Mitigating Included- and Omitted-Variable Bias in Estimates of Disparate Impact*

Jongbin Jung, Sam Corbett-Davies, Johann D. Gaebler[1], Ravi Shroff[2], and Sharad Goel[3]

[1]Department of Statistics, Harvard University
[2]Department of Applied Statistics, Social Science, and Humanities, New York University
[3]Harvard Kennedy School, Harvard University

Abstract

Managers, employers, policymakers, and others often seek to understand whether decisions are biased against certain groups. One popular analytic strategy is to estimate disparities after adjusting for observed covariates, typically with a regression model. This approach, however, suffers from two key statistical challenges. First, omitted-variable bias can skew results if the model does not adjust for all relevant factors; second, and conversely, included-variable bias—a lesser-known phenomenon—can skew results if the set of covariates includes irrelevant factors. Here we introduce a new, three-step statistical method, which we call risk-adjusted regression, to address both concerns in settings where decision makers have clearly measurable objectives. In the first step, we use all available covariates to estimate the value, or inversely, the *risk*, of taking a certain action, such as approving a loan application or hiring a job candidate. Second, we measure disparities in decisions after adjusting for these risk estimates alone, mitigating the problem of included-variable bias. Finally, in the third step, we assess the sensitivity of results to potential mismeasurement of risk, addressing concerns about omitted-variable bias. To do so, we develop a novel, non-parametric sensitivity analysis that yields tight bounds on the true disparity in terms of the average gap between true and estimated risk—a single interpretable parameter that facilitates credible estimates. We demonstrate this approach on a detailed dataset of 2.2 million police stops of pedestrians in New York City, and show that traditional statistical tests of discrimination can substantially underestimate the magnitude of disparities.

1 Introduction

Studies of discrimination generally start by assessing whether certain groups, particularly those defined by race and gender, receive favorable decisions more often than others. For example, one might examine whether loan applications from white candidates are granted more often than those from racial minorities, or whether male employees are promoted more often than women. Although observed disparities may be the result of bias, it is also possible that they stem from statistical differences between groups. In particular, if some groups contain disproportionately many qualified members, then one would also expect those groups to receive disproportionately many favorable decisions, even in the absence of discrimination.

To tease apart these two possibilities—group differences versus discrimination—the most popular statistical approach is ordinary linear or logistic regression. In the banking context, for instance, one could examine race-contingent lending rates after adjusting for relevant factors, such as income and credit history. Disparities that persist after accounting for such factors are often interpreted as evidence of discrimination. This basic statistical strategy has been used in numerous studies to test for bias across domains, including education (Espenshade et al., 2004; Grossman et al., 2023b), employment (Polachek, 2008), criminal justice (Abrams, 2014; Gaebler et al., 2022; Gelman et al., 2007; MacDonald and Raphael, 2019; Rehavi and Starr, 2012), and medicine (Balsa et al., 2005).

*We thank Josh Grossman, Julian Nyarko, and João M. Souto-Maior for helpful conversations. All data and replication code are available at: https://github.com/jgaeb/rar-repro. The sensitivity analysis is implemented in the rar R package available on CRAN.

Despite the ubiquity of such regression-based tests for discrimination, the approach suffers from two serious statistical limitations. First, the well-known problem of omitted-variable bias arises when decisions are based in part on relevant factors that correlate with group membership, but which are omitted from the regression (Angrist and Pischke, 2008). For example, if lending officers consider an applicant's payment history, and if payment history correlates with race but is not recorded in the data (and thus cannot be included in the regression), the results of the regression can suggest discrimination where there is none, or *vice versa*. Unfortunately, omitted-variable bias is the rule rather than the exception. It is generally prohibitive to measure every variable relevant to a decision, and it is likely that most unmeasured variables are at least weakly correlated with demographic attributes, skewing results.

A second problem with regression-based tests is what Ayres (2005, 2010) calls *included-variable bias*, an issue as important as omitted-variable bias in studies of discrimination but one that receives far less attention. To take an extreme example, it is problematic to include control variables in a regression that are obvious proxies for legally protected attributes—such as vocal register as a proxy for gender—when examining the extent to which observed disparities stem from group differences in qualification. Including such proxies will typically lead one to underestimate the true magnitude of discrimination in decisions. But what counts as a "proxy" is not always clear. For example, given existing patterns of residential segregation, one might argue that ZIP codes are a proxy for race, and thus should be excluded when testing for racial bias. But one could also argue that ZIP code provides legitimate information relevant to a decision, and so excluding it would lead to omitted-variable bias. Ayres (2010) proposes a middle ground, suggesting that potential proxies should be included, but their coefficients capped to a "justifiable" level; in practice, however, it is difficult to determine and defend specific constraints on regression coefficients.[1]

Here we present a statistically principled and logistically straightforward method for measuring discrimination that addresses both omitted- and included-variable bias. Our method, which we call risk-adjusted regression, proceeds in three steps. In the first step, we use all available information, including potential proxies of protected traits, to estimate the value—or, equivalently, the risk—of taking a particular action. For example, in the lending context, we might estimate an applicant's risk of default if granted a loan, conditional on all available covariates. In the second step, we assess disparities by regressing decisions (e.g., loan offers) on individual-level risk estimates and protected traits alone, allowing us to measure the extent to which similarly qualified individuals are treated differently. This strategy can be seen as formalizing the coefficient-capping procedure of Ayres (2010)—with covariates used only to the extent that they are statistically justified by risk—and thus circumvents the problem of included-variable bias. Finally, we assess the sensitivity of results to potential mismeasurement of risk. In particular, we derive tight analytic bounds on risk-adjusted disparities as a function of the extent to which risk estimates differ from true risk.

To demonstrate this approach, we examine 2.2 million stops of pedestrians conducted by the New York City Police Department between 2008 and 2011. After adjusting for a stopped individual's statistical risk of carrying a weapon—based in part on detailed behavioral indicators recorded by officers—we find that stopped Black and Hispanic pedestrians are searched for weapons substantially more often than stopped white individuals. We find that these risk-adjusted disparities are considerably larger than disparities suggested by a standard regression that adjusts for all available covariates, underscoring the importance of accounting for included-variable bias. Finally, we show that our results are robust to potentially large errors in risk estimates.

[1] Such problems have prompted a search for alternatives to regression-based approaches. Most prominently, Becker (1993) proposed the outcome test, which is based not on the rate at which decisions are made, but on the success rate of those decisions. In the context of banking, Becker argued that even if, hypothetically, racial minorities were less creditworthy than white applicants, minorities who were granted loans should still be found to repay their loans at the same rate as white applicants who were granted loans. If loans to minorities had a higher repayment rate than loans to white borrowers, it suggests that lenders applied a double standard (intentionally or not), granting loans only to exceptionally qualified minorities—potentially violating disparate impact laws. The outcome-based approach has been applied almost as broadly as simple regression to study discrimination in, for example, policing (Ayres, 2002; Goel et al., 2016b, 2017; Knowles et al., 2001), lending (Berkovec et al., 2018), and scientific publication (Smart and Waldfogel, 1996). Outcome tests, however, have their own significant statistical shortcomings, most notably the problem of infra-marginality (Ayres, 2002; Pierson et al., 2018; Simoiu et al., 2017), which can lead the test to incorrectly suggest an absence of discrimination (Pierson et al., 2020).

2 Theories of Discrimination

There are two main legal doctrines of discrimination in the United States: disparate treatment and disparate impact. Here we describe the conceptual underpinnings of these theories. We further connect these ideas to standard statistical tests of discrimination, highlighting the problem of included-variable bias.

2.1 The Jurisprudence of Discrimination

Disparate treatment derives force from the Equal Protection Clause of the U.S. Constitution's Fourteenth Amendment, and it prohibits government agents from acting with "discriminatory purpose" (Washington v. Davis, 1976). Although equal protection law bars policies undertaken with animus, it allows for the limited use of protected attributes to further a compelling government interest. For example, until recently, certain affirmative action programs for college admissions were legally permissible to further the government's interest in promoting diversity. In 2023, the U.S. Supreme Court overturned the legality of such affirmative action programs, ruling that it was unlawful to explicitly consider race in college admissions decisions (SFFA v. Harvard, 2023).

The most widespread statistical test of such intentional discrimination is ordinary linear or logistic regression, in which one estimates the likelihood of favorable (or unfavorable) decisions across groups defined by race, gender, or other legally protected traits. In this approach, the investigator adjusts for all potentially relevant risk factors, excluding only clear proxies for the protected attributes. Barring omitted-variable bias, non-zero coefficients on the protected traits suggest those factors influenced the decision maker's actions; in the absence of a compelling justification, such evidence is suggestive of a discriminatory purpose. It is difficult—and perhaps impossible—to rigorously define the *influence*, or causal effect, of largely immutable traits like race on decisions (Greiner and Rubin, 2011; VanderWeele and Robinson, 2014), but a regression of this type is nevertheless considered a reasonable first step to identify discriminatory motive, both by criminologists and by legal scholars (Fagan, 2010; Gaebler et al., 2022). For an equal protection claim to succeed in court, however, one typically needs additional documentary evidence (e.g., acknowledgement of an illegitimate motive) to bolster the statistical evidence.

In contrast to disparate treatment, the disparate impact doctrine is concerned with the effects of a policy, not a decision maker's intentions, and it is the primary form of discrimination we study in this paper. Under the disparate impact standard, a practice may be deemed discriminatory if it has an unjustified adverse effect on protected groups, even in the absence of explicit categorization or animus. The doctrine stems from statutory rules, rather than constitutional law, and applies only in certain contexts, such as employment (via Title VII of the 1964 Civil Rights Act) and housing (via the Fair Housing Act of 1968). Apart from federal statutes, some states have passed more expansive disparate impact laws, including Illinois and California.

The disparate impact doctrine was formalized in the landmark U.S. Supreme Court case *Griggs v. Duke Power Co.* (1971). In 1955, the Duke Power Company instituted a policy that mandated employees have a high school diploma to be considered for promotion, which had the effect of drastically limiting the eligibility of Black employees. The Court found that this requirement had little relation to job performance, and thus deemed it to have an unjustified disparate impact. Importantly, the employer's motivation for instituting the policy was irrelevant to the Court's decision; even if enacted without discriminatory purpose, the policy was deemed discriminatory in its effects and hence illegal.

More specifically, the legal test of disparate impact developed over half a century of case law has three principal elements (Grossman et al., 2023a):

1. **Adverse impact:** The plaintiff first must establish that the policy disproportionately impacts the minority group.

2. **No Justification:** Next, the defendant must establish that the adverse impact has a substantial justification rooted in a legitimate policy goal.

3. **Less discriminatory alternative:** Even if the disparate impact is justified, the plaintiff can nevertheless prevail if they demonstrate that there is an alternative feasible policy with less adverse impact on the minority group.

Our concern here is with the second element, namely, whether the adverse impact has some legitimate policy justification.

As discussed above, the standard statistical test for disparate treatment is a "kitchen sink" regression, where one examines the residual explanatory power of protected group status after including all other available covariates as controls. That approach, however, is ill-suited to assess whether practices are rationally justified, which is the relevant standard in disparate impact claims. Ayres (2005) makes the point persuasively in the context of the original *Griggs* decision:

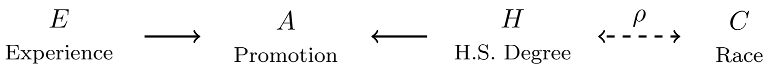
> "One could imagine running a regression to test whether an employer was less likely to hire African American applicants than white applicants. It would be possible to control in this regression for whether the applicant had received a high-school diploma. Under the facts of *Griggs*, such a control would likely have reduced the racial disparity in the hiring rates. But including in the regression a variable controlling for applicants' education would be inappropriate. The central point of *Griggs* was to determine whether the employer's diploma requirement had a disparate racial impact. The possibility that including a diploma variable would reduce the estimated race effect in the regression would in no way be inconsistent with a theory that the employer's diploma requirement disparately excluded African Americans from employment."

In short, by including educational status in the regression, one would mask the policy's unjustified disparate impact.

To assess claims of *unjustified* disparate impact—in *Griggs* and beyond—one would ideally compare decision rates for similarly *qualified* groups of applicants (e.g., similarly qualified white and Black candidates). Unfortunately, if one does not, or cannot, adjust for sufficiently many covariates, omitted-variable bias may skew results; conversely, if one does adjust for a rich set of covariates, included-variable bias may corrupt conclusions.[2]

2.2 A formal illustration of included-variable bias

We give a simple formal illustration of the statistical phenomenon at issue in *Griggs*. Consider the following data-generating process, depicted by the DAG below.



Here, C indicates the race of an employee, with $C = 0$ representing a White employee and $C = 1$ a Black employee; and H indicates whether an employee has a high-school degree. We assume that H and C have a joint distribution given by

$$\Pr(H = 1, C = 1) = \Pr(H = 0, C = 0) = \frac{1+\rho}{4}, \qquad \Pr(H = 1, C = 0) = \Pr(H = 0, C = 1) = \frac{1-\rho}{4}.$$

In particular, it follows that both H and C are marginally $\text{Bern}(\frac{1}{2})$ with correlation ρ. Importantly, we need not assume that race C "causes" an employee to have a high school degree but only that race and having a high school degree have some correlation ρ (although, in this particular example, a causal mechanism like discriminatory enrollment policies is plausible).

Further, $E \sim \text{Unif}(0, 1)$ denotes an employee's level of experience, interpreted as the proportion of employees who have been employed for less time than them. Finally, by A, we denote whether the employee was promoted. Recalling the details of *Griggs*, we assume that

$$\Pr(A = 1 \mid E, H) = E \cdot H.$$

[2] A related body of work conceptualizes discrimination as race-specific differences in decision error rates (Arnold et al., 2020, 2021; Bohren et al., 2022; Coston et al., 2020; Hardt et al., 2016). For example, one might consider differences in loan rejection rates between white and Black applicants who in reality would repay their loans (i.e., a counterfactual false negative rate). Recent work, however, has observed that by conditioning on *ex post* (potential) outcomes—as opposed to *ex ante* risk, as we do—such approaches suffer from infra-marginality and associated statistical issues, rendering them problematic measures of common legal and policy understandings of discrimination (Ayres, 2005; Corbett-Davies et al., 2023; Mayson, 2018; Nilforoshan et al., 2022).

That is, only employees with high school degrees are promoted, and their probability of promotion depends on their experience.

