD_iE, produced similar estimates of racial disparities. To the extent that it allows researchers to focus on calibration of racial prediction rather than maximizing individual predictive accuracy, BIRD_iE may alleviate some privacy concerns and reduce incentives to collect and link more personal data in an attempt to further increase accuracy. We view this as a welcome change, consistent with the public’s preference for focusing on accurate disparity information.

However, BIRD_iE does allow for the creation of improved BISG probabilities that incorporate the outcome variable and thus can be more accurate as well as better calibrated. While this accuracy gain is a purely statistical phenomenon based on variables already present in the individual dataset under study, researchers should be cautious, for example, in releasing these improved racial predictions publicly. We urge practitioners to view racial prediction tools as a means to the end of accurate disparity estimation, and treat the intermediate probabilistic predictions with appropriate care.

6.3 Future Research

Much work remains to be done in accurately and reliably measuring racial disparities. First, the BISG probabilities themselves can be improved. Approaches to doing so include Imai et al. (2022), which accounts for some Census measurement error while remaining computationally tractable, and Greengard and Gelman (2023), which rakes BISG margins to improve calibration. More work on identifying and producing data sources which can be used as BISG inputs, rather than relying solely on Census tabulations, will also pay dividends for BISG quality.

Beyond the BISG probabilities, further empirical analyses could determine useful additional variables to condition on, which could allow analysts to weaken the required assumption. Additional study, possibly combined with qualitative research, could identify causal pathways that might threaten the assumptions that BISG and BIRD_iE rely on, and develop data sources, like our auxiliary 1930 Census data, that could be used to evaluate the plausibility of those assumptions in real-world analyses, and their effect on numerical conclusions. Finally, the BIRD_iE model could also be extended to directly model more complex types of outcome variables.

References

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022). The 2020 Cen-

- sus Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, (Special Issue 2). <https://hdsr.mitpress.mit.edu/pub/7evz361i>.
- Alao, R., Bogen, M., Miao, J., Mironov, I., and Tannen, J. (2021). How Meta is working to assess fairness in relation to race in the US across its products and systems. Technical report, Technical Report. Meta AI.
- Anderson, M. and Fienberg, S. E. (1999). *Who Counts?: The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation.
- Argyle, L. and Barber, M. (2022). Misclassification and bias in predictions of individual ethnicity from administrative records.
- Åslund, O. and Skans, O. N. (2012). Do anonymous job application procedures level the playing field? *ILR Review*, 65(1):82–107.
- Barreto, M. A. (2007). ¡Sí se puede! Latino candidates and the mobilization of Latino voters. *American Political Science Review*, 101(3):425–441.
- Bearer-Friend, J. (2019). Should the irs know your race? the challenge of colorblind tax data. *Tax L. Rev.*, 73:1.
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.
- Brown, D. A. (2022). *The whiteness of wealth: How the tax system impoverishes Black Americans—and how we can fix it*. Crown.
- Budiman, A., Cilluffo, A., and Ruiz, N. G. (2019). Key facts about Asian origin groups in the US. *Washington, DC: Pew Research Center*.
- Bun, M., Desfontaines, D., Dwork, C., Naor, M., Nissim, K., Roth, A., Smith, A., Steinke, T., Ullmanand, J., and Vadhan, S. (2021). Statistical inference is not a privacy violation.
- Buttice, M. K. and Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Political analysis*, 21(4).
- Chen, J. (2018). Fair lending needs explainable models for responsible recommendation. *arXiv preprint arXiv:1809.04684*.
- Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348.
- Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J., and Ho, D. (2023). How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 667–686.
- Cho, W. T. and Manski, C. F. (2008). Oxford handbook of political methodology. chapter Cross-Level/Ecological Inference, pages 547–569. Oxford University Press.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- Congressional Research Service (2017). Tax deductions for individuals: A summary. Technical Report R42872.

- Conway, D. A. and Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination. *Journal of Business & Economic Statistics*, 1(1):75–85.
- Cronin, J. A., DeFilippes, P., and Fisher, R. (2023). Tax expenditures by race and hispanic ethnicity: An application of the U.S. Treasury Department’s race and Hispanic ethnicity imputation. Technical Report 122, Office of Tax Analysis, U.S. Department of the Treasury.
- Decter-Frain, A. (2022). How should we proxy for race/ethnicity? Comparing Bayesian Improved Surname Geocoding to machine learning methods. *arXiv preprint arXiv:2206.14583*.
- DeLuca, K. and Curiel, J. A. (2022). Validating the applicability of Bayesian inference with surname and geocoding to congressional redistricting. *Political Analysis*, pages 1–7.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaa05580.
- Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43(5p1):1722–1736.
- Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., and Goldin, J. (2023). Measuring and mitigating racial disparities in tax audits.
- Fiscella, K. and Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research*, 41(4p1):1482–1500.
- Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.
- Gay, C. (2001). The effect of black congressional representation on political participation. *American Political Science Review*, 95(3):589–602.
- Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational Bayes. *Journal of machine learning research*, 19(51).
- Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review*.
- Greene, W. H. (1984). Reverse regression: The algebra of discrimination. *Journal of Business & Economic Statistics*, 2(2):117–120.
- Greengard, P. and Gelman, A. (2023). BISG: When inferring race or ethnicity, does it matter that people often live near their relatives?
- Greiner, J. D. and Quinn, K. M. (2009). $R \times C$ ecological inference: Bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):67–81.
- Guage, C., Goldin, J. S., Ouyang, D., and Ho, D. E. (2023). Enabling data disaggregation of Asian subgroups: A dataset of Wikipedia names. *Working Paper*.
- Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