A natural quantity to consider is the average difference in promotion rates for "similarly situated" Black and White employees—that is, the average difference in promotion rates for employees with the same experience level and high school degree. In this case, it is straightforward to show[3] that

$$\mathbb{E}[\Pr(A = 1 \mid E, H, C = 1) - \Pr(A = 1 \mid E, H, C = 0)] = 0. \tag{1}$$

Since race does not factor into promotion decisions, the average difference in hiring rates for similarly situated Black and White employees is zero, suggesting an absence of disparate *treatment*. However, if, following the Court's reasoning in *Griggs*, one were to consider the average difference in promotion rates for employees with the same *experience level*—regardless of whether they had a high school degree—one would obtain

$$\mathbb{E}[\Pr(A = 1 \mid E, C = 1) - \Pr(A = 1 \mid E, C = 0)] = \frac{\rho}{2}, \tag{2}$$

Following the facts of *Griggs*, we expect $\rho < 0$, indicating that Black employees are less likely than White employees to have attained a high-school degree. As a result, as shown in Eq. (2), Black employees at the same level of experience are less likely to be promoted. In particular, this estimand captures the disparate *impact* of the promotion policy. This stylized example illustrates the importance of conditioning on the appropriate variables to avoid included-variable bias in studies of discrimination. Here, high-school degree attainment is inappropriate to condition on because it has little relationship to job performance.

Non-parametric estimands like those shown in Eqs. (1) and (2) can be statistically challenging to estimate in practice, so it is common instead to use linear regression or other parametric models to approximate these estimands. Imagine if an analyst were to fit a linear kitchen-sink regression model of A given experience level (E), high school degree (H), the interaction between these two terms ($E \cdot H$), and race (C),

$$\Pr(A = 1 \mid E, H, C) = \beta_0 + \beta_E \cdot E + \beta_H \cdot H + \beta_{E:H} \cdot E \cdot H + \beta_C \cdot C.$$

Here, β_C is typically interpreted as the magnitude of "discrimination." In this example, $\hat{\beta}_C$ tends to zero in large samples, suggesting a lack of disparate *treatment*; see Appendix A. But, despite fitting a correctly specified model of decisions, the analyst would fail to detect the disparate impact of the hiring policy.

Imagine, instead, that the analyst regressed decisions only on years of experience (E) and race (C), as the Court indicated they should, i.e.,

$$\Pr(A = 1 \mid E, C) = \beta'_0 + \beta'_E \cdot E + \beta'_C \cdot C.$$

In this case, as with the non-parametric estimand in Eq. (2), $\hat{\beta}'_C$ tends to $\frac{\rho}{2}$, reflecting the fact that it correctly captures the disparate impact of the promotion policy. (In this specific example, the parametric and non-parametric estimands are identical, but this is not always the case.)

Following the Court's ruling, we have assumed that any use of education is *unjustified*, while use of experience is *justified* when making promotion decisions. As a result, the key expression of interest is that given in Eq. (2), which quantifies disparities in decision rates after conditioning on experience (E) alone. In practice, however, all covariates are usually at least weakly informative about one's qualifications. In the context of *Griggs*, high school degree attainment, even among employees with the same level of experience, is likely at least somewhat predictive of job performance—making it unclear how one should formally define a measure of disparate impact. In the subsequent section, we introduce one useful way of conceptualizing disparate impact in this more realistic setting.

3 A Statistical Approach to Assessing Disparate Impact

We now formally describe our approach to measuring disparate impact—a procedure we call risk-adjusted regression. The data generating process consists of draws of tuples $(X, \tilde{X}, C, W, A)$ where

$$X \in \mathcal{X}, \qquad \tilde{X} \in \tilde{\mathcal{X}}, \qquad C \in \{1, \ldots, m\}, \qquad W \in \{0, 1\}, \qquad \text{and} \quad A \in \{0, 1\}.$$

[3] See Appendix A for proof.

Here A is a binary decision, X is the set of covariates on which that decision is based, and C represents membership in some protected class. (In general, we make no assumption about whether C is encoded in X.) The binary variable W represents a latent property of interest to the decision maker. However, we assume the decision is made based on X alone, i.e.,

$$A \perp\!\!\!\perp W \mid X.$$

Finally, we assume that an analyst observes the tuple $(\tilde{X}, C, A, A \cdot W)$, based on which they seek to estimate disparate impact in the decision process. Here $\tilde{X}$ represents an alternative set of covariates available to the analyst that may differ from $\tilde{X}$ in arbitrary ways. Additionally, the term $A \cdot W$ reflects the fact that W is only observed by the analyst when action $A = 1$ occurs.

As a concrete example, consider estimating risk-adjusted racial disparities in police searches of pedestrians for weapons—the application we discuss in more detail below. In this case, A_i, X_i, C_i, and W_i, respectively, indicate whether the i-th stopped pedestrian was searched, the information available to the officer when deciding whether to conduct a search, the stopped individual's race, and whether the individual was in possession of a weapon. Further, $\tilde{X}_i$ denotes the information available to the analyst from administrative records of the i-th stop. The fact that the analyst observes $A_i \cdot W_i$ means that they know whether searched individuals were carrying a weapon, but not whether unsearched individuals were carrying a weapon, which is generally the case. Although we do not consider them further here, this general framing applies to a wide variety of settings, such as hiring (where W represents a prospective employee's productivity and A whether or not they were hired) and lending (where W represents the amount a prospective borrower will repay and A represents the bank's lending decision).

Given this setup, we define *ex ante risk* to be:

$$R = \Pr(W = 1 \mid X). \tag{3}$$

In our policing example, R_i is the probability that the i-th stopped individual, with covariates X_i, is carrying a weapon.[4] To ensure that various quantities of interest are well-defined, we assume a variant of the overlap condition holds (Rosenbaum and Rubin, 1983a,b):

$$0 < \Pr(C = j \mid R) < 1, \qquad \text{for all } j = 1, \ldots, m.$$

We note that this version of the overlap condition is immediately implied by an analogue of the stronger standard overlap condition: $0 < \Pr(C = j \mid X) < 1$.

Our goal is to quantify whether decisions systematically differ for individuals at the same level of risk. Many estimands summarizing such potential differences are possible, but, for simplicity, we consider the following non-parametric quantity:

$$\mathbb{E}[\Pr(A = 1 \mid C = j, R) - \Pr(A = 1 \mid C = 1, R)]. \tag{4}$$

Eq. (4) defines risk-adjusted disparities to be the difference in the probability of taking an action for group $C = j$ relative to the reference group $C = 1$, after accounting for potential differences in risk across groups. Our overlap assumption ensures that this quantity is well-defined.

To facilitate computation—and, in particular, the sensitivity analysis presented below—we approximate the estimand in (4) by a linear probability model. Specifically, we estimate $\Pr(A = 1 \mid C, R)$ as

$$\sum_{j=1}^{m} \hat{\beta}_j \cdot \mathbf{1}(C = j) + \hat{\beta}_R \cdot R, \tag{5}$$

where $\mathbf{1}(C = j)$ indicates membership in group j and $\hat{\beta}_j$ is its corresponding fitted coefficient. The difference between fitted coefficients, $\hat{\beta}_j - \hat{\beta}_1$, yields an estimate of (4)—see Appendix C for further details on the statistical quality of this estimator.

[4] To allow for the possibility of statistical discrimination (Arrow, 1973; Chaudhuri and Sethi, 2008; Phelps, 1972) in the decision maker's course of action, risk may be conditioned on group membership as well as other observables—formally, by including C among the covariates X. However, whether the inclusion of race or other protected attributes in risk models is legally appropriate is uncertain. Importantly, our theoretical and empirical results remain essentially unchanged whether or not one assumes X encodes C; see Figures 15 through 21 in Appendix E.

Under this model, if $\hat{\beta}_j$ is greater than $\hat{\beta}_1$, this indicates that members of group j are more likely to receive action $A = 1$ than members of the base group with similar estimated risk. In our policing example, this means that members of group j are searched more often than members of the reference group who were equally likely to be carrying a weapon, and we would say that such elevated search rates are *unjustified* by risk. We note, however, that $\hat{\beta}_j > \hat{\beta}_1$ does not imply *intentional* discrimination—as in *Griggs*, unjustified disparate impact is possible even under a facially neutral policy undertaken without animus.

4 Sensitivity analysis

Our definition of disparate impact depends on the true risks, R_i. But, in practice, analysts observe only partial information on X and W, and so at best can construct only imperfect estimates of risks, $\hat{R}_i$. For instance, in our running police example, officers might base search decisions in part on subtle behavioral cues that are not documented in the data; and analysts typically would not know whether individuals who were *not* searched were in fact carrying a weapon. In Section 5, we discuss various approaches to estimating risk in light of these challenges. Whatever approach one adopts, the accuracy of one's conclusions depends critically on the accuracy of one's estimates of risk. To address this issue, we develop a novel sensitivity analysis (implemented in the `rar` R package on CRAN) that draws inspiration from methods for sensitivity analysis popular in the causal inference literature (Jung et al., 2020; Rosenbaum and Rubin, 1983a)—though we emphasize that our framework does not itself involve estimating causal effects.

To start, we assume that there is a known constant $\epsilon \geq 0$ such that

$$\frac{1}{n} \sum_{i=1}^{n} |R_i - \hat{R}_i| \leq \epsilon; \tag{6}$$

that is, that the true risks and the estimated risks differ on average by at most ϵ.[5] Given this assumption, the goal of our sensitivity analysis is to understand how different our estimate of disparate impact, $\hat{\beta}_j - \hat{\beta}_1$, could be if we had access to the true risk R_i instead of the estimated risk $\hat{R}_i$. In practice, an analyst would examine the robustness of conclusions to different choices of ϵ.

We could attempt to bound our estimate of disparate impact by searching over all possible choices of R_i satisfying the constraint in Eq. (6). But such an approach is overly conservative, as the observed data themselves rule out possible values of R_i. In our policing example, for instance, we expect the average true risk of searched individuals to approximately equal the proportion of searched individuals carrying a weapon. Since, by assumption, the observed data tell us about this latter quantity, they constrain the possible risk distributions we must consider.

To formalize our approach, a key quantity to consider is the average (true) risk on each stratum,

$$\frac{1}{n_{j,a}} \sum_{i \in \mathcal{S}_{j,a}} R_i, \tag{7}$$

where

$$\mathcal{S}_{j,a} = \{\, i : C_i = j, A_i = a \,\}, \qquad n_{j,a} = |\mathcal{S}_{j,a}|,$$

i.e., $\mathcal{S}_{j,a}$ denotes the stratum containing all those individuals in group j for whom the decision maker took action a, and $n_{j,a}$ is its size. By the law of large numbers, the average risk on stratum $\mathcal{S}_{j,a}$ is approximated by $\mathbb{E}[R \mid C = j, A = a]$. Further,

$$\begin{aligned}
\mathbb{E}[R \mid C = j, A = a] &= \mathbb{E}[\mathbb{E}[W \mid X] \mid C = j, A = a] \\
&= \mathbb{E}[\mathbb{E}[W \mid X, U] \mid C = j, A = a] \\
&= \mathbb{E}[W \mid C = j, A = a],
\end{aligned}$$

where the first equality follows by definition, the second by the fact that $W \perp\!\!\!\perp U \mid X$, and the third by the law of iterated expectations, since C is a function of X and A is a function of X and U. When $A = 1$, by our

[5] These differences could arise due to either measurement or modelling error. However, since measurement error tends to vanish as the sample size increases, it is less of a concern, and is better captured through bootstrapped bounds on the sensitivity analysis, as we discuss below.

assumption in Section 3, the above quantity is identified by data available to the analyst. In particular, in our policing example, a consistent estimator of $\mathbb{E}[W \mid C = j, A = 1]$ is the proportion of searched individuals possessing a weapon, among those individuals belonging to group j. We denote by ρ_j the analyst's estimate of $\mathbb{E}[W \mid C = j, A = 1]$ based on the observed data.[6]

Finally, putting the pieces together, for each particular j^* we seek the largest and smallest values of $\hat{\beta}_{j^*} - \hat{\beta}_1$, as estimated with R_i, subject to two key constraints: (1) the average absolute deviation between R_i and $\hat{R}_i$ is less than ϵ; and (2) R_i is consistent with the observed data on each stratum $\mathcal{S}_{j,1}$. This is succinctly expressed as an optimization problem (henceforth **"the base problem"**):

$$\begin{aligned} \underset{\mathbf{R}\in\mathbb{R}^n}{\text{Optimize}} \quad & \hat{\beta}_{j^*} - \hat{\beta}_1 \\ \text{s.t. } \tfrac{1}{n}\textstyle\sum_{i=1}^n |R_i - \hat{R}_i| &\le \epsilon \\ \tfrac{1}{n_{j,1}}\textstyle\sum_{i\in\mathcal{S}_{j,1}} R_i &= \rho_j,\ (j = 1,\ldots,m) \\ R_i &\le u_i,\ \ (i = 1,\ldots,n) \\ R_i &\ge \ell_i,\ \ (i = 1,\ldots,n) \end{aligned} \tag{8}$$

where, by "Optimize," we simply mean both maximize and minimize, and $\hat{R}_i$ and ρ_j are fixed by the data as discussed above.[7] The additional upper and lower bounds on R_i arise from the fact that the risks must be probabilities, i.e., $0 \le R_i \le 1$, though our approach accommodates tighter, individual-specific bounds $\ell_i \le R_i \le u_i$ set by the analyst.

Below, we give a polynomial time approximation to the solution of the base problem when the constraints are "sortable." Sortability is a mild hypothesis which holds in the typical case, $\ell_i = 0$ and $u_i = 1$ for all i. It also holds if, e.g., ℓ_i and u_i differ from $\hat{R}_i$ by the additive constant Γ on the log odds scale, as is common in many kinds of sensitivity analysis (e.g., Rosenbaum, 2002).

Definition 4.1 (Sortability). We say that the *constraints are sortable* if there exists a permutation π of $\{1,\ldots,n\}$ such that $\hat{R}_{\pi(1)} \le \ldots \le \hat{R}_{\pi(n)}$, $\ell_{\pi(1)} \le \ldots \le \ell_{\pi(n)}$, and $u_{\pi(1)} \le \ldots \le u_{\pi(n)}$.

There are two main obstacles to solving the base problem above. First, the dimension of the search space is $n - m$, where n is the number of observations and m is the number of groups. Second, the regression coefficients $\hat{\beta}_j$ are non-convex functions of the true risk vector $\mathbf{R} = (R_1,\ldots,R_n)$. Addressing these twin challenges involves a detailed analysis of the underlying geometry of the problem. Here we present an outline of our approach that illustrates the key ideas, with the full exposition in Appendix B.

It is useful to think of the base problem as an optimization over subproblems (henceforth **"the parameterized problem"**) where the average (true) risk is assumed known and equal to some τ_j for *unobserved* strata $\mathcal{S}_{j,0}$ as well. Using the closed form of OLS regression, we can show that solving the parameterized problem reduces to—although is not equivalent to—solving the following optimization problem (henceforth **"the simplified problem"**):

$$\begin{aligned} \underset{\mathbf{R}\in\mathbb{R}^n}{\text{Optimize}} \quad & \frac{1}{n}\sum_{i=1}^n R_i^2 \\ \text{s.t. } \tfrac{1}{n}\textstyle\sum_{i=1}^n |R_i - \hat{R}_i| &\le \epsilon, \\ \tfrac{1}{n_{j,1}}\cdot\textstyle\sum_{i\in\mathcal{S}_{j,1}} R_i &= \rho_j,\ (j = 1,\ldots,m) \\ \tfrac{1}{n_{j,0}}\cdot\textstyle\sum_{i\in\mathcal{S}_{j,0}} R_i &= \tau_j,\ (j = 1,\ldots,m) \\ R_i &\le u_i,\ \ (i = 1,\ldots,n) \\ R_i &\ge \ell_i.\ \ (i = 1,\ldots,n) \end{aligned} \tag{9}$$

[6] One might alternatively estimate $\mathbb{E}[W \mid C = j, A = 1]$ by the average *estimated* risks on the stratum, namely $\frac{1}{|\mathcal{S}_{j,1}|}\cdot\sum_{i\in\mathcal{S}_{j,1}} \hat{R}_i$. This alternative, which we adopt in our empirical analysis in Section 5, has the interpretive advantage that our sensitivity analysis exactly recovers the correct answer when $\epsilon = 0$.

[7] It may seem inappropriate to fix the average risks on the observed stratum $\mathcal{S}_{j,1}$ to be exactly ρ_j, since ρ_j is estimated with error. One could instead allow ρ_j to vary within some range (e.g., a 95% confidence interval), solving the parameterized problem for various ρ_j as well as τ_j, as detailed below. However, doing so doubles the dimension of the search space, which is computationally costly, and does not fully take advantage of available information about the distribution of the estimation error. A more principled and computationally feasible approach to dealing with estimation error is to bootstrap confidence intervals for the bounds following Zhao et al. (2019), as we do in Section 5.2 below, where ρ_j is re-estimated in each bootstrapped resample, and the base problem is solved with corresponding equality constraints.

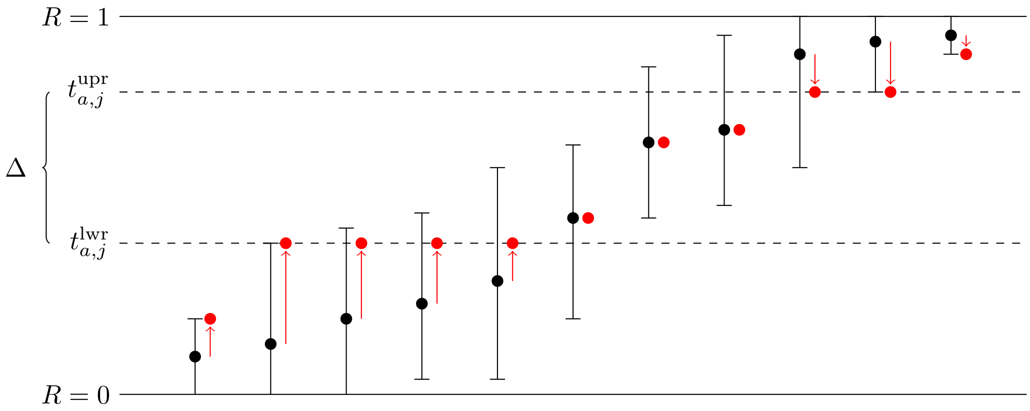


Figure 1: *Minimization normal form. Black dots represent estimated risks $\hat{R}_i$, the error bars represent the allowable range $[\ell_i, u_i]$, and red dots indicate the corresponding value of the normal form vector R_i. Here Δ denotes the gap between the upper and lower thresholds.*

An efficient method for solving the simplified problem then yields an efficient method for solving the base problem simply by searching over this much smaller m-dimensional parameter space of $(\tau_1, \ldots, \tau_m)$. In practice, m typically equals two or three, and so this final optimization step can be solved either using a grid search, as we do in Section 5, or using other optimization techniques for low-dimensional non-convex problems.

The objective of the simplified problem, $\frac{1}{n}\sum_{i=1}^{n} R_i^2$, is convex, as are the other constraints on $\mathbf{R}$, and so, in principle, minimization in the simplified problem can be achieved efficiently and accurately using interior-point methods (Boyd and Vandenberghe, 2004). However, by carefully examining the KKT conditions (Karush, 1939; Kuhn and Tucker, 1951), it is possible to see that the minimizing risk vector must adhere to a special form, which we term "minimization normal form." Specifically, there exist thresholds $t^{\text{lwr}}_{j,a}$ and $t^{\text{upr}}_{j,a}$ for each stratum such that the minimizing R_i is obtained by "pulling up" low estimated risks to $t^{\text{lwr}}_{j,a}$, "pulling down" high estimated risks to $t^{\text{upr}}_{j,a}$, and leaving unchanged intermediate risks. An example of a risk vector in minimization normal form is shown in Figure 1.

This observation reduces the dimension of the search space from $n - 2m$ to $2m$; that is, by the above, one need only search over the $2m$ thresholds. One can, however, do better. A careful examination of the KKT conditions reveals that while the thresholds $t^{\text{lwr}}_{j,a}$ and $t^{\text{upr}}_{j,a}$ may vary by stratum, the gap between them $\Delta = t^{\text{upr}}_{j,a} - t^{\text{lwr}}_{j,a}$ must be the same across strata. Moreover, once Δ is fixed, the thresholds $t^{\text{lwr}}_{j,a}$ and $t^{\text{upr}}_{j,a}$ are determined by the strata-specific average risks ρ_j and τ_j. As a result, finding the minimizing risk vector for the simplified problem reduces to optimizing over a single parameter, Δ, which, in our setting, can be done very efficiently. In particular, we give an $O(n \cdot \log(m))$ algorithm that finds the exact solution.

Maximization requires more care, since the maximization problem is non-convex, and the general problem of maximizing a quadratic objective over a convex set is NP-hard (Sahni, 1974). Using a similar, but more delicate, version of the techniques used to minimize the simplified problem, we give an $O(m \cdot (\epsilon/\gamma)^2 + n)$-time algorithm for approximating the maximum of the simplified problem. The key steps described above for optimizing the base problem are given in Algorithm 1.

Putting these results together, we have the following theorem, the proof of which is given in Appendix B.

Theorem 4.2. *Suppose the constraints are sortable. Consider Algorithm 1 with step-size parameters η and γ and maximum average absolute deviation ϵ. Let δ^* denote the true optimum of the base problem, and let $\delta^\dagger$ denote the value of the objective returned by Algorithm 1. Then, there exists a constant c_0 such that Algorithm 1 runs in time at most*

$$c_0 \left(n \cdot \log(n) + m/\gamma^2\right) \cdot \eta^{-m}.$$

Moreover, there exists a constant c_1 and a problem-specific constant $V(\epsilon)$ (i.e., depending on $\hat{R}_i$, ℓ_i, u_i, and

Algorithm 1 (Sensitivity analysis)

Input: The estimated risk vector $\hat{\mathbf{R}}$, the lower and upper bounds $\boldsymbol{\ell}$ and $\mathbf{u}$, the average absolute difference ϵ, and the step-size parameters η and γ.
Output: The minimum and maximum values of $\hat{\beta}_j - \hat{\beta}_1$ for all groups $j = 1, \ldots, m$.

```
Set $\rho_j \leftarrow \frac{1}{|\mathcal{S}_{j,1}|} \cdot \sum_{i \in \mathcal{S}_{j,1}} \hat{R}_i$ for $j = 1, \ldots, m$
Generate a grid of $(\tau_1, \ldots, \tau_m)$ with step size $\eta$
repeat
    Calculate $\mathbf{R}^*_{\min}$ minimizing the simplified problem with $\rho_1, \ldots, \rho_m$ and $\tau_1, \ldots, \tau_m$
    Calculate $\mathbf{R}^*_{\max}$ maximizing the simplified problem to within $2nm\gamma$ with $\rho_1, \ldots, \rho_m$ and $\tau_1, \ldots, \tau_m$
    Calculate $\hat{\beta}_j - \hat{\beta}_1$ using both $\mathbf{R}^*_{\min}$ and $\mathbf{R}^*_{\max}$
until The grid of $(\tau_1, \ldots, \tau_m)$ is exhausted
return The largest and smallest differences $\beta_j - \beta_1$ observed for all $j = 2, \ldots, m$
```

ϵ) such that

$$|\delta^* - \delta^\dagger| \leq \frac{c_1 \cdot m(\eta + \gamma)}{V(\epsilon)^2}.$$

We give the definition of $V(\epsilon)$ in Eq. (32) in Appendix B. Roughly, $V(\epsilon)$ is related to the within-group variance of the estimated risks V,

$$V = \frac{1}{n} \sum_{j=1}^{m} n_j \cdot \textsc{Var}\left((\hat{R}_i)_{i \in \mathcal{G}_j} \right).$$

In particular, $V(\epsilon) \geq V - 4\epsilon$. Moreover, $V(\epsilon)$ is positive whenever the base problem has a finite solution. In our principal application, it takes a few seconds on standard hardware for the algorithm to run for 100 values of ϵ with $m = 3$ race groups and $n > 10^6$ observations with negligible approximation error.

As we noted previously, our approach to sensitivity analysis for risk-adjusted regression is related to sensitivity analysis methods developed in the causal inference literature, particularly to recent work on L_2 sensitivity bounds (Huang and Pimentel, 2022; Zhang and Zhao, 2022). Our approach, however, is distinct in important ways that make it uniquely advantageous in our setting. First, methods for sensitivity analysis in the causal inference literature that can be adapted to our setting often involve a large number of parameters—making them difficult to credibly calibrate—or require parametric assumptions on the form of the confounding itself (e.g., Cinelli and Hazlett, 2020; Jung et al., 2020; Rosenbaum, 2002; Rosenbaum and Rubin, 1983a). In contrast, the single parameter in our sensitivity analysis—ϵ, the average absolute deviation between the true and estimated risks—is straightforward to reason about and explain, particularly to policymakers or other non-technical stakeholders.

Further, most past sensitivity methods that do involve a single parameter would, if suitably adapted to our setting, be parameterized in terms of an L_∞-bound Γ on the log odds ratio between true and estimated risks (e.g., Dorn and Guo, 2022; Zhao et al., 2019). However, in realistic models of confounding, the change in the odds ratio typically cannot be bounded in this way for all units: some—potentially very small—set of units will have an estimated risk of approximately 0 but a true risk of approximately 1 or *vice versa*. As a result, one may need to set Γ to be very large to strictly satisfy the assumptions of the approach, yielding sensitivity bounds that are, in practice, quite conservative. More recent methods involving L_2 bounds avoid this problem (Huang and Pimentel, 2022; Zhang and Zhao, 2022), but would not yield tight bounds on the estimand of interest in our setting.

5 An Application to Policing

We apply our approach above to investigate the "stop-and-frisk" practices of the New York City Police Department (NYPD).[8] Police officers in the United States may stop and question pedestrians if they have

[8] Reproduction materials for this analysis are available at https://github.com/jgaeb/rar-repro.

"reasonable and articulable" suspicion of criminal activity; officers may additionally conduct a "frisk" (i.e., a brief pat-down of one's outer garments) if they believe the stopped individual is carrying a weapon. Although a policy of stopping and frisking individuals is not inherently illegal, in *Floyd v. City of New York* (2013), a federal district court ruled that the NYPD carried out such stops with racial animus, violating the Equal Protection Clause of the Fourteenth Amendment.

The court in *Floyd* was interested in assessing claims of disparate treatment; here we re-analyze the data with a focus on disparate impact. We specifically consider frisk decisions, as they have a clear goal of ensuring officer safety by recovering weapons, and a well-measured outcome—whether a weapon was in fact found. We study 2.2 million pedestrian stops that occurred between 2008 and 2011. For each stop, we have detailed information on the date, time, and location of the stop; the demographics of the stopped individual (e.g., age, gender, and race); the suspected crime; the reasons prompting the stop (e.g., "furtive movements" or "suspicious bulge"); and additional circumstances surrounding the stop (e.g., evasive responses to questioning, witness reports, or evidence of criminal activity in the vicinity).[9]

To start, we note that 1.7% of frisks turn up a weapon. White pedestrians are frisked in 44% of police stops, whereas Black and Hispanic pedestrians are frisked in 57% and 58% of stops, respectively, a 13–14 percentage point gap. These raw disparities are computed without adjusting for any potentially explanatory variables, and so represent an extreme case of omitted-variable bias. At the other extreme is the "kitchen sink regression", which adjusts for all observed pre-frisk covariates in a standard linear probability model. In this case, stopped Black and Hispanic pedestrians are 3–4 percentage points more likely to be frisked, relative to white pedestrians with similar recorded characteristics. These kitchen-sink disparities are suggestive of disparate treatment (and similar evidence was indeed presented to the court in *Floyd* to support such an allegation), but they may understate the extent to which the policy imposes an unjustified disparate impact on racial minorities, due to included-variable bias.

5.1 Estimating risk-adjusted disparities

The key ingredient in applying risk-adjusted regression is estimating the risk $R = \Pr(W = 1 \mid X)$, as in Eq. (3), where W indicates whether a stopped individual has a weapon and X is the information available to the officer when making their frisk decision. There are, however, two challenges to estimating this quantity. First, the information available to officers typically differs from that recorded in administrative datasets available to analysts. Second, W remains unobserved for individuals who were not frisked. These challenges can be mitigated, but not eliminated. Our approach is thus to estimate these risks as best we can, and then gauge the robustness of our results to estimation error. For ease of exposition, we adopt a simple estimation strategy, regressing weapon possession (W) on the covariates $\tilde{X}$ in the recorded data, fit on the subset of individuals who were, in reality, frisked.[10] These estimates suffer from some degree of omitted-variable bias and selection effects, an issue we return to in our sensitivity analysis below.