- Hajnal, Z. and Trounstine, J. (2005). Where turnout matters: The consequences of uneven turnout in city politics. *The Journal of Politics*, 67(2):515–535.
- Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.
- Imai, K., Lu, Y., and Strauss, A. (2008). Bayesian and likelihood inference for 2×2 ecological tables: an incomplete-data approach. *Political Analysis*, 16(1):41–69.
- Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully Bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):1–10.
- Kallus, N., Mao, X., and Zhou, A. (2021). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*.
- Kenny, C., McCartan, C., Kuriwaki, S., Simko, T., and Imai, K. (2023a). Evaluating bias and noise induced by the u.s. census bureau’s privacy protection methods. *arXiv preprint arXiv:2306.07521*.
- Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E., Simko, T., and Imai, K. (2023b). Comment: The essential role of policy evaluation for the 2020 census disclosure avoidance system. *Harvard Data Science Review*, Special Issue 2: Differential Privacy for the 2020 U.S. Census:1–16.
- Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T., Simko, T., and Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US census. *Science Advances*, 7(41):eabk3283.
- Kermani, A. and Wong, F. (2021). Racial disparities in housing returns. Technical report, National Bureau of Economic Research.
- King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.
- King, G., Tanner, M. A., and Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge University Press.
- Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1):419–441.
- Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.
- Laird, N. (1993). The EM algorithm. In *Computational Statistics*, volume 9 of *Handbook of Statistics*, pages 509–520. Elsevier.
- Lee, D. D. I. and Velez, Y. (2023). Ethical use of administrative data in inequality research.
- Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.
- Moran, B. I. and Whitford, W. (1996). A black critique of the internal revenue code. *Wis. L. Rev.*, page 751.
- Park, J., Malachi, E., Sternin, O., and Tevet, R. (2009). Subtle bias against Muslim job applicants in personnel decisions. *Journal of Applied Social Psychology*, 39(9):2174–2190.

- Rosenman, E. T., Olivella, S., and Imai, K. (2023). Race and ethnicity data for first, middle, and last names. *Scientific Data*, 10(299):1–11.
- Ruggles, S., Fitch, C. A., Goeken, R., Hacker, J. D., Nelson, M. A., Roberts, E., Schouweiler, M., and Sobek, M. (2021). Ipums ancestry full count data: 1930 5
- Strmic-Pawl, H. V., Jackson, B. A., and Garner, S. (2018). Race counts: Racial and ethnic data on the U.S. census and the implications for tracking inequality. *Sociology of Race and Ethnicity*, 4(1):1–13.
- Sullivan, L., Meschede, T., Shapiro, T., and Fernanda Escobar, M. (2017). Misdirected investments: How the mortgage interest deduction drives inequality and the racial wealth gap. Technical report, Institute on Assets and Social Policy and the National Low Income Housing Coalition.
- Tzioumis, K. (2018). Demographic aspects of first names. *Scientific Data*, 5(1):1–9.
- U.S. Census Bureau (2014). Frequently occurring surnames from the 2010 census. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. Accessed: 2022-03-25.
- U.S. Census Bureau (2022). Census Bureau releases estimates of undercount and overcount in the 2020 census. *March 10, 2022, Release Number CB22-CN.02*.
- Van Ryn, M. and Fu, S. S. (2003). Paved with good intentions: do public health and human service providers contribute to racial/ethnic disparities in health? *American journal of public health*, 93(2):248–255.
- Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.
- Voicu, I. (2018). Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5(1):1–13.
- Wakefield, J. (2004). Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–425.
- Williams, D. R. and Jackson, P. B. (2005). Social sources of racial disparities in health. *Health affairs*, 24(2):325–334.
- Zest AI (2020). Zest race predictor (zrp). Available at <https://github.com/zestai/zrp/>.
- Zhang, Y. (2018). Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197.

A Proofs of Propositions

Corollary 2.2.1 (Underestimation of racial disparity). *Let $y \in \mathcal{Y}$. If race is binary (so $\mathcal{R} = \{0, 1\}$), and $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$, then*

$$\hat{\mu}_{Y|R}^{(wtd)}(y \mid 1) - \hat{\mu}_{Y|R}^{(wtd)}(y \mid 0) < \mathbb{P}(Y = y \mid R = 1) - \mathbb{P}(Y = y \mid R = 0).$$

Proof. For notational simplicity, let $\mu_r = \mathbb{P}(Y = y \mid R = r)$ and $\hat{\mu}_r = \hat{\mu}_{Y|R}^{(wtd)}(y \mid r)$ for $r \in \{0, 1\}$. Since $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, necessarily $\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = 1\} \mid G, X, S)] > 0$ and $\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = 1\} \mid G, X, S)] < 0$. We note that the corollary could be stated under this more general condition, but was not for expositional clarity. Thus by Theorem 2.2, $\hat{\mu}_1 - \mu_r < 0$ and $\hat{\mu}_1 - \mu_r > 0$. Then

$$\begin{aligned} \hat{\mu}_1 - \hat{\mu}_0 &= \hat{\mu}_1 - \mu_1 + \mu_1 - \mu_0 + \mu_0 - \hat{\mu}_0 \\ &= (\hat{\mu}_1 - \mu_1) - (\hat{\mu}_0 - \mu_0) + (\mu_1 - \mu_0) \\ &< \mu_1 - \mu_0, \end{aligned}$$

as claimed. \square

Theorem 3.1 (Nonparametric Identification). *For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then, under Assumption CI-YS, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, G = g, X = x)$ are identified if and only if both \mathbf{P} and the augmented matrix $(\mathbf{P} \ \mathbf{b})$ have rank $|\mathcal{R}|$.*

Proof. Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y \mid R, G, X$, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned} \mathbb{P}(Y = y \mid G = g, X = x, S = s) &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x, S = s) \mathbb{P}(R = r \mid G = g, X = x, S = s) \\ &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x) \mathbb{P}(R = r \mid G = g, X = x, S = s). \end{aligned}$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, G = g, X = x)$. These parameters are identified if and only if this system has a unique solution, i.e. if the coefficient matrix \mathbf{P} has rank $|\mathcal{R}|$ and so does the augmented matrix $(\mathbf{P} \ \mathbf{b})$. \square