We start by dividing the original stops into two subsets: (1) the approximately 1 million stops that occurred in 2008 and 2009, which we use to train a risk model; and (2) the remaining 1.2 million stops that occurred in 2010 and 2011, to which we apply the fitted risk model to estimate disparities in frisk decisions. To estimate risk of weapon possession, we use gradient boosted decision trees, a non-linear model popular in the machine learning community for its predictive performance, restricting to stops in the first subset of data in which a frisk was conducted. Predictive performance and model checks presented in Figure 14 in Appendix E indicate that the model yields predictions with reasonable performance and that predictions are well-calibrated across groups. We use this fitted model to produce an estimate of *ex ante* risk $\hat{R}_i$ for every pedestrian stopped in the second subset of the data, including those who were not frisked. These risk estimates are shown in Figures 12 and 13

Figure 2 shows frisk rates as a function of estimated risk, disaggregated by race. At every level of risk,

[9] This information is recorded in a standardized way on UF-250 forms that officers are required to complete after each stop. A copy of the form can be found online.

[10] Our approach to estimating risk is not the only one, or even the best in any given instance. For example, one could, in theory, partially address the selection problem by requiring that some subset of individuals be randomly frisked, as is sometimes done at roadside checkpoints and airports—though this raises legal, ethical, and logistical difficulties. Alternatively, one could look for officers who frisk almost everyone they stop—similar to recent approaches for studying judicial decision-making (e.g., Dobbie et al., 2018; Kleinberg et al., 2018). Both of these alternatives come with their own limitations, meaning that sensitivity analysis is still critical to credible inference.

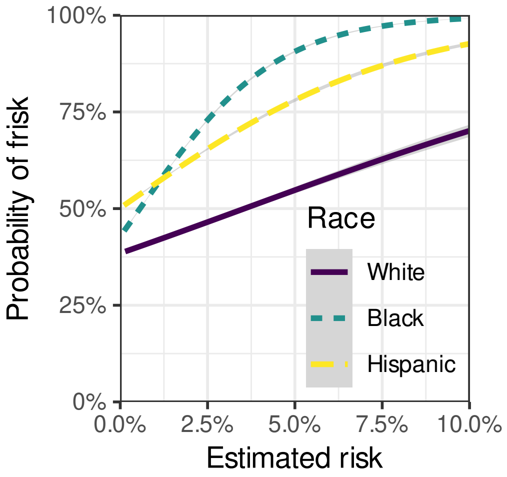


Figure 2: *Frisk rates vs. risk, as estimated via logistic regression curves fit separately for each race group. Across risk levels, stopped Black and Hispanic pedestrians are frisked substantially more frequently than comparably risky white individuals, indicative of disparate impact.*

stopped Black and Hispanic pedestrians are frisked at a much higher rate than stopped white individuals, a gap that is suggestive of disparate impact in frisk decisions. To add quantitative detail to these qualitative results, we now compute risk-adjusted disparities, fitting the linear probability model in Eq. (5) on the second half of the NYPD data, computing estimates for the Black-white disparity $\hat{\beta}_{\text{Black}} - \hat{\beta}_{\text{White}}$ and Hispanic-white disparity $\hat{\beta}_{\text{Hispanic}} - \hat{\beta}_{\text{White}}$. Figure 3 shows the results, together with the raw disparities and those estimated from a kitchen-sink model. We find that stopped Black and Hispanic pedestrians were about 15 percentage points more likely to be frisked than white pedestrians who were equally likely to be carrying a weapon. Further, the risk-adjusted disparities are in fact *greater* than the raw disparities in frisk rates. To understand why, we note that stopped white pedestrians were, on average, more likely to be carrying a weapon yet less likely to be frisked than racial minorities; as a result, the risk-adjusted gap in frisk rates is even larger than the raw, unadjusted gap. Finally, we see that the kitchen-sink regression dramatically underestimates the extent of disparate impact faced by minorities. In this case, the kitchen-sink model adjusts for a variety of features—including whether the stopped individual made "furtive movements"—that are correlated with race but are poor predictors of weapon possession, skewing estimates of disparate impact.[11]

5.2 Sensitivity analysis

The disparities computed above aim to circumvent included-variable bias by adjusting for each individual's estimated risk. But our estimates of risk may themselves be skewed if officers observe factors that are predictive of risk but are not recorded in our data. To account for this possibility, Figure 4 displays worst-case bounds on our estimate of disparate impact as a function of the mismatch between true and estimated risk—operationalized in terms of ϵ, as defined in Eq. (6); see Appendix C for further details on how the bounds are computed. To ease interpretation, the horizontal axis in the plot is expressed in terms of the *relative* average absolute difference between the true and estimated risks: ϵ divided by the overall weapon recovery rate among frisked individuals (1.7%). Our analysis shows that large risk-adjusted disparities remain even if we allow the true risk to differ considerably from our risk estimates. In particular, we would find large disparate impacts for both Black and Hispanic pedestrians even if the true risks differed from our risk

[11] We exclude hair color and eye color from the kitchen-sink model, since these are obvious proxies for race that are effectively unrelated to risk, and would thus be excluded in most traditional legal and statistical analyses of discrimination. As we would expect, including these variables as controls exacerbates the problem of included-variable bias, but our results show that such bias can occur even if obviously problematic variables are excluded.

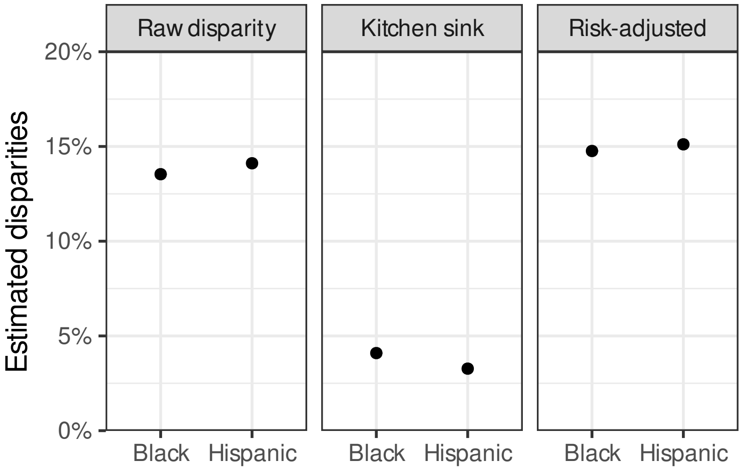


Figure 3: *Racial gaps in frisk rates adjusting for different sets of covariates, where the y-axis shows the percentage point difference relative to stopped white individuals. The left panel shows the raw disparities in frisk rates. As a measure of discrimination, raw disparities suffer from omitted-variable bias: there may, in theory, be legitimate reasons why Black and Hispanic pedestrians are more likely to be frisked. The middle panel shows the estimated race effects in a kitchen-sink regression, adjusting for all pre-frisk covariates. These estimates suffer from included-variable bias because they adjust for features that are correlated with race but unrelated to risk. The right panel shows the results of our risk-adjusted regression, adjusting exclusively for estimated risk of weapon possession. In all cases, estimated standard errors are less than 0.2 percentage points, and so are not visible in the plot.*

estimates by 50% of the base rate.

It is impossible to know the precise extent to which our risk estimates differ from the true risk. To understand the plausible magnitude of the discrepancy, we conduct a simulation in which we remove a key set of risk-relevant covariates from the data, estimate risk based on the reduced information, and then measure differences between the original and new risk estimates. Specifically, we remove variables listed in two sections of the UF-250 stop forms that describe the "circumstances" prompting the encounter. These sections each consist of 10 binary variables—including, for example, "fits description", "actions indicative of casing", and "changing direction at sight of officer"—that are crucial for establishing the legal basis of the stop. We then compute the average absolute difference between our original risk estimates based on the full, uncensored data and the risk estimates based on the redacted data.[12] We find that this value is 0.7 percentage points—or 44% of the base rate—which we take as an estimate of ϵ in a scenario with severe unobserved confounding. As shown in Figure 4, this level of error (indicated by the dashed vertical lines) would yield an estimate of disparate impact that is, at a minimum, greater than 7.5 percentage points for both Black and Hispanic individuals. This sensitivity analysis suggests our results are robust to substantial unobserved confounding.

6 Discussion

We have sought to develop a simple, intuitive framework for addressing the most serious concerns of included- and omitted-variable bias in disparate impact studies. On a detailed dataset of police stops, we found that these concerns are more than hypothetical possibilities. In particular, regressions that adjust for all available covariates—in line with common legal and statistical convention—can substantially skew estimates

[12] As above, we use gradient boosted decision trees to estimate the probability of frisk conditional on the remaining, uncensored covariates.

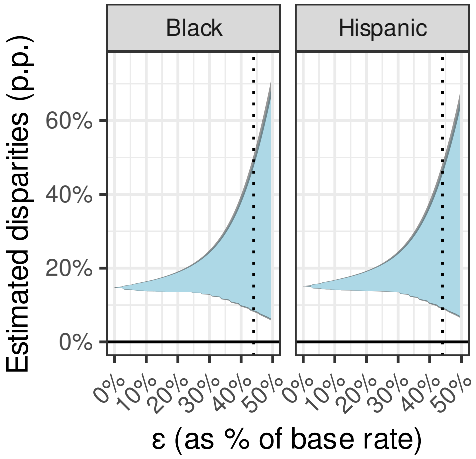


Figure 4: *Sensitivity of the risk-adjusted disparities in frisk decisions to mismeasurement of risk. The blue bands bound our estimates of disparate impact as a function of the average absolute difference between the true and estimated risks ϵ, relative to the base rate (1.7%). The dotted line at 44% ($\epsilon = 0.7$ p.p.) corresponds to a simulated situation with severe confounding. The small grey bands along the top and bottom of the blue bands represent 95% percentile bootstrapped confidence intervals ($N = 1000$; Zhao et al., 2019). The step size for the grid search over average risks of the unobserved strata was $\eta = 0.05$ p.p. for all groups, and $\gamma = 0.01$ p.p. was the approximation parameter for the maximization routine; see Appendix B.*

of disparate impact.

Our risk-adjusted regression framework is subject to some important limitations. First, our method requires access to an outcome to estimate risk. In some instances, this information is not available to analysts. In other cases, it is not even clear how to rigorously define the relevant outcome. For example, in college admissions, decision makers often care about multiple factors in hard to quantify ways (Garg et al., 2021; Grossman et al., 2023b). Second, to credibly estimate risk, analysts need sufficiently rich covariate data. Our sensitivity analysis helps mitigate omitted-variable bias, but it cannot replace better data. Third, and relatedly, analysts may not have access to race or other relevant protected characteristics, complicating the analysis. There has, however, been recent progress on estimating disparities using auxiliary data sources (e.g., Kallus et al., 2022; McCartan et al., 2023). Finally, we have modeled decisions using linear probability models with constant slope across groups. It is straightforward to estimate risk-adjusted disparities with more flexible decision models. However, relaxing this assumption makes it harder to gauge the sensitivity of the inferred disparities to inaccurate risk estimates.

Throughout our analysis, we have estimated "disparate impact" by a regression coefficient on protected-group identity in a model that adjusts for estimated risk. This procedure is analogous to current practice in discrimination studies, where we simply replace the usual set of control variables with a single variable capturing risk. As seen by our formalization of disparate impact in Eq. (4), we are effectively measuring a particular weighted average of differences in decision rates across individuals of similar risk. While intuitively reasonable, this definition raises subtle questions of law and policy.

Consider, for example, Figure 2, where we plot race-specific frisk rates as a function of risk. Stopped Black and Hispanic pedestrians are frisked more often than stopped white pedestrians at every level of risk. As a result, one would find that racial minorities face disparate impact regardless of how one averages across risk levels; the precise number might change, but the qualitative conclusion would remain the same. However, comparing Black and Hispanic pedestrians, the direction of the disparity depends on the risk level one considers.[13] Low-risk Hispanic individuals are frisked more often than low-risk Black individuals, but

[13] Such a comparison between minority groups is unusual in disparate impact cases, but it illustrates the underlying theoretical

high-risk Black individuals are frisked more often than their high-risk Hispanic counterparts. Consequently, a conclusion of disparate impact between Black and Hispanic pedestrians would depend heavily on the precise definition applied. The analysis is further complicated if the risk distributions differ substantially between groups. If, hypothetically, stopped Hispanic pedestrians were mostly low-risk and stopped Black pedestrians mostly high-risk, majorities of both groups could argue that they were treated more harshly than members of the other group who were equally likely to be carrying a weapon.[14]

The crossing of risk curves that we see in Figure 2 is a potentially widespread phenomenon, and, to our knowledge, disparate impact law has not yet resolved the underlying conceptual ambiguity it invokes. Many discussions of disparate impact tacitly assume that policies either consistently harm or help groups defined by protected traits. Such thinking can be seen in the original *Griggs* ruling, where the Supreme Court aimed to proscribe policies that acted as "built-in headwinds" for racial minorities. But, formally, disparate impact law concerns facially race-neutral policies, not intentional discrimination, and there is no theoretical or empirical guarantee that such policies will adversely impact all members of a particular group.

A related issue is the extent to which concern for unjustified disparities compels decision makers to act optimally. This concern connects most closely to the third element of legal tests of disparate impact outlined in Section 2 above, namely, the existence of a less discriminatory alternative policy. For example, Figure 2 suggests that officers are only marginally responsive to risk, with the lowest-risk individuals still frisked more than 40% of the time. If, instead, officers frisked only the people with high probability of carrying a weapon, they could frisk far fewer individuals—and, in particular, far fewer minority individuals—while recovering almost the same number of weapons (Goel et al., 2016a). A more efficient frisk strategy could thus reduce the burdens of policing on racial minorities while still maintaining public safety. Such efficiency is indeed one of the aims of statistical risk assessment tools that are now used in the criminal justice system and beyond to guide high-stakes decisions (Chouldechova et al., 2018; Corbett-Davies et al., 2023; Goel et al., 2021; Monahan and Skeem, 2016; Shroff, 2017). If these tools are shown to reduce racial disparities, are policymakers obliged—legally or ethically—to adopt them? The role of efficiency in disparate impact claims has largely gone unanswered by the courts, adding yet another subtlety to defining and measuring disparities. Researchers have only recently taken up these questions (Bohren et al., 2022; Corbett-Davies et al., 2023; Elzayn et al., 2023; Grossman et al., 2023a; Raghavan, 2023).