Theorem 3.2 (Unbiasedness of OLS Estimator). *If Assumptions CI-SG, ACC, and CI-YS hold, and the identification conditions in Theorem 3.1 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,*

$$\mathbb{E}[\hat{\mu}_{Y|R}^{(p-ols)}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

Proof. Fix $y \in \mathcal{Y}$ and define $m_{gxr} = \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x]$. Then under Assumptions CI-SG, ACC, and CI-YS,

$$\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, S = s]$$

$$\begin{aligned}
&= \sum_{r \in \mathcal{R}} \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x] \mathbb{P}(R = r \mid G = g, X = x, S = s) \\
&= \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r,
\end{aligned}$$

where as in the main text $\hat{\mathbf{p}}$ is the (random) vector of BISG probabilities. In fact, since the right-hand side depends on S only through $\hat{\mathbf{p}}$, we have

$$\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, \hat{\mathbf{p}}] = \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r.$$

So the conditional expectation of $\mathbf{1}\{Y = y\}$ given X, G , and the BISG probabilities $\hat{\mathbf{p}}$ is linear in those probabilities, with coefficients m_{gxr} . Consequently, the OLS estimate $\hat{\mu}_{Y|RGX}^{(\text{ols})}(y \mid \cdot, g, x)$ will be unbiased for m_{gxr} , by the standard results.

Now, we can expand $\mathbb{P}(Y = y \mid R = r)$ as

$$\begin{aligned}
\mathbb{P}(Y = y \mid R = r) &= \sum_{x \in \mathcal{X}, g \in \mathcal{G}} \mathbb{P}(Y = y \mid R = r, G = g, X = x) \mathbb{P}(G = g, X = x \mid R = r) \\
&= \sum_{x \in \mathcal{X}, g \in \mathcal{G}} m_{gxr} q_{gx|r}.
\end{aligned}$$

Since $\hat{\mu}_{Y|RGX}^{(\text{ols})}(y \mid \cdot, g, x)$ is unbiased for m_{gxr} , by the linearity of expectation the poststratified estimator $\hat{\mu}_{Y|R}^{(\text{p-ols})}(y \mid r)$ is unbiased for $\mathbb{P}(Y = y \mid R = r)$. \square

Theorem 3.3 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators). *For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\hat{\mu}_{Y|R}^{(\text{wtd})}(y \mid \cdot) = \hat{\mu}_{Y|R}^{(\text{ols})}(y \mid \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j \mid G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k \mid G_i, X_i, S_i) = 0$ and vice versa) or $\hat{\mu}_{Y|R}^{(\text{wtd})}(y \mid j) = \hat{\mu}_{Y|R}^{(\text{wtd})}(y \mid k)$.*

Proof. Fix a $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$. The weighting estimator of $\mathbb{P}(Y = y \mid R = r)$ within the set of individuals with $G_i = g$ and $X_i = x$ may be written

$$\hat{\mu}_{Y|RGX}^{(\text{wtd})}(y \mid r, g, x) = \frac{\hat{\mathbf{P}}_{I(xg)r}^\top \mathbf{1}\{\mathbf{Y}_{I(xg)} = y\}}{\hat{\mathbf{P}}_{I(xg)r}^\top \mathbf{1}} = \frac{\left\| \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{1}\{\mathbf{Y}_{I(xg)} = y\}) \right\|}{\left\| \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{1}) \right\|},$$

the ratio of the projected length of the outcome vector $\mathbf{1}\{\mathbf{Y} = y\}$ and the constant vector $\mathbf{1}$ onto $\hat{\mathbf{P}}_{I(xg)r}$. We can write the OLS estimator as

$$\hat{\mu}_{Y|R}^{(\text{ols})} = (\hat{\mathbf{P}}_{I(xg)r}^\top \hat{\mathbf{P}}_{I(xg)r})^{-1} \hat{\mathbf{P}}_{I(xg)r}^\top \mathbf{1}\{\mathbf{Y}_{I(xg)} = y\} = \text{coord}_{\hat{\mathbf{P}}_{I(xg)r}}(\text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{1}\{\mathbf{Y}_{I(xg)} = y\})),$$

where $\text{coord}_{\hat{\mathbf{P}}_{I(xg)r}}$ is the function that returns the coordinates of its input vector in the $\hat{\mathbf{P}}_{I(xg)r}$ basis (by assumption $\hat{\mathbf{P}}_{I(xg)r}$ has rank $|\mathcal{R}|$ and so its columns are linearly independent). To make the comparison even easier, notice that we can break the projection $\text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}$ into two steps, writing it instead as $\text{proj}_{\hat{\mathbf{P}}_{I(xg)r}} = \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}} \circ \text{proj}_{\hat{\mathbf{P}}}$. Letting $\mathbf{Y}_{\text{proj}} = \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{1}\{\mathbf{Y}_{I(xg)} = y\})$, then, we can rewrite our estimators as

$$\hat{\mu}_{Y|R}^{(\text{wtd})}(y \mid r) = \frac{\left\| \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{Y}_{\text{proj}}) \right\|}{\left\| \text{proj}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{1}) \right\|} \quad \text{and} \quad \hat{\mu}_{Y|R}^{(\text{ols})}(y \mid r) = \text{coord}_{\hat{\mathbf{P}}_{I(xg)r}}(\mathbf{Y}_{\text{proj}})_r.$$

Now, since the individual BISG probabilities are nonnegative and sum to 1, a pair $j, k \in \mathcal{R}$ of races has perfectly discriminating BISG probabilities if and only if the corresponding columns of $\hat{\mathbf{P}}$ are orthogonal, i.e., $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = I$. Begin by writing \mathbf{Y}_{proj} in terms of the $\hat{\mathbf{P}}_{I(xg)}$ basis, so

$$\mathbf{Y}_{\text{proj}} = \sum_{j \in \mathcal{R}} c_j \hat{\mathbf{P}}_{I(xg)j},$$

and thus $\hat{\mu}_{Y|R}^{(\text{ols})}(y | j) = c_j$. Without loss of generality, suppose the c_j are numbered as $c_1 \geq c_2 \geq \dots \geq c_{|\mathcal{R}|}$. We can also expand $\mathbf{1}$ in the same basis. Since the individual probabilities must sum to one, in fact we have $\mathbf{1} = \sum_{j \in \mathcal{R}} \hat{\mathbf{P}}_{I(xg)j}$.

For the forward direction, we assume $\hat{\mu}_{Y|R}^{(\text{wtd})}(y | j) = \hat{\mu}_{Y|R}^{(\text{ols})}(y | j) = c_j$; multiplying out the denominator of the weighting estimator, we have $\hat{\mathbf{P}}_{I(xg)j}^\top \mathbf{Y} = c_j \hat{\mathbf{P}}_{I(xg)j}^\top \mathbf{1}$ for all j ; substituting the basis expansions of \mathbf{Y}_{proj} and $\mathbf{1}$, this yields

$$\sum_{k \in \mathcal{R}} c_k \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = \sum_{k \in \mathcal{R}} c_j \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k}, \quad \text{so} \quad \sum_{k \in \mathcal{R}} (c_j - c_k) \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = 0.$$

Now fix $j \in J_1 = \arg \max_j c_j$; this relation still holds, but now every term in the sum is nonnegative and in particular $c_j > c_k$ for all $k \notin J_1$. Therefore we must have $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = 0$ for all $k \notin J_1$. Then fix $j \in J_2 = \arg \max_{j \notin J_1} c_j$; since $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)l} = 0$ for all $l \in J_1$, every term in the sum is still nonnegative and in particular $c_j > c_k$ for all $k \notin J_1 \cup J_2$. Therefore we must have $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = 0$ for all $k \notin J_1 \cup J_2$. Proceeding this way through all sets of common values in the c_j we find that for all $j, k \in \mathcal{R}$, either $c_j = c_k$ or $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = 0$.

For the reverse direction, fix $j \in \mathcal{R}$ and let $J = \{k \in \mathcal{R} : c_k = c_j\}$, so that by assumption $\hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k} = 0$ for all $k \notin J$. Then by the above basis expansion, $\hat{\mu}_{Y|RGX}^{(\text{ols})}(y | j) = c_j$, and

$$\hat{\mu}_{Y|R}^{(\text{wtd})}(y | j) = \frac{\sum_{k \in \mathcal{R}} c_k \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k}}{\sum_{k \in \mathcal{R}} \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k}} = \frac{c_j \sum_{k \in J} \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k}}{\sum_{k \in J} \hat{\mathbf{P}}_{I(xg)j}^\top \hat{\mathbf{P}}_{I(xg)k}} = c_j = \hat{\mu}_{Y|R}^{(\text{ols})}(y | j). \quad \square$$

Theorem 3.4 (Nonparametric Identification Under Assumption CI-YSF). *Let $f : \mathcal{S} \rightarrow \mathbb{R}^d$, $d < |\mathcal{S}|$, with range $f(\mathcal{S})$. For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r | G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y | G = g, X = x, S = s)$. Then under Assumption CI-YSF, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y | R, f(\mathcal{S}) = z, G = g, X = x)$ are identified if and only if both \mathbf{P} and the augmented matrix $(\mathbf{P} \ \mathbf{b})$ have rank $|\mathcal{R}|$.*

Proof. The argument is identical to the proof of Theorem 3.1.

Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y | f(\mathcal{S}), R, G, X$, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned} \mathbb{P}(Y = y | G = g, X = x, S = s) &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y | R = r, f(\mathcal{S}) = f(s), G = g, X = x, S = s) \\ &\quad \times \mathbb{P}(R = r | G = g, X = x, S = s) \\ &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y | R = r, f(\mathcal{S}) = f(s), G = g, X = x) \end{aligned}$$

$$\times \mathbb{P}(R = r \mid G = g, X = x, S = s).$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $z \in f(\mathcal{S})$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, f(\mathcal{S}) = s, G = g, X = x)$. These parameters are identified if and only if this system has a unique solution, i.e. if the coefficient matrix \mathbf{P} has rank $|\mathcal{R}|$ and so does the augmented matrix $(\mathbf{P} \ \mathbf{b})$. \square

B Additional Small-area Accuracy Evaluation

We evaluate the small-area estimates with two additional measures. First, we calculate the root-mean-square error (RMSE) of the estimated conditional probabilities by race within each geographic area, and then average this across all geographic areas. This captures the overall accuracy of the estimates. Second, to measure how well each method captures relative differences between geographic areas, we calculate the correlation between the estimated and true conditional probabilities across all geographic areas by race. As in the main text, we remove area-race cells with fewer than 5 voters. A set of estimates which uniformly underestimates the proportion of Black voters which are registered Democrats, but which otherwise correctly orders geographic areas according to their proportion of Black Democrats, will score high on the correlation measure but also higher in RMSE.

Figure B1 summarizes our results, which closely track the findings of Section 4.4. The BIRDiE models are more accurate at all geographic levels and for both Black and White voters. The weighting estimator performs the worst of all the methods.