By foregrounding the role of risk in understanding disparities, we have aimed to clarify some of the thorny conceptual issues at the heart of disparate impact analysis. While there are still important unresolved questions, we believe that our statistical approach provides practitioners with a tractable way to assess disparities in many domains while avoiding some important pitfalls of traditional methods. Looking forward, we hope this work spurs further theoretical and empirical research on discrimination at the intersection of statistics, economics, law, and public policy.

issue.

[14] Some scholars have similarly investigated interactions between race and other decision-making criteria. For example, Espenshade et al. (2004) find that preferences for underrepresented minorities in college admissions is greatest for applicants with SAT scores in the 1200–1300 range, and the effect is attenuated at lower scores. That analysis, however, found no score ranges where minority applicants faced an absolute disadvantage relative to white applicants with equal scores. We do not know of any research that has found a change in the direction of the disparities like we see between Black and Hispanic pedestrians in Figure 2.

A Formalization of Griggs

We begin by proving the results from Section 2.2 for the parametric formulation of the problem, and then use the parametric results to prove the non-parametric claims.

A.1 Linear Regression Formulation

Recall that the OLS coefficient estimates are given in this case by

$$\left(\sum_{i=1}^{n}\begin{bmatrix} 1 & X_{0,i} & X_{1,i} & X_{0,i}\cdot X_{1,i} & C_i \\ X_{0,i} & X_{0,i}^2 & X_{0,i}\cdot X_{1,i} & X_{0,i}^2\cdot X_{1,i} & X_{0,i}\cdot C_i \\ X_{1,i} & X_{0,i}\cdot X_{1,i} & X_{1,i}^2 & X_{0,i}\cdot X_{1,i}^2 & X_{1,i}\cdot C_i \\ X_{0,i}\cdot X_{1,i} & X_{0,i}^2\cdot X_{1,i} & X_{0,i}\cdot X_{1,i}^2 & X_{0,i}^2\cdot X_{1,i}^2 & X_{0,i}\cdot X_{1,i}\cdot C_i \\ C_i & X_{0,i}\cdot C_i & X_{1,i}\cdot C_i & X_{0,i}\cdot X_{1,i}\cdot C_i & C_i^2 \end{bmatrix}\right)^{-1}\left(\sum_{i=1}^{n}\begin{bmatrix} A_i \\ X_{0,i}\cdot A_i \\ X_{1,i}\cdot A_i \\ X_{0,i}\cdot X_{1,i}\cdot A_i \\ C_i\cdot A_i \end{bmatrix}\right)$$

Note that because of the matrix inversion, we can divide both sums by n. Therefore, applying the continuous mapping theorem and the strong law of large numbers, we see that this converges to

$$\begin{bmatrix} 1 & \mathbb{E}[E] & \mathbb{E}[H] & \mathbb{E}[E\cdot H] & \mathbb{E}[C] \\ \mathbb{E}[E] & \mathbb{E}[E^2] & \mathbb{E}[E\cdot H] & \mathbb{E}[E^2\cdot H] & \mathbb{E}[E\cdot C] \\ \mathbb{E}[H] & \mathbb{E}[E\cdot H] & \mathbb{E}[H^2] & \mathbb{E}[E\cdot H^2] & \mathbb{E}[H\cdot C] \\ \mathbb{E}[E\cdot H] & \mathbb{E}[E^2\cdot H] & \mathbb{E}[E\cdot H^2] & \mathbb{E}[E^2\cdot H^2] & \mathbb{E}[E\cdot H\cdot C] \\ \mathbb{E}[C] & \mathbb{E}[E\cdot C] & \mathbb{E}[H\cdot C] & \mathbb{E}[E\cdot H\cdot C] & \mathbb{E}[C^2] \end{bmatrix}^{-1}\begin{bmatrix} \mathbb{E}[A] \\ \mathbb{E}[E\cdot A] \\ \mathbb{E}[H\cdot A] \\ \mathbb{E}[E\cdot H\cdot A] \\ \mathbb{E}[C\cdot A] \end{bmatrix}.$$

Now, recalling the joint and marginal distributions of these variables, we have that

$$\mathbb{E}[E]=\frac{1}{2},\qquad \mathbb{E}[H]=\frac{1}{2},\qquad \mathbb{E}[C]=\frac{1}{2}.$$

Moreover,

$$\mathbb{E}[E^2]=\frac{1}{3},\qquad \mathbb{E}[H^2]=\frac{1}{2},\qquad \mathbb{E}[C^2]=\frac{1}{2}.$$

By the definition of covariance,

$$\rho=\frac{\mathbb{E}[H\cdot C]-\mathbb{E}[H]\cdot\mathbb{E}[C]}{\sqrt{\textsc{Var}(H)\cdot\textsc{Var}(C)}}=\frac{\mathbb{E}[H\cdot C]-\frac{1}{4}}{\sqrt{\frac{1}{4}\cdot\frac{1}{4}}}=4\cdot\mathbb{E}[H\cdot C]-1,$$

whence $\mathbb{E}[H\cdot C]=\frac{\rho+1}{4}$. By independence,

$$\begin{aligned} \mathbb{E}[E\cdot H]&=\mathbb{E}[E]\cdot\mathbb{E}[H]=\frac{1}{4}, & \mathbb{E}[E\cdot C]&=\mathbb{E}[E]\cdot\mathbb{E}[C]=\frac{1}{4}, \\ \mathbb{E}[E^2\cdot H]&=\mathbb{E}[E^2]\cdot\mathbb{E}[H]=\frac{1}{6}, & \mathbb{E}[E\cdot H^2]&=\mathbb{E}[E\cdot H]=\frac{1}{4}, \\ \mathbb{E}[E^2\cdot H^2]&=\mathbb{E}[E^2\cdot H]=\frac{1}{6}, & \mathbb{E}[E\cdot H\cdot C]&=\mathbb{E}[E]\cdot\mathbb{E}[H\cdot C]=\frac{\rho+1}{8}, \end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E}[\mathbb{E}[A \mid E, H]] \\ &= \mathbb{E}[E \cdot H] \\ &= \frac{1}{4},\end{aligned} \qquad \begin{aligned}\mathbb{E}[E \cdot A] &= \mathbb{E}[\mathbb{E}[E \cdot A \mid E, H]] \\ &= \mathbb{E}[E^2 \cdot H] \\ &= \frac{1}{6},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[H \cdot A] &= \mathbb{E}[\mathbb{E}[H \cdot A \mid E, H]] \\ &= \mathbb{E}[E \cdot H^2] \\ &= \frac{1}{4},\end{aligned} \qquad \begin{aligned}\mathbb{E}[E \cdot H \cdot A] &= \mathbb{E}[\mathbb{E}[E \cdot H \cdot A \mid E, H]] \\ &= \mathbb{E}[E^2 \cdot H^2] \\ &= \frac{1}{6},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[C \cdot A] &= \mathbb{E}[\mathbb{E}[C \cdot A \mid E, H, C]] \\ &= \mathbb{E}[E \cdot H \cdot C] \\ &= \frac{\rho + 1}{8}.\end{aligned}$$

Thus, we have that

$$\begin{bmatrix} 1 & \mathbb{E}[E] & \mathbb{E}[H] & \mathbb{E}[E \cdot H] & \mathbb{E}[C] \\ \mathbb{E}[E] & \mathbb{E}[E^2] & \mathbb{E}[E \cdot H] & \mathbb{E}[E^2 \cdot H] & \mathbb{E}[E \cdot C] \\ \mathbb{E}[H] & \mathbb{E}[E \cdot H] & \mathbb{E}[H^2] & \mathbb{E}[E \cdot H^2] & \mathbb{E}[H \cdot C] \\ \mathbb{E}[E \cdot H] & \mathbb{E}[E^2 \cdot H] & \mathbb{E}[E \cdot H^2] & \mathbb{E}[E^2 \cdot H^2] & \mathbb{E}[E \cdot H \cdot C] \\ \mathbb{E}[C] & \mathbb{E}[E \cdot C] & \mathbb{E}[H \cdot C] & \mathbb{E}[E \cdot H \cdot C] & \mathbb{E}[C^2] \end{bmatrix}^{-1}$$

equals

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{6} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \frac{\rho+1}{4} \\ \frac{1}{4} & \frac{1}{6} & \frac{1}{4} & \frac{1}{6} & \frac{\rho+1}{8} \\ \frac{1}{2} & \frac{1}{4} & \frac{\rho+1}{4} & \frac{\rho+1}{8} & \frac{1}{2} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{7\rho+9}{\rho+1} & -12 & -\frac{6\rho+8}{\rho+1} & 12 & -\frac{2}{\rho+1} \\ -12 & 24 & 12 & -24 & 0 \\ -\frac{6\rho+8}{\rho+1} & 12 & \frac{12*\rho^2-16}{\rho^2-1} & -24 & \frac{4\rho}{\rho^2-1} \\ 12 & -24 & -24 & 48 & 0 \\ -\frac{2}{\rho+1} & 0 & \frac{4\rho}{\rho^2-1} & 0 & -\frac{4}{\rho^2-1} \end{bmatrix}$$

and so the OLS regression coefficients converge almost surely to

$$\begin{bmatrix} \frac{7\rho+9}{\rho+1} & -12 & -\frac{6\rho+8}{\rho+1} & 12 & -\frac{2}{\rho+1} \\ -12 & 24 & 12 & -24 & 0 \\ -\frac{6\rho+8}{\rho+1} & 12 & \frac{12*\rho^2-16}{\rho^2-1} & -24 & \frac{4\rho}{\rho^2-1} \\ 12 & -24 & -24 & 48 & 0 \\ -\frac{2}{\rho+1} & 0 & \frac{4\rho}{\rho^2-1} & 0 & -\frac{4}{\rho^2-1} \end{bmatrix} \begin{bmatrix} \frac{1}{4} \\ \frac{1}{6} \\ \frac{1}{4} \\ \frac{1}{6} \\ \frac{\rho+1}{8} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

On the other hand, if the analyst were to drop high school graduation from their regression, then we would have that the regression coefficients converge almost surely to

$$\begin{bmatrix} 1 & \mathbb{E}[E] & \mathbb{E}[C] \\ \mathbb{E}[E] & \mathbb{E}[E^2] & \mathbb{E}[E \cdot C] \\ \mathbb{E}[C] & \mathbb{E}[E \cdot C] & \mathbb{E}[C^2] \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[A] \\ \mathbb{E}[E \cdot A] \\ \mathbb{E}[C \cdot A] \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{3} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{4} \\ \frac{1}{6} \\ \frac{\rho+1}{8} \end{bmatrix} = \begin{bmatrix} -\frac{\rho}{4} \\ \frac{1}{2} \\ \frac{\rho}{2} \end{bmatrix}.$$

A.2 Nonparametric Formulation

Recall the two different non-parametric estimands:

$$\mathbb{E}[\mathbb{E}[A \mid E, H, C = 1] - \mathbb{E}[A \mid E, H, C = 0]]$$

and

$$\mathbb{E}[\mathbb{E}[A \mid E, C = 1] - \mathbb{E}[A \mid E, C = 0]].$$

We note that in the first case, since $\mathbb{E}[A \mid E, H, C] = E \cdot H$, we have that

$$\mathbb{E}[\mathbb{E}[A \mid E, H, C = 1] - \mathbb{E}[A \mid E, H, C = 0]] = \mathbb{E}[E \cdot H - E \cdot H] = 0.$$

In the second case, using the expectations calculated above, we have that

$$\begin{aligned} \mathbb{E}[A \mid E, C = c] &= \mathbb{E}[A \mid E, H = 1, C = c] \cdot \Pr(H = 1 \mid E, C = c) \\ &\qquad + \mathbb{E}[A \mid E, H = 0, C = c] \cdot \Pr(H = 0 \mid E, C = c) \\ &= E \cdot \Pr(H = 1 \mid E, C = c) \\ &= E \cdot \Pr(H = 1 \mid C = c) \\ &= E \cdot \frac{\mathbb{E}[H \cdot C]}{\mathbb{E}[C]} \\ &= E \cdot \frac{\frac{1-(-1)^c \cdot \rho}{4}}{\frac{1}{2}} \\ &= E \cdot \frac{1 - (-1)^c \cdot \rho}{2}. \end{aligned}$$

Here we have used the fact that $A = 0$ when $H = 0$ in the second equality, the independence of H from E and C in the third, and various expectations calculated above in the remaining equalities. Thus, it follows that

$$\mathbb{E}[\mathbb{E}[A \mid E, C = 1] - \mathbb{E}[A \mid E, C = 0]] = \mathbb{E}\left[E \cdot \frac{1+\rho}{2} - E \cdot \frac{1-\rho}{2}\right] = \mathbb{E}[E \cdot \rho] = \frac{\rho}{2}.$$

This exactly equals β_C' in the second regression.

B Sensitivity Analysis

In what follows, we formally justify the sensitivity analysis laid out in Section 4, deriving algorithms to efficiently solve the optimization problem it gives rise to and rigorously verifying their correctness and runtimes. This discussion is structured in three parts. In the first part, by introducing a new low-dimensional family of parameters, we show how to reduce the base optimization problem of Eq. (8) into the simplified problem of Eq. (9). In the second and third parts, we show how to efficiently find the minimum and maximum values of the simplified optimization problem, respectively.

Remark B.1. In the main text, to for expositional clarity, we parameterized our optimization problems in terms of *average* risks. However, it is mathematically more natural to work with the *total* risks. Therefore, throughout this appendix we let ρ_j and τ_j denote the *total* risk in each stratum, i.e., we rewrite the relevant constraints of the base problem and the simplified problem as follows:

$$\sum_{i \in \mathcal{S}_{j,1}} R_i = \rho_j, \qquad \sum_{i \in \mathcal{S}_{j,0}} R_i = \tau_j.$$

B.1 Simplifying the objective

To simplify the optimization problem in Eq. (8), we can avail ourselves of the fact that OLS has a closed-form solution. We augment the notation in Section 4 as follows:

$$\mathcal{G}_j = \mathcal{S}_{j,0} \cup \mathcal{S}_{j,1} = \{\, i : C_i = j \,\},$$

i.e., $\mathcal{G}_j$ denotes the set of indices belonging to the j-th group.

Lemma B.2. *Let $n_j = |\mathcal{G}_j| > 0$ denote the number of individuals in the j-th group; let σ_j denote the search rate in group j, i.e., $\sigma_j = \frac{1}{n_j}\sum_{i \in \mathcal{G}_j} A_i$; and let r_j and t_j denote*

$$r_j = \sum_{i \in \mathcal{S}_{j,1}} R_i, \qquad t_j = \sum_{i \in \mathcal{S}_{j,0}} R_i,$$

i.e., the total risk of the searched and unsearched individuals belonging to the j-th group, respectively.