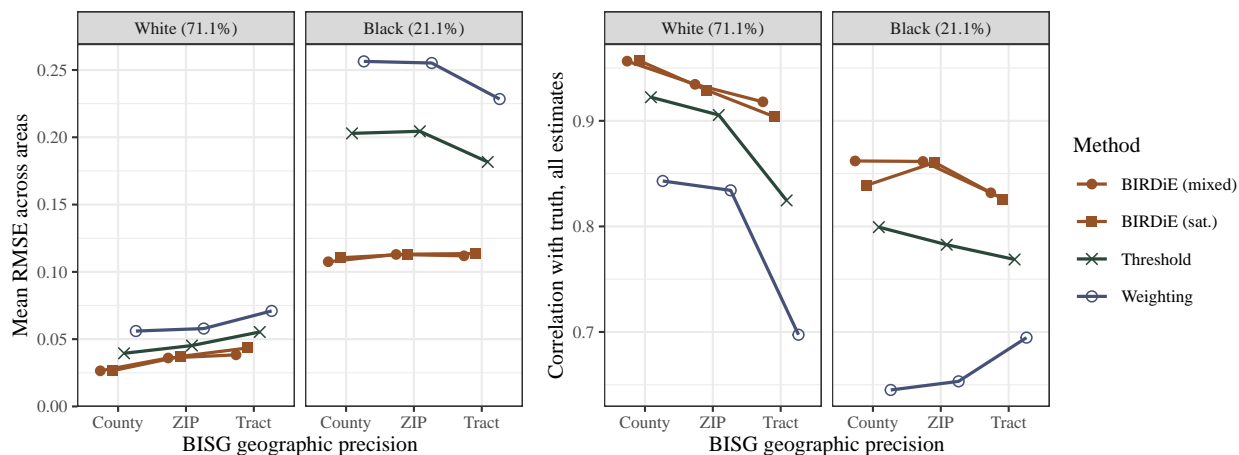


Figure B1: Accuracy of small-area estimates by race, as measured by root-mean-square error (RMSE; lower is better) and the correlation between the estimates and ground truth (higher is better).

C Sensitivity Analysis

C.1 Local sensitivity analysis

In this section, we develop a sensitivity analysis that assesses how the bias in BISG race probabilities affect the estimates of racial disparities. In particular, we consider a setting where Assumptions CI-SG and ACC may be violated but Assumption CI-YS still holds. For example, consider the existence of unobserved confounder that affects some or all of the variables except the outcome, i.e., (R, S, G, X) . This leads to the violation of Assumption CI-SG, but Assumption CI-YS continues to be satisfied so long as such unobserved confounder

does not affect the outcome. Unfortunately, even inaccurate BISG predictions can still lead to biased estimates of racial disparities.

Specifically, if either the Census data are inaccurate, or the conditional independence relation does not hold $S \perp\!\!\!\perp G, X \mid R$, then the BISG predictions $\hat{\mathbf{P}}$ will differ from the “true” individual race probabilities $\mathbf{P}^* = \mathbb{P}(R \mid G, X, S)$. Our goal is to quantify how an error in these probabilities $\mathbf{P}^* - \hat{\mathbf{P}}$ shifts the posterior and hence the estimates of racial disparities.

Denote by π_δ the posterior constructed using the error-corrected BISG race probabilities $\hat{\mathbf{P}}_i + \delta_i$ as the input probabilities for the model (see Equation (3)), where π_{δ^*} is the true posterior with $\delta_i^* := \mathbf{P}_i^* - \hat{\mathbf{P}}_i$. Estimating how π and π_δ differ in general is difficult, but we focus on the settings where δ is small enough to make a linear approximation appropriate. In sum, we aim to quantify how the small error in BISG probabilities can alter the estimates of racial disparities.

For clarity, in this section we will use $\theta_{rG_iX_iY_i}$ to denote the model parameter or function thereof that represents $\pi(Y_i \mid R_i = r, G_i, X_i)$. This mirrors the notation of most of the specific models discussed in Section 3.2 above. Then define the following perturbation weight, which represents the ratio of posterior based on the biased and error-corrected BISG race probabilities:

$$w(\Theta, \delta^*) := \prod_{i=1}^N \left(1 + \frac{\theta_{\cdot G_i X_i Y_i}^\top \delta_i^*}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i} \right) \propto \frac{\pi_{\delta^*}(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})}{\pi(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})},$$

Then, using a local linear approximation, we write the bias for a particular quantity of interest $g(\Theta)$ as

$$\begin{aligned} \mathbb{E}_{\pi_{\delta^*}}[g(\Theta)] - \mathbb{E}_\pi[g(\Theta)] &= \left. \frac{d \mathbb{E}_{\pi_\delta}[g(\Theta)]}{d\delta} \right|_{\delta=0}^\top \delta^* + o(\|\delta^*\|) \\ &= \text{Cov}_\pi \left(g(\Theta), \left. \frac{d \log w(\Theta, \delta)}{d\delta} \right|_{\delta=0} \right)^\top \delta^* + o(\|\delta^*\|), \end{aligned} \quad (6)$$

where the second equality is due to Theorem 2.1 of (Giordano et al., 2018; see also the idea of *local sensitivity* from Gustafson, 1996).

With this representation, we can bound the total error in $\mathbb{E}_\pi[g(\Theta)]$ for sufficiently small δ as the following theorem shows.

Theorem C.1 (Bias Bound). *Define $\tilde{\vartheta}_{ir} := \frac{\theta_{rG_iX_iY_i}}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}$. Then for any input probabilities with total error $\|\delta^*\|^2 =$*

$$\sum_{i=1}^n \|\delta_i^*\|^2 \leq \Delta^2,$$

$$|\mathbb{E}_{\pi^*}[g(\Theta)] - \mathbb{E}_\pi[g(\Theta)]| \lesssim \Delta \|\text{Cov}_\pi(g(\Theta), \tilde{\vartheta})\|, \quad (7)$$

as $\Delta \rightarrow 0$.

Proof. This is immediate from (6) once we compute

$$\begin{aligned} \frac{d \log w(\Theta, \delta)}{d\delta_{ir}} &= \frac{d}{d\delta_{ir}} \sum_{i=1}^N \log \left(1 + \frac{\theta_{\cdot G_i X_i Y_i}^\top \delta_i}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i} \right) \\ &= \frac{d}{d\delta_{ir}} \log \left(1 + \frac{\theta_{\cdot G_i X_i Y_i}^\top \delta_i}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i} \right) \\ &= \frac{1}{1 + \frac{\theta_{\cdot G_i X_i Y_i}^\top \delta_i}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}} \times \frac{\theta_{rG_iX_iY_i}}{\theta_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i} \end{aligned}$$

$$= \frac{\theta_{rG_i X_i Y_i}}{\theta_{\cdot G_i X_i Y_i}^\top (\hat{\mathbf{P}}_i + \delta_i)}$$

and evaluate at $\delta = 0$, since the worst-case bias for a fixed total error can be obtained by having the maximum allowable δ point in the direction of the gradient of $\log w(\Theta, \delta)$. \square

The theorem shows that once researchers choose the amount of total error Δ , then the bound on the shift in a quantity of interest can be computed readily from posterior draws. It is important to note that both δ^* and $\text{Cov}_\pi(g(\Theta), \hat{\mathcal{Y}})$ are vectors whose dimension depends on the sample size N . Thus, all else being equal, their norms will each grow as \sqrt{N} . However, each entry $\text{Cov}_\pi(g(\Theta), \hat{\mathcal{Y}}_{ir})$ will tend to shrink as N increases, since each observation exerts less leverage on the overall posterior. Thus the overall impact of the sample size on the bound in Equation (7) may depend on specific features of the data. In particular, and as should be expected, the error is not guaranteed to vanish as $N \rightarrow \infty$. Practitioners should evaluate Equation (7) under a range of plausible Δ to understand how robust their findings are to worst-case linear violations of Assumptions CI – SG and ACC.

C.2 OLS sensitivity analysis

For a different understanding of the effect of a particular δ , we can derive a result on the error in conditional probability estimates under the OLS estimator and particular configurations of δ . Unlike Theorem C.1, this result holds across all sizes of δ , and not just asymptotically as $\|\delta\| \rightarrow 0$. However, it applies to the OLS estimator, which, while unbiased, we do not recommend in practice. Despite this difference, we expect many of the qualitative conclusions to hold for BIRDIE models. The effect of any particular δ can of course be calculated directly by re-fitting the model to new race probabilities.

Here, we will work with a fixed $y \in \mathcal{Y}$ and among the subset of individuals with a particular $g \in \mathcal{G}$ and $x \in \mathcal{X}$. Then for notational simplicity we let $\hat{\mu}^{(\text{ols})}$ be the vector of estimates of $\mathbb{P}(Y = y \mid R, G = g, X = x)$, and μ the corresponding true probabilities. Similarly, we write $\hat{\mathbf{P}}$ for the matrix of individual race probability estimates for the subset of individuals with $g \in \mathcal{G}$ and $x \in \mathcal{X}$; elsewhere in the text this would be notated $\hat{\mathbf{P}}_{\mathcal{I}(xg)}$

Proposition C.2 (OLS Bias from incorrect $\hat{\mathbf{P}}$). *Let Assumption CI-YS hold. If the OLS estimator $\hat{\mu}^{(\text{ols})}$ is calculated using race probabilities $\hat{\mathbf{P}}$ which differ from the true probabilities $\mathbf{P}^* = \hat{\mathbf{P}} + \delta$, then its bias satisfies*

$$\mathbb{E}[\hat{\mu}^{(\text{ols})}] - \mu = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \delta \mu$$

Proof. We can write the OLS estimate as

$$\hat{\mu}^{(\text{ols})} = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \mathbf{1}\mathbf{Y} = y.$$

As shown in Theorem 3.2, under Assumption CI-YS, $\mathbb{P}(Y = y \mid R = r, G = g, X = x)$ is linear in the true \mathbf{P}^* . Thus letting $\varepsilon = \mathbf{1}\mathbf{Y} = y - \mathbb{P}(Y = y \mid R = r, G = g, X = x)$, we can substitute and find

$$\begin{aligned} \hat{\mu}^{(\text{ols})} &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\mathbf{P}^* \mu + \varepsilon) \\ &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top ((\hat{\mathbf{P}} + \delta) \mu + \varepsilon). \end{aligned}$$

Taking an expectation, since $\mathbb{E}[\varepsilon] = 0$ we find

$$\begin{aligned} \mathbb{E}[\hat{\mu}^{(\text{ols})}] &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top ((\hat{\mathbf{P}} + \delta) \mu) \\ &= \mu + (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\delta \mu); \end{aligned}$$

rearrangement yields the result. \square

Informally, for BISG error δ to cause problems with the OLS estimate, two things must happen. First, within individuals it must be “correlated” (i.e., have nonzero inner product) with the true conditional probabilities μ . Since δ_i must always sum to zero, practically, this means that positive BISG errors must tend to occur in racial groups which have a relatively high occurrence of outcome y : $\mathbb{P}(Y = y \mid R = r, G = g, X = x) > \mathbb{P}(Y = y \mid G = g, X = x)$. Second, the vector $\delta\mu$ (where each entry measures this “correlation” between errors and relative frequencies) must be correlated the BISG probabilities themselves \hat{P} . For example, if $\delta_i\mu$ is positive and tends to be larger for individuals with a high BISG probability of being Hispanic, then the overall OLS estimator the conditional probability of $Y = y$ among Hispanics will be biased upwards.

While Proposition C.2 applies within a (G, X) cell, if the same conditions hold across all (G, X) combinations, then the overall poststratified estimator will be similarly biased.

D Surname Groupings for North Carolina Robustness Analysis

We classify every surname in the voter file into one of nine groups, each containing surnames from one or more of 22 surname groups that we provide in the replication data and software. These groups are organized mainly around different regions of the world and different waves of immigration to the United States. To create the surname groups, each individual in the 1930 Census data was classified into one of the 22 groups. Then among the set of individuals with each surname, the group with the highest number of individuals relative to the whole population was assigned to that surname. For example, while most people named “Smith” fall into the Anglosphere group (containing 3rd or more generation White U.S. residents as of 1930, as well as immigrants from the U.K., Canada, Australia, etc.), there are relatively more Smiths among Black people than any other of the 22 groups. Thus “Smith” is assigned to the Black surname group. The full code for creating the 22 surname groupings from the 1930 Census data is available in the replication materials.