If R_i is constant within each of the m groups, then the OLS estimate need not exist due to collinearity. Otherwise, the OLS estimate of the coefficients $\hat{\boldsymbol{\beta}}$ in Eq. (5) satisfies

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \\ 0 \end{bmatrix} + \frac{\frac{1}{n}\left[\sum_{j=1}^m \sigma_j \cdot t_j - (1-\sigma_j)\cdot r_j\right]}{\frac{1}{n}\left[\sum_{i=1}^n R_i^2 - \sum_{j=1}^m \frac{(r_j+t_j)^2}{n_j}\right]} \begin{bmatrix} \frac{r_1+t_1}{n_1} \\ \vdots \\ \frac{r_m+t_m}{n_m} \\ -1 \end{bmatrix}. \tag{10}$$

Proof. Recall that our design matrix and outcome variables take the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}(C_1 = 1) & \dots & \mathbf{1}(C_1 = m) & R_1 \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}(C_n = 1) & \dots & \mathbf{1}(C_n = m) & R_n \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}.$$

The OLS estimate of the coefficients $\hat{\boldsymbol{\beta}}$ is then simply $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.

Note that by definition, we have that

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} n_1 & & & \sum_{i\in\mathcal{G}_1} R_i \\ & \ddots & & \vdots \\ & & n_m & \sum_{i\in\mathcal{G}_m} R_i \\ \sum_{i\in\mathcal{G}_1} R_i & \dots & \sum_{i\in\mathcal{G}_m} R_i & \sum_{i=1}^n R_i^2 \end{bmatrix} = \begin{bmatrix} n_1 & & & r_1+t_1 \\ & \ddots & & \vdots \\ & & n_m & r_m+t_m \\ r_1+t_1 & \dots & r_m+t_m & \sum_{i=1}^n R_i^2 \end{bmatrix}.$$

Similarly,

$$\mathbf{X}^\top\mathbf{Y} = \begin{bmatrix} \sum_{i\in\mathcal{G}_1} A_i \\ \vdots \\ \sum_{i\in\mathcal{G}_m} A_i \\ \sum_{i=1}^n R_i \cdot A_i \end{bmatrix} = \begin{bmatrix} n_1\cdot\sigma_1 \\ \vdots \\ n_m\cdot\sigma_m \\ \sum_{j=1}^m r_j \end{bmatrix}.$$

Note that $\mathbf{X}^\top\mathbf{X}$ has the form

$$\begin{bmatrix} D & v^\top \\ v & \sum_{i=1}^n R_i^2, \end{bmatrix}$$

where D is diagonal. Then, we see that since D is invertible, it follows that $\mathbf{X}^\top\mathbf{X}$ is invertible if the Schur complement of the D is invertible, i.e., non-zero; that is, if

$$\sum_{i=1}^{n} R_i^2 - \sum_{j=1}^{m} \frac{(r_j+t_j)^2}{n_j} \neq 0.$$

Recalling the definitions of r_j and t_j, we see that the left-hand expression equals

$$\sum_{j=1}^{m} \sum_{i \in \mathcal{G}_j} \left(R_i - \frac{r_j + t_j}{n_j} \right)^2 = \sum_{j=1}^{m} n_j \cdot \textsc{Var}\left((R_i)_{i \in \mathcal{G}_j} \right).$$

Now, the variance of a set is only zero if all the elements of the set are equal, i.e., if for all $i \in \mathcal{G}_j$, R_i equals the average risk in that group, $\frac{r_j+t_j}{n_j}$.

Assume this is not the case, i.e., that $\mathbf{X}^\top\mathbf{X}$ is invertible. Since $\mathbf{X}^\top\mathbf{X}$ is an arrowhead matrix, we can invert it using the Sherman-Morrison Formula. In particular, we obtain that

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & & & \\ & \ddots & & \\ & & \frac{1}{n_m} & \\ & & & 0 \end{bmatrix} + \frac{1}{\sum R_i^2 - \sum_{j=1}^{m} \frac{(r_j+t_j)^2}{n_j}} \begin{bmatrix} \frac{r_1+t_1}{n_1} \\ \vdots \\ \frac{r_m+t_m}{n_m} \\ -1 \end{bmatrix} \begin{bmatrix} \frac{r_1+t_1}{n_1} \\ \vdots \\ \frac{r_m+t_m}{n_m} \\ -1 \end{bmatrix}^\top . \tag{11}$$

Combining this with the expression derived for $\mathbf{X}^\top\mathbf{Y}$ above and dividing the numerator and denominator by n yields that

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \\ 0 \end{bmatrix} + \frac{\frac{1}{n}\left[\sum_{j=1}^{m} \sigma_j \cdot t_j - (1-\sigma_j) \cdot r_j \right]}{\frac{1}{n}\left[\sum R_i^2 - \sum_{j=1}^{m} \frac{(r_j+t_j)^2}{n_j} \right]} \begin{bmatrix} \frac{r_1+t_1}{n_1} \\ \vdots \\ \frac{r_m+t_m}{n_m} \\ -1 \end{bmatrix} . \tag{12}$$

□

We note that Lemma B.2 provides the following useful condition on when the base problem has a meaningful solution:

Corollary B.3. *The maximum and minimum values of the base problem are both finite if and only if it is not the case that $\ell_i \leq \rho_j \leq u_i$ for all $i \in \mathcal{G}_j$ and $j = 1, \ldots, m$, and in addition*

$$\frac{1}{n} \sum_{j=1}^{m} \sum_{i \in \mathcal{G}_j} \left| \hat{R}_i - \frac{\rho_j}{n_{j,1}} \right| \leq \epsilon.$$

Proof. The proof follows immediately from noting that since the averages of the risks on the observed strata $\mathcal{S}_{j,1}$ are fixed at ρ_j, the risks for the whole j-th group can only all be equal if the risks of the observed *and* the unobserved individuals are uniformly equal to ρ_j. □

The following corollary completes the simplification of the optimization problem in Eq. (8) in the case where the average true risk in the unobserved strata is also assumed known (**"the parameterized problem"** of Section 4),[15] i.e., of

$$\begin{array}{rl} \underset{\mathbf{R} \in \mathbb{R}^n}{\text{Optimize}} & \hat{\beta}_{j^*} - \hat{\beta}_1 \\ \text{s.t. } \frac{1}{n}\sum_{i=1}^{n} & |R_i - \hat{R}_i| \leq \epsilon, \\ & \sum_{i \in \mathcal{S}_{j,1}} R_i = \rho_j, \ (j = 1, \ldots, m) \\ & \sum_{i \in \mathcal{S}_{j,0}} R_i = \tau_j, \ (j = 1, \ldots, m) \\ & R_i \leq u_i, \ \ (i = 1, \ldots, n) \\ & R_i \geq \ell_i. \ \ (i = 1, \ldots, n) \end{array} \tag{13}$$

[15] See Remark B.1 above for the slight difference in notation between here and Section 4.

Corollary B.4. *The optimal solutions of the parameterized problem in Eq. (13), if they are both well-defined, are also the optimal solutions of the simplified problem in Eq. (9).*

Proof. Let $\hat{\boldsymbol{\beta}}$ be as in Eq. (5), and r_j and t_j as in Lemma B.2.[16] Then, it follows directly from Eq. (10) that we have that for all $\mathbf{R}$,

$$\hat{\beta}_j - \hat{\beta}_1 = \sigma_j - \sigma_1 + \frac{\frac{1}{n}\left[\sum_{j=1}^{m} \sigma_j \cdot t_j - (1-\sigma_j)\cdot r_j\right]}{\frac{1}{n}\left[\sum_{i=1}^{n} R_i^2 - \sum_{j=1}^{m} \frac{(r_j+t_j)^2}{n_j}\right]} \left(\frac{r_j + t_j}{n_j} - \frac{r_1 + t_1}{n_1}\right).$$

Since the constraints of the parameterized problem fix $r_j = \rho_j$ and $t_j = \tau_j$ for all feasible $\mathbf{R}$, the objective function of the parameterized problem has the form

$$a + \frac{b}{x - c}, \tag{14}$$

where

$$\begin{aligned}
a &= \sigma_j - \sigma_1, \\
b &= \frac{1}{n}\left[\sum_{j=1}^{n} \sigma_j \cdot \tau_j - (1-\sigma_j)\cdot \rho_j\right] \cdot \left(\frac{\rho_j + \tau_j}{n_j} - \frac{\rho_1 + \tau_1}{n_1}\right), \\
c &= \frac{1}{n}\sum_{j=1}^{m} \frac{(\rho_j + \tau_j)^2}{n_j}, \\
x &= \frac{1}{n}\sum_{i=1}^{n} R_i^2.
\end{aligned}$$

If $b = 0$, then the objective does not depend on $\mathbf{R}$, and so there is nothing to prove. Otherwise, by continuity and the fact that the feasible region is convex, and hence connected, and compact, the range of the objective in Eq. (9) over the feasible region is a closed interval I. As shown in the proof of Lemma B.2 and Corollary B.3, $x \geq c$, and $x > c$ if the optima are both well-defined. Consequently, we can assume the interval is strictly positive, i.e., $I \subseteq (c, \infty)$. The derivative of Eq. (14) exists everywhere on I and is, moreover, non-zero and continuous, and so $a + \frac{b}{x-c}$ is strictly monotone on I. Therefore, the maximum and minimum must occur at the endpoints, i.e., at the maximum and minimum values of x. □

We note that since b is not guaranteed to be positive, the maximum of the simplified problem in Eq. (9) does not necessarily correspond to the maximum of the base problem in Eq. (8) and *vice versa*.

B.2 Minimization

In the case of minimization, the simplified problem in Eq. (9) defines a convex optimization problem, which, as noted in Section 4 above, can consequently be solved efficiently using standard interior-point methods. However, these results can be improved in a problem-specific way to yield a simpler, and, in practice, more computationally tractable method of finding exact solutions. Moreover, they parallel the solution method for maximization, which is not a convex problem, and hence to which standard interior point methods cannot be applied.

B.2.1 Optimizing over a single stratum

We begin by solving a simpler version of the problem, where we restrict to a single stratum. In the case of minimization—as in the case of maximization, as we show below—solutions to the minimization problem must be in a certain normal form. While searching over all risk vectors is prohibitively difficult, searching over normal form risk vectors can be done straightforwardly in linear time.

[16] In particular, strictly speaking, $\hat{\boldsymbol{\beta}}$, r_j, and t_j should be written $\hat{\boldsymbol{\beta}}(\mathbf{R})$, $r_j(\mathbf{R})$, and $t_j(\mathbf{R})$, since they depend on $\mathbf{R}$.

Definition B.5 (Minimization normal form). We say that a risk vector $\mathbf{R}$ is in *minimization normal form* if there exist thresholds $t_{a,j}^{\text{lwr}}$ and $t_{a,j}^{\text{upr}}$ such that for $i \in \mathcal{S}_{j,a}$,

$$R_i = \begin{cases} \min(u_i, t_{a,j}^{\text{lwr}}) & \hat{R}_i \leq t_{a,j}^{\text{lwr}}, \\ \hat{R}_i & t_{a,j}^{\text{lwr}} \leq \hat{R}_i \leq t_{a,j}^{\text{upr}}, \\ \max(\ell_i, t_{a,j}^{\text{upr}}) & t_{a,j}^{\text{upr}} \leq \hat{R}_i, \end{cases}$$

and, in addition, $t_{a,j}^{\text{upr}} - t_{a,j}^{\text{lwr}}$ is equal to some Δ for all $a = 0, 1$ and $j = 1, \ldots, m$.

In the context of a single stratum, we refer to t^{upr} and t^{lwr} for notational simplicity. We also assume that $\mathcal{S}_{j,a} = \{1, \ldots, n\}$ for the same reason.

Minimization normal form results from "pushing" $\hat{R}_i$ up to t^{lwr} or down to t^{upr} as far as possible. An example of a risk vector in minimization normal form is shown in Figure 1. As the name suggests, risk vectors minimizing the objective must be in minimization normal form.

Lemma B.6. *Consider the optimization problem (**"the single stratum minimization problem"**)*

$$\begin{aligned} \underset{\mathbf{R} \in \mathbb{R}^n}{\text{Minimize}} \quad & \frac{1}{n} \sum_{i=1}^{n} R_i^2 \\ \text{s.t. } & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} |R_i - \hat{R}_i| \leq \epsilon, \\ & \textstyle\sum_{i=1}^{n} R_i = \mu, \\ & R_i \leq u_i, \ (\forall i) \\ & R_i \geq \ell_i. \ (\forall i) \end{aligned} \tag{15}$$

Suppose that $\mathbf{R}^*$ is a minimizer. Then, $\mathbf{R}^*$ is unique and in minimization normal form. Moreover, $\mathbf{R}^*$ exhausts the L_1 budget, in the sense that either $\frac{1}{n}\|\mathbf{R}^* - \hat{\mathbf{R}}\|_1 = \epsilon$ or $t^{\text{lwr}} = t^{\text{upr}}$.

Proof. The proof proceeds by examining the first-order KKT conditions to derive weak necessary conditions that solutions must satisfy. These conditions are then strengthened to minimization normal form by directly comparing the objective function at different points satisfying the weak conditions. Finally, we show that the minimum is the point satisfying the "strengthened" conditions which is most distant from $\hat{\mathbf{R}}$, i.e., that exhausts the budget.

The problem is simplified by rewriting the first constraint as a collection of linear (and hence everywhere differentiable) constraints. In particular, $\frac{1}{n}\|\mathbf{R} - \hat{\mathbf{R}}\|_1 \leq \epsilon$ is equivalent to the 2^n constraints of the form $\mathbf{S}^\top(\mathbf{R} - \hat{\mathbf{R}}) \leq \epsilon \cdot n$ for all $\mathbf{S} \in \{-1, 1\}^n$. We assume throughout, without loss of generality, that $\ell_i < u_i$.

Weak conditions Let $\mathbf{e}_i$ denote the i-th standard basis vector and $\mathbf{1} = \sum_{i=1}^{n} \mathbf{e}_i$. Note that:

- The gradient of the objective is $\frac{2}{n} \cdot \mathbf{R}$;
- The gradient of $\mathbf{S}_k^\top(\mathbf{R} - \hat{\mathbf{R}}) - \frac{\epsilon}{n}$, where $\{\mathbf{S}_1, \ldots, \mathbf{S}_{2^n}\} = \{-1, 1\}^n$, is $\mathbf{S}_k$;
- The gradient of $\sum_{i=1}^{n} R_i$ is $\mathbf{1}$;
- The gradient of $R_i - u_i$ is $\mathbf{e}_i$ $\mathbf{e}_i$;
- The gradient of $\ell_i - R_i$ is $-\mathbf{e}_i$.

The first-order necessary KKT conditions therefore require that[17]

$$2 \cdot \mathbf{R}^* - \lambda \cdot \mathbf{1} + \sum_{i=1}^{n} \mu_{0,i} \cdot \mathbf{e}_i - \sum_{i=1}^{n} \mu_{1,i} \cdot \mathbf{e}_i + \sum_{k=1}^{2^n} \nu_k \cdot \mathbf{S}_k = 0 \tag{16}$$

[17] For notational simplicity, we have multiplied through by n and absorbed constants into the corresponding Lagrange multipliers.

for some arbitrary λ, and for non-negative $\mu_{0,i}$, $\mu_{1,i}$, and ν_k for $k = 1, \ldots, 2^n$ satisfying complementary slackness, i.e., such that $\mu_{0,i} \cdot (R_i^* - u_i) = 0$, $\mu_{1,i} \cdot (\ell_i - R_i^*) = 0$, and $\nu_k \cdot (\mathbf{S}_k^\top (\mathbf{R}^* - \hat{\mathbf{R}}) - n \cdot \epsilon) = 0$ for all $i = 1, \ldots, n$ and $k = 1, \ldots, 2^n$.

These consequences allow us to prove the following weak characterization of $\mathbf{R}^*$. Let $\Delta = \sum_{k=1}^{2^n} \nu_k$. Then, for all i,

$$R_i^* \in \left\{ \hat{R}_i, u_i, \ell_i, \frac{1}{2} \cdot [\lambda - \Delta], \frac{1}{2} \cdot [\lambda + \Delta] \right\}. \tag{17}$$

For, from (16), we get that

$$2 \cdot R_i^* - \lambda + \mu_{0,i} - \mu_{1,i} + \sum_{k=1}^{2^n} \nu_k \cdot S_{k,i} = 0, \tag{18}$$

where $S_{k,i}$ refers to the k-th component of $\mathbf{S}_k$. If $R_i^* = \hat{R}_i$, there is nothing to prove. Therefore, assume that $R_i^* \neq \hat{R}_i$. We note that by complementary slackness, $\nu_k > 0$ only if $S_{k,i} \cdot (R_i^* - \hat{R}_i) \geq 0$ for all $i = 1, \ldots, n$. For, supposing without loss of generality that $R_i^* > \hat{R}_i$, if $S_{k,i} \cdot (R_i^* - \hat{R}_i) < 0$, then, since $\mathbf{S}_k + 2\mathbf{e}_i \in \{-1, 1\}^n$, we would have that

$$(\mathbf{S}_k + 2\mathbf{e}_i)^\top (\mathbf{R}^* - \hat{\mathbf{R}}) = \epsilon \cdot n + 2(R_i^* - \hat{R}_i) > \epsilon \cdot n,$$

which violates the constraints. Hence $S_{k,i} = S_{k',i}$ for all k and k' such that $\nu_k, \nu_{k'} > 0$. Therefore, our expression simplifies to

$$2 \cdot R_i^* - \lambda + \mu_{0,i} - \mu_{1,i} + \Delta \cdot s = 0,$$

where $s = \pm 1$. By complementary slackness, if $\mu_{0,i} > 0$, then $R_i^* = u_i$; if $\mu_{1,i} > 0$, then $R_i^* = \ell_i$. Therefore, we need only consider the case where $\mu_{0,i} = \mu_{1,i} = 0$, i.e., where

$$2 \cdot R_i^* - \lambda + s \cdot \Delta = 0,$$

whence $R_i^* = \frac{1}{2}[\lambda + s \cdot \Delta]$. Therefore, for all i, (17) holds.

Strengthening conditions Next, we strengthen (17) to minimization normal form. First, we show that if $R_{i_0}^* > \hat{R}_{i_0}$ and $R_{i_0}^* > R_{i_1}^*$, then $R_{i_1}^* = u_{i_1}$. For, suppose not. Then, there exists some $\delta > 0$ such that (1) $R_{i_0}^* - \delta > \hat{R}_{i_0}$, (2) $R_{i_1}^* + \delta < u_{i_1}$, and (3) $R_{i_0}^* - \delta > R_{i_1}^*$. Define $\mathbf{R}' = \mathbf{R}^* + \delta \cdot (\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$. Then, we note that

$$\begin{aligned}
\|\mathbf{R}' - \hat{\mathbf{R}}\|_1 - \|\mathbf{R}^* - \hat{\mathbf{R}}\|_1 &= \sum_{i=1}^n |R_i' - \hat{R}_i| - |R_i^* - \hat{R}_i| \\
&= |R_{i_0}' - \hat{R}_{i_0}| - |R_{i_0}^* - \hat{R}_{i_0}| + |R_{i_1}' - \hat{R}_{i_1}| - |R_{i_1}^* - \hat{R}_{i_1}| \\
&= (R_{i_0}^* - \delta - \hat{R}_{i_0}) - (R_{i_0}^* - \hat{R}_{i_0}) + |R_{i_1}^* + \delta - \hat{R}_{i_1}| - |R_{i_1}^* - \hat{R}_{i_1}| \\
&\leq -\delta + |R_{i_1}^* - \hat{R}_{i_1}| + \delta - |R_{i_1}^* - \hat{R}_{i_1}| \\
&= 0,
\end{aligned}$$

so $\mathbf{R}'$ satisfies the L_1-distance constraint. Here, we have used the fact that both $R_{i_0}^*$ and $R_{i_0}^* - \delta$ are greater than $\hat{R}_{i_0}$ in the third equality, and the triangle inequality in the inequality. The remaining constraints also hold by construction, so $\mathbf{R}'$ is feasible.

Since $\mathbf{R}'$ is feasible, we will arrive at a contradiction if we can show that it achieves a greater objective. However,

$$\begin{aligned}
\sum_{i=1}^n (R_i^*)^2 - (R_i')^2 &= (R_{i_0}^*)^2 - ((R_{i_0}^*)^2 + \delta^2 - 2 \cdot R_{i_0}^* \cdot \delta) + (R_{i_1}^*)^2 \\
&\qquad - ((R_{i_1}^*)^2 + \delta^2 + 2 \cdot (R_{i_1}^*) \cdot \delta) + \sum_{i \neq i_0, i_1} (R_i^*)^2 - (R_i^*)^2 \\
&= 2 \cdot \delta \cdot (R_{i_0}^* - R_{i_1}^* - \delta) \\
&> 0,
\end{aligned}$$

where the inequality follows from the fact that $R^*_{i_0} - \delta > R^*_{i_1}$. Therefore $\mathbf{R}^*$ is not a minimum, contrary to hypothesis. In exactly the same manner, we see that if $R^*_{i_0} < \hat{R}_i$ and $R^*_{i_1} < R^*_{i_0}$, then $R^*_{i_1} = \ell_{i_1}$.

Combining the claim—i.e., that if $R^*_{i_0} > \hat{R}_{i_0}$ and $R^*_{i_0} > R^*_{i_1}$, then $R^*_{i_1} = u_{i_1}$—with (17) yields that for all i, if $R^*_i \leq \frac{1}{2} \cdot [\lambda - \Delta]$, then $R^*_i = \min(u_i, \frac{1}{2} \cdot [\lambda - \Delta])$; if $R^*_i \geq \frac{1}{2} \cdot [\lambda + \Delta]$, then $R^*_i = \max(\ell_i, \frac{1}{2} \cdot [\lambda + \Delta])$; and otherwise $R^*_i = \hat{R}_i$. Taking $t^{\text{lwr}} = \frac{1}{2} \cdot [\lambda - \Delta]$ and $t^{\text{upr}} = \frac{1}{2} \cdot [\lambda + \Delta]$ gives the result.

Uniqueness and budget exhaustion If the budget is not exhausted, then, by complementary slackness, $\nu_k = 0$ for all $k = 1, \ldots, 2^k$, whence $\Delta = 0$. This immediately implies that $t^{\text{lwr}} = t^{\text{upr}} = \frac{1}{2}\lambda$.

To see uniqueness, suppose that $\mathbf{R}^{(0)}$ and $\mathbf{R}^{(1)}$ are two distinct minima, both in normal form, with corresponding thresholds t^{lwr}_k and t^{upr}_k for $k = 0, 1$. We note that since $\sum_{i=1}^n R^{(0)}_i = \sum_{i=1}^n R^{(1)}_i = \mu$, we must have, without loss of generality, that $t^{\text{lwr}}_0 \leq t^{\text{lwr}}_1$ and $t^{\text{upr}}_0 \geq t^{\text{upr}}_1$, with at least one of the two inequalities strict. However, it follows that $\|\mathbf{R}^{(0)} - \hat{\mathbf{R}}\|_1 < \|\mathbf{R}^{(1)} - \hat{\mathbf{R}}\|_1$, which is impossible by the preceding paragraph. □

Lemma B.6 suggests a natural algorithm for solving the optimization problem in Eq. (9): we can sweep over all possible risk vectors in minimization normal form simply by sweeping over t^{lwr}. In particular, as we increase t^{lwr}, the sum constraint—i.e., that $\sum_{i=1}^n R_i = \mu$—forces t^{upr} to decrease so as to counterbalance it exactly.

More precisely, consider an index i to be "active" if one of the thresholds is between ℓ_i and u_i. If k^{lwr} represents the number of indices that are active because of t^{lwr}—and k^{upr} is also defined accordingly—then, locally, if we increase t^{lwr} at unit rate, the sum of the risks increases at the rate of k^{lwr}. Therefore, the rates of increase of the two thresholds, r^{lwr} and r^{upr}, must satisfy[18]

$$r^{\text{lwr}} \cdot k^{\text{lwr}} = r^{\text{upr}} \cdot k^{\text{upr}}. \tag{19}$$

In particular, we know that we have reached the minimum—and hence solved the optimization problem—once either the budget has been exhausted or the two thresholds have become equal. For fixed r^{lwr} and r^{upr}, this would be straightforward to calculate; however, these rates can, in principle, change whenever k^{upr} and k^{lwr} change. We call these change points "distinguished point"—that is, points at which a threshold t equals ℓ_i, $\hat{R}_i$, or u_i for some i. Therefore, the algorithm consists of repeating the following steps:

1. Determining the "next" distinguished points that either the lower or upper threshold will reach.

2. Checking whether the budget will be exhausted before that point is reached.

3. If not, advancing to that point and recalculating k^{lwr} or k^{upr} as appropriate.

In addition, since $\sum_{i=1}^n R_i$ may not equal μ, a "preprocessing" step, where t^{lwr} is increased or t^{upr} is decreased may be needed to find the "initial" risk vector in minimization normal form satisfying the sum constraint.

We make two observations about the informal algorithm sketch given above needed for its extension to the case of multiple strata. First, we note that every risk vector $\mathbf{R}^*$ in minimization normal form is the minimizer for the corresponding single stratum minimization problem with $\epsilon = \frac{1}{n}\|\mathbf{R}^* - \hat{\mathbf{R}}\|_1$ and $\mu = \sum_{i=1}^n R^*_i$. Consequently, although intended to find the minimizer for a specific ϵ, the algorithm sketched above actually sweeps over the minimizers for *all possible* ϵ. As a result, it is not any harder to solve a single instance of the single stratum minimization problem than it is to solve it generally for all possible ϵ, given some fixed μ. Secondly, while the natural way to parameterize the risk vectors in the discussion above is in terms of t^{lwr}, when optimizing over multiple strata, it is actually more natural to reparameterize the risk vectors in terms of the gap $\Delta = t^{\text{upr}} - t^{\text{lwr}}$. If in addition to Eq. (19) we impose the condition that $r^{\text{lwr}} + r^{\text{upr}} = 1$, so that Δ decreases at a constant rate, then we have that[19]

$$r^{\text{lwr}} = \frac{k^{\text{upr}}}{k^{\text{lwr}} + k^{\text{upr}}}, \qquad r^{\text{upr}} = \frac{k^{\text{lwr}}}{k^{\text{lwr}} + k^{\text{upr}}}. \tag{20}$$

[18] If either k^{lwr} and k^{upr}, then the corresponding threshold can be changed at any rate without affecting the sum constraint. However, this is because changing that threshold in this case does not actually change the risk vector, since no indices are active, and so we can ignore this case in the subsequent discussion.

[19] We note that by Lemma B.6, for every δ there is a unique $\mathbf{R}_\delta$, but that the reverse is not necessarily true, since there may not be any i such that $R_i = t^{\text{lwr}}$ or t^{upr}, in which case different thresholds can be chosen without altering the underlying risk vector.

To solve the single stratum minimization problem across the whole range of possible ϵ, let $\mathbf{R}_\Delta$ be the unique solution we can associate with a given gap Δ between t^{lwr} and t^{upr}.[20] Then, we can study two functions:

$$\epsilon(\Delta) = \|\mathbf{R}_\Delta - \hat{\mathbf{R}}\|_1, \qquad \Sigma(\Delta) = \sum_{i=1}^{n} R_{\Delta,i}^2 - \hat{R}_i^2.$$

By the preceding discussion we readily derive that $\epsilon(\Delta)$ is a piecewise linear function of Δ, whose slope at a given value of Δ is equal to[21]

$$K = 2 \cdot \frac{k^{\text{lwr}} \cdot k^{\text{upr}}}{k^{\text{lwr}} + k^{\text{upr}}},$$

and $\Sigma(\Delta)$ is a piecewise quadratic function of Δ satisfying

$$\begin{aligned}
\Sigma(\Delta + t) &= \Sigma(\Delta) + k^{\text{lwr}} \cdot [(t^{\text{lwr}} + r^{\text{lwr}} \cdot t)^2 - (t^{\text{lwr}})^2] + k^{\text{upr}} \cdot [(t^{\text{upr}} - r^{\text{upr}} \cdot t)^2 - (t^{\text{lwr}})^2] \\
&= \Sigma(\Delta) + k^{\text{lwr}} \cdot r^{\text{lwr}} \cdot t \cdot (r^{\text{lwr}} \cdot t + 2t^{\text{lwr}}) + k^{\text{upr}} \cdot r^{\text{upr}} \cdot t \cdot (r^{\text{upr}} \cdot t - 2t^{\text{upr}}) \\
&= \Sigma(\Delta) - 2 \cdot \frac{k^{\text{lwr}} \cdot k^{\text{upr}}}{k^{\text{lwr}} + k^{\text{upr}}} \cdot (t^{\text{upr}} - t^{\text{lwr}}) \cdot t + \frac{k^{\text{lwr}}(k^{\text{upr}})^2 + k^{\text{upr}}(k^{\text{lwr}})^2}{(k^{\text{lwr}} + k^{\text{upr}})^2} \cdot t^2 \\
&= \Sigma(\Delta) - K\Delta t + \frac{K}{2} t^2,
\end{aligned}$$

as long as t is sufficiently small that the number of active indices does not change. It follows that $\epsilon(\Delta)$ and $\Sigma(\Delta)$ are fully determined by the following collections:

$$\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_N), \qquad 1 = \Delta_1 < \cdots < \Delta_N = 0,$$

where Δ_k represents the k-th value of Δ at which k^{lwr} and k^{upr} change; and

$$\mathbf{K} = (K_1, \ldots, K_N), \qquad K_1 = 0, \quad K_N = 0,$$

where K_k denotes the value of K beginning at $\Delta = \Delta_k$. From $\boldsymbol{\Delta}$ and $\mathbf{K}$, the Δ^* such that $\|\mathbf{R}_{\Delta^*} - \hat{\mathbf{R}}\|_1 = \epsilon$ can be calculated in linear time. For completeness, this algorithm is given in Algorithm 2.