Because of the demographics of the United States, as well as limitations of the source data there is more geographic specificity in the surname groupings for some regions (e.g., Europe) than for others (e.g., South America and Africa). We collapse the 22 surname groups to nine for the robustness analysis in Section 4 based on the demographics of North Carolina specifically and to minimize the computational burden of performing the robust analysis.

The 50 roughly most frequent surnames in each group, along with a brief description of the group, are listed below. We stress that for the purposes of sensitivity analysis, the surname groups need only be correlated with countries of origin and racial subgroups. Perfect alignment is neither possible nor necessary.

Anglosphere and Black surname group. Surnames which are relatively more prevalent among 3rd-or-more generation White U.S. residents and Black U.S. residents in 1930.

1. SMITH	11. LEWIS	21. HALL	31. COLLINS	41. COX
2. WILLIAMS	12. ROBINSON	22. CAMPBELL	32. STEWART	42. WARD
3. BROWN	13. WALKER	23. MITCHELL	33. MORRIS	43. RICHARDSON
4. JONES	14. ALLEN	24. CARTER	34. COOK	44. WATSON
5. DAVIS	15. WRIGHT	25. ROBERTS	35. ROGERS	45. BROOKS
6. TAYLOR	16. SCOTT	26. PHILLIPS	36. MORGAN	46. WOOD
7. MOORE	17. HILL	27. EVANS	37. COOPER	47. JAMES
8. JACKSON	18. GREEN	28. TURNER	38. BAILEY	48. BENNETT
9. WHITE	19. ADAMS	29. PARKER	39. REED	49. GRAY
10. CLARK	20. BAKER	30. EDWARDS	40. HOWARD	50. HUGHES

First wave European immigration surname group. Surnames associated with German, Nordic, and Irish immigrants.

1. JOHNSON	11. BURNS	21. CARROLL	31. SCHULTZ	41. HIGGINS
2. ANDERSON	12. OLSON	22. RILEY	32. PEARSON	42. OCONNOR
3. NELSON	13. WAGNER	23. BURKE	33. BARRETT	43. QUINN
4. MURPHY	14. MEYER	24. LARSON	34. BECK	44. SWANSON
5. PETERSON	15. SCHMIDT	25. CARLSON	35. POWERS	45. FITZGERALD
6. KELLY	16. RYAN	26. OBRIEN	36. LEONARD	46. CHRISTENSEN
7. SULLIVAN	17. DUNN	27. LYNCH	37. BENSON	47. MANNING
8. MURRAY	18. KELLEY	28. HANSON	38. LYONS	48. MCLAUGHLIN
9. MCDONALD	19. HANSEN	29. WEBER	39. MCCARTHY	49. DOYLE
10. KENNEDY	20. CUNNINGHAM	30. WALSH	40. ERICKSON	50. BRADY

Second wave European immigration surname group. Surnames associated with Eastern European, Italian, Jewish, Russian, Greek, and other Southern European immigrants.

1. FOX	11. ZIMMERMAN	21. KLINE	31. KATZ	41. NICHOLAS
2. NICHOLS	12. KLEIN	22. BERGER	32. MARINO	42. ROSENBERG
3. HOFFMAN	13. GROSS	23. STEIN	33. BRUNO	43. ROSSI
4. NEWMAN	14. GOODMAN	24. RAYMOND	34. MOSER	44. SINGER
5. SCHNEIDER	15. SHERMAN	25. FRIEDMAN	35. GOLDSTEIN	45. ABRAMS
6. KELLER	16. WOLF	26. LEVY	36. GOLDBERG	46. ACKERMAN
7. GREGORY	17. KRAMER	27. NOVAK	37. KAPLAN	47. HELLER
8. SCHWARTZ	18. NICHOLSON	28. KAUFMAN	38. KESSLER	48. STERN
9. COHEN	19. WEISS	29. LEVINE	39. ROMANO	49. SCHAFER
10. BECKER	20. RUSSO	30. LEHMAN	40. FINK	50. SHAPIRO

East Asian surname group. Surnames associated with Chinese, Japanese, and Korean immigrants.

1. LEE	11. BOWEN	21. HORNE	31. HAN	41. LIANG
2. YOUNG	12. LIU	22. XIONG	32. LAU	42. SUN
3. WONG	13. PAUL	23. LIM	33. MA	43. JUNG
4. WANG	14. CHAN	24. TANG	34. PUCKETT	44. ZHOU
5. PARK	15. TODD	25. CHO	35. CHIN	45. GEE
6. MAY	16. ZHANG	26. CHENG	36. GIL	46. ZHAO
7. JOSEPH	17. LANG	27. KANG	37. XU	47. SHIN
8. LOWE	18. YU	28. LAW	38. SONG	48. OHARA
9. CHANG	19. CHOI	29. CRAFT	39. KAY	49. ZHU
10. LIN	20. MOON	30. NG	40. STROUD	50. YEE

South Asian surname group. Surnames associated with Indian and Southwest Asian immigrants.

1. WILSON	11. CARR	21. DAVID	31. MOHAMED	41. SAMUEL
2. THOMAS	12. SINGH	22. HOWE	32. BOGGS	42. SEWELL
3. PATEL	13. BISHOP	23. HAHN	33. KUMAR	43. HASSAN
4. STEVENS	14. MANN	24. GOOD	34. WESTON	44. SADLER
5. WOODS	15. FRANCIS	25. JOHN	35. BEATTY	45. PINTO
6. SHAW	16. GILL	26. OSBORN	36. SWAIN	46. MAJOR
7. FERGUSON	17. YATES	27. ABRAHAM	37. GOMES	47. BARNHART
8. RAY	18. MARSH	28. RODRIGUES	38. JACOB	48. CARMICHAEL
9. WILLIS	19. ROY	29. PEREIRA	39. TOLBERT	49. MUHAMMAD
10. GEORGE	20. KAUR	30. SHARMA	40. PAYTON	50. GUPTA

Southeast Asian and Pacific surname group. Surnames associated with Southeast Asian and Pacific Islander immigrants, including Vietnamese and Filipino immigrants.

1. MILLER	11. SILVA	21. HUANG	31. PHAN	41. HOANG
2. MARTIN	12. SANTOS	22. WU	32. VO	42. CASH
3. KING	13. GREENE	23. ROWE	33. VU	43. BUI
4. NGUYEN	14. LI	24. BAUTISTA	34. LU	44. CHU
5. KIM	15. LE	25. HOUSTON	35. NGO	45. SINCLAIR
6. LONG	16. YANG	26. LAM	36. TAN	46. SORIANO
7. TRAN	17. LITTLE	27. HUYNH	37. HONG	47. ZHENG
8. CHEN	18. MORAN	28. HO	38. DANG	48. LESLIE
9. WEBB	19. PHAM	29. CHUNG	39. DO	49. ANGEL
10. GORDON	20. RAMSEY	30. TRUONG	40. LY	50. DUONG

Non-Cuban Hispanic surname group. Surnames associated with Mexican and Latin American immigrants, not including Cuban immigrants, and Puerto Rican residents.

1. GARCIA	11. RIVERA	21. MENDOZA	31. CASTRO	41. ALVARADO
2. RODRIGUEZ	12. GOMEZ	22. RUIZ	32. FERNANDEZ	42. DELGADO
3. MARTINEZ	13. DIAZ	23. CASTILLO	33. VARGAS	43. PENA
4. HERNANDEZ	14. CRUZ	24. GONZALES	34. GUZMAN	44. CONTRERAS
5. LOPEZ	15. REYES	25. VASQUEZ	35. MENDEZ	45. SANDOVAL
6. PEREZ	16. MORALES	26. ROMERO	36. MUNOZ	46. GUERRERO
7. SANCHEZ	17. GUTIERREZ	27. MORENO	37. SALAZAR	47. RIOS
8. RAMIREZ	18. ORTIZ	28. HERRERA	38. GARZA	48. ESTRADA
9. TORRES	19. RAMOS	29. MEDINA	39. SOTO	49. ORTEGA
10. FLORES	20. CHAVEZ	30. AGUILAR	40. VAZQUEZ	50. NUNEZ

Cuban surname group. Surnames associated with Cuban immigrants.

1. GONZALEZ	11. SUAREZ	21. CRANE	31. SARGENT	41. MARRERO
2. ALVAREZ	12. CONNER	22. FRYE	32. GORE	42. VALDES
3. JIMENEZ	13. SANTANA	23. PARRA	33. ZIEGLER	43. OLIVA
4. BOWMAN	14. DECKER	24. MAYO	34. TOMLINSON	44. MCCLENDON
5. DAVIDSON	15. SKINNER	25. DAVIES	35. LOWRY	45. QUEEN
6. ACOSTA	16. ABBOTT	26. BLANCO	36. PAGAN	46. MCCORD
7. MOLINA	17. GARRISON	27. WITT	37. LORD	47. CRESPO
8. MIRANDA	18. PONCE	28. CARRASCO	38. CARBAJAL	48. CORNEJO
9. CASTANEDA	19. PALACIOS	29. ALONSO	39. BETANCOURT	49. DUMAS
10. BALL	20. SLOAN	30. HAINES	40. PATINO	50. BUENO

“Other” surname group. Surnames not associated with one of the other categories, including those associated with later Western European immigration, Middle Eastern & North African-associated surnames, Native-associated surnames and Afro-Caribbean-associated surnames.

1. PERRY	11. WELCH	21. SIMON	31. FRANCO	41. MCKENZIE
2. HENRY	12. DAY	22. CUMMINGS	32. HAMMOND	42. BEIL
3. HUNT	13. STANLEY	23. CHANDLER	33. CLARKE	43. COCHRAN
4. ROSE	14. HOPKINS	24. SHARP	34. WATERS	44. NASH
5. PIERCE	15. LAMBERT	25. BARBER	35. FRANK	45. BRYAN
6. PETERS	16. NORRIS	26. GRIFFITH	36. ANDRADE	46. MEYERS
7. KNIGHT	17. WALTERS	27. PACHECO	37. LLOYD	47. CARSON
8. RICHARDS	18. STEELE	28. CROSS	38. FRENCH	48. WILKINSON
9. MORRISON	19. BUSH	29. GOODWIN	39. OWEN	49. ATKINSON
10. JACOBS	20. WOLFE	30. MULLINS	40. CHARLES	50. VINCENT

E Estimating mortgage rates by race

The decennial census reports the number of homeowners by race, and the number of homeowners with a mortgage, but does not report the number of mortgages by race. Thus we are left to infer the mortgage-race distribution from this marginal information. Fortunately, both ownership-race and mortgage-ownership marginals are reported at fine geographic levels.

We therefore produce estimates of mortgage rates by race at the ZCTA level, then aggregate these estimates to the nation. Stratifying by ZCTA means that any variation in mortgage rates by race that is explained by geographic variation will be captured.

Within each ZCTA we estimate the number of mortgages by race by taking the fraction of homeowners with a mortgage and multiplying by the number of homeowners of each racial group. Nationwide, we find no association between the racial composition of a ZCTA and the fraction of homeowners with a mortgage. While not dispositive because of possible aggregation bias, this finding nevertheless suggests that after controlling for geography, little variation by racial group in the fraction of homeowners with a mortgage remains.

Table E1 reports our nationwide estimates of the fraction of each racial group with a mortgage.

Table E1: Census-reported number of households, and fraction that own their home, by race, and estimated fraction that have a mortgage, by race.

Race	Total households	Fraction owner-occupied	Fraction with mortgage
White	82,343,859	72.2%	49.9%
Black	13,797,354	44.6%	31.7%
Hispanic	14,822,017	49.6%	33.7%
Asian	4,580,883	58.1%	44.4%
Native	758,975	57.2%	32.0%
Other	1,788,360	49.3%	35.6%