Thus, a complete solution for the optimization problem for a single stratum across *all possible* ϵ requires only calculating $\boldsymbol{\Delta}$ and $\mathbf{K}$. As described above, this can be calculated by moving the thresholds toward each other at the prescribed rates r^{lwr} and r^{upr}, updating the rates each time an index is activated or deactivated, until the gap between the thresholds is zero. Similar to evaluating $\epsilon(\Delta)$ and $\Sigma(\Delta)$, this can be completed in linear time, as described in Algorithm 4. (As above, we note that if $\frac{1}{n}\sum_{i=1}^n \hat{R}_i \neq \mu$, it may be necessary to do a preprocessing step, also in linear time, as shown in Algorithm 3.)

B.2.2 Optimizing over all strata

With a complete solution to the problem of minimizing the sum of squares for a single stratum, the problem of minimizing the sum of squares across all strata is straightforward. We begin by characterizing solutions to the minimization problem in the general case.

Lemma B.7. *Consider the simplified minimization problem in Eq.* (9)*. Suppose that* $\mathbf{R}^*$ *is a solution. Then, the restriction of* $\mathbf{R}^*$ *to any stratum is in minimization normal form. Moreover,* $\mathbf{R}^*$ *exhausts the* L_1 *budget, in that either* $\frac{1}{n}\|\mathbf{R} - \hat{\mathbf{R}}\|_1 = \epsilon$ *or* $t^{\text{lwr}} = t^{\text{upr}} = \mu_{a,j}$*, where* $\mu_{a,j}$ *equals either* ρ_j *or* τ_j *depending on whether* a *equals* 1 *or* 0*.*

Proof. The proof is virtually identical to the proof of Lemma B.6. The only difference is that the first-order necessary KKT conditions take the form

$$2 \cdot \mathbf{R}^* - \left(\sum_{j=1}^{m} \lambda_{0,j} \cdot \mathbf{1}_{0,j} + \lambda_{1,j} \cdot \mathbf{1}_{1,j} \right) + \sum_{i=1}^{n} \mu_{0,i} \cdot \mathbf{e}_i - \sum_{i=1}^{n} \mu_{1,i} \cdot \mathbf{e}_i + \sum_{k=1}^{2^n} \nu_k \cdot \mathbf{S}_k = 0, \tag{21}$$

[20] For every Δ there is a unique $\mathbf{R}_\Delta$. In particular, any other distinct risk vector in minimization normal form with the same gap must be strictly greater or lesser, as argued at the end of Lemma B.6, meaning that it cannot satisfy the sum constraint. However, the reverse is not necessarily true—that is, $\mathbf{R}_\Delta$ may equal $\mathbf{R}_{\Delta'}$ for $\Delta \neq \Delta'$—since achieving a given gap may require that $k^{\text{lwr}} = k^{\text{upr}} = 0$, in which case different thresholds can be chosen without altering the underlying risk vector.

[21] Or zero if both k^{lwr} and k^{upr} are zero.

where $\mathbf{1}_{a,j} = \sum_{C_i=c_j, A_i=a} \mathbf{e}_i$. Restricting to a single i gives the following minor variant of Eq. (18):

$$2 \cdot R_i^* - \lambda_{a,j} + \mu_{0,i} - \mu_{1,i} + \sum_{k=1}^{n} \nu_k \cdot S_{k,i} = 0.$$

The proof then proceeds identically; the only difference is to note that while $\lambda_{a,j}$ varies by stratum, $\Delta = \sum_{k=1}^{2^n} \nu_k$ does not. □

It follows that solving the minimization problem across all strata can be carried out in almost the same way as across a single stratum:

1. Construct piecewise linear functions $\epsilon_{a,j}(\Delta)$ and piecewise quadratic functions $\Sigma_{a,j}(\Delta)$ for each stratum;

2. Note that because the sums $\sum_{a=0}^{1} \sum_{j=1}^{m} \epsilon_{a,j}(\Delta)$ and $\sum_{a=0}^{1} \sum_{j=1}^{m} \Sigma_{a,j}(\Delta)$ are also piecewise linear and quadratic, respectively, they can also be evaluated using Algorithm 2;

3. Find Δ^* such that $\sum_{a=0}^{1} \sum_{j=1}^{m} \epsilon_{a,j}(\Delta^*) = \epsilon$, and evaluate $\sum_{a=0}^{1} \sum_{j=1}^{m} \Sigma_{a,j}(\Delta^*)$.

Consequently, beyond the machinery established in Algorithms 2, 3, and 4, we only need a way of calculating the sums of these piecewise functions. This can be carried out straightforwardly using a variation on the standard merge-sort algorithm, which we give in Algorithm 5 for completeness.

Putting this all together, we obtain the following lemma summarizing the results of this section.

Lemma B.8. *If $\boldsymbol{\ell}$, $\hat{\mathbf{R}}$, and $\boldsymbol{u}$ have been sorted, then there exists an $O(\log(m) \cdot n)$ algorithm solving the simplified minimization problem (Eq. (9)).*

Proof. The proof is straightforward. We begin by noting that Algorithm 2 requires linear time: the index i increases by one in the while loop on lines 6 through 10 on each iteration, and must be less than the length of the input. Similarly, in Algorithm 3, the while loop on lines 11 through 22 increments i^{lwr} by one on each loop, and limits it to three times the length of the input. The same analysis applies to Algorithm 4, in which either i^{upr} or i^{lwr} increases by one on each iteration. We note that the run time of Algorithm 5 is proportional to the sum of the lengths of its inputs. Using a standard divide-and-conquer approach, we can accomplish the combination of the $2m$ strata in m total calls, where each element of the input appears in $1 + \lceil \log_2(m) \rceil$ of the function calls, for a total time complexity of $O(\log(m) \cdot n)$. Putting this all together, the total complexity is $O(\log(m) \cdot n)$. □

B.3 Maximization

Even in the simplified problem in Eq. (9), maximization is not a convex problem. In general, maximizing a quadratic objective over a linearly constrained convex set is NP-hard (Sahni, 1974). However, the restricted forms of the objective and constraints allow us efficiently to solve the optimization problem exactly in the case of a single stratum and approximately in the case of multiple strata.

B.3.1 Optimizing over a single stratum

As in the case of minimization, risk vectors that maximize the single-stratum maximization problem have a special normal form.

Definition B.9 (Maximization normal form). We say that a risk vector $\mathbf{R}$ is in *maximization normal form* if there exist indices ("pivots") i^{lwr} and i^{upr} such that

$$R_i = \begin{cases} \ell_i & i < i^{\text{lwr}}, \\ \hat{R}_i & i^{\text{lwr}} < i < i^{\text{upr}}, \\ u_i & i^{\text{upr}} < i. \end{cases}$$

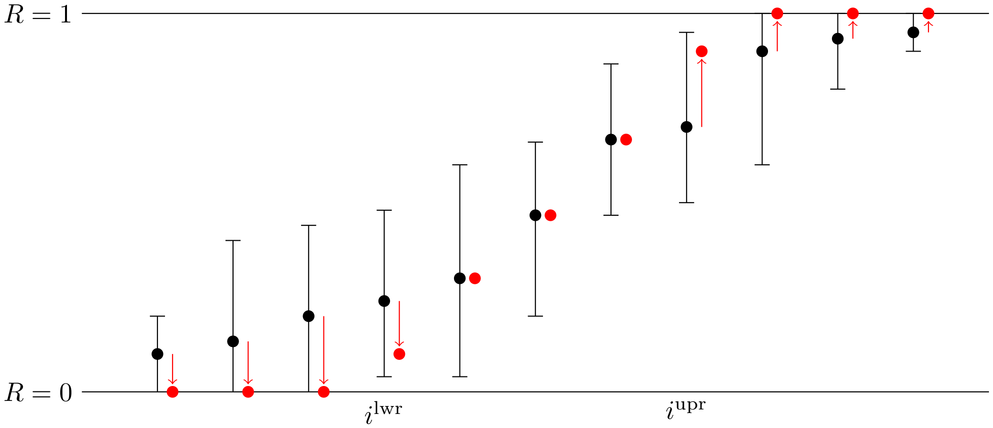


Figure 5: *Maximization normal form. Black dots represent estimated risks $\hat{R}_i$, the error bars represent the allowable range $[\ell_i, u_i]$, and red dots indicate the corresponding value of the normal form vector R_i. The pivots are indicated on the x-axis.*

We additionally require that the risk values at the pivots themselves satisfy $R_{i^{\text{lwr}}} \in [\ell_{i^{\text{lwr}}}, \hat{R}_{i^{\text{lwr}}}]$ and $R_{i^{\text{upr}}} \in [\hat{R}_{i^{\text{upr}}}, u_{i^{\text{upr}}}]$.[22,23]

Maximization normal form is similar to minimization normal form, except that instead of begin pushed toward the thresholds, values at indices *below* the lower pivot are pushed down, and values at indices *above* the upper pivot are pushed up. An illustration of maximum normal form is given in Figure 5.

An important difference from the case of minimization is that studying the KKT conditions does not yield the normal form directly; instead, it yields a weaker characterization.

Definition B.10 (Weak normal form). We say that a risk vector is in *weak normal form* if there exist thresholds t^{lwr} and t^{upr} such that

$$R_i = \begin{cases} \ell_i & R_i < t^{\text{lwr}}, \\ t^{\text{lwr}} & R_i = t^{\text{lwr}}, \\ \hat{R}_i & t^{\text{lwr}} < R_i < t^{\text{upr}}, \\ t^{\text{upr}} & R_i = t^{\text{upr}}, \\ u_i & t^{\text{upr}} < R_i, \end{cases}$$

and there exists at most one lower index i^{lwr} such that $R_{i^{\text{lwr}}} = t^{\text{lwr}}$; and similarly, there exists at most one upper index i^{upr} such that $R_{i^{\text{upr}}} = t^{\text{upr}}$

We note that, in the case of *weak* maximization normal form, the indices i^{lwr} and i^{upr} are analogous to the pivots in maximization normal form; however, they do not share the key property, which is that all of the indices $i < i^{\text{lwr}}$ index risks that have been pushed down all the way to ℓ_i and all of the indices $i > i^{\text{upr}}$ index risks that have been pushed all the way up to u_i; see Figure 6.

Lemma B.11. *Consider the single-stratum maximization problem—i.e., maximization in Eq. (15). Suppose that $\mathbf{R}^*$ is a solution. Then, $\mathbf{R}^*$ is in weak normal form. Moreover, either $\frac{1}{n}\|\mathbf{R}^* - \hat{\mathbf{R}}\|_1 = \epsilon$ or R_i^* equals ℓ_i or u_i for all but (at most) one i.*

Proof. The proof is quite similar to the proof of Lemma B.6. We note that the first-order KKT conditions

[22] Since it involves the ordering of the indices as well as the risk values, the appropriate generalization of maximization normal form is notationally awkward to express when multiple strata are involved. Fortunately, in contrast to the case of minimization, we will not have occasion to use such a generalization, and so do not give its definition.

[23] We note that we allow i^{lwr} and i^{upr} to take the values 0 or $n+1$ in addition to $\{1, \ldots, n\}$ to cover cases where, e.g., $R_i = \ell_i$ for all $i = 1, \ldots, n$.

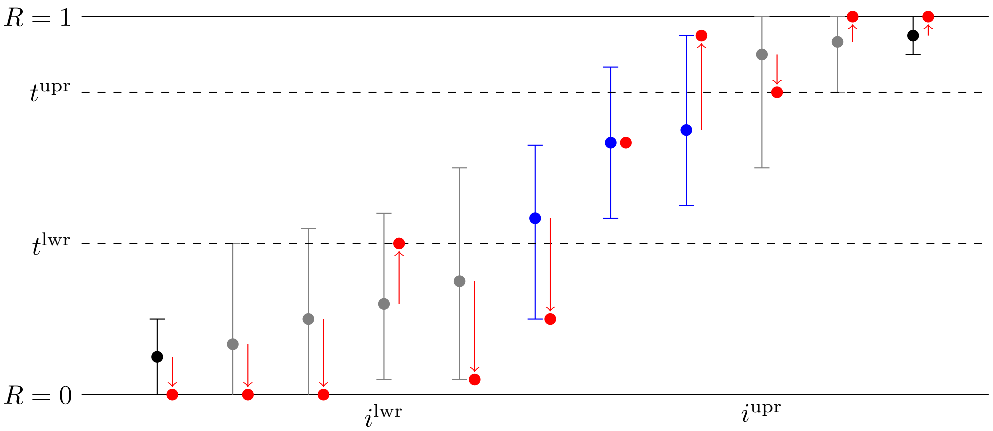


Figure 6: *Weak maximization normal form. Black, blue, and grey dots represent estimated risks $\hat{R}_i$, the error bars represent the allowable range $[\ell_i, u_i]$, and red dots indicate the corresponding value of the normal form vector R_i. The pivots are indicated on the x-axis. The risk R_i at indices where the estimated risk is between the thresholds and the bounds straddle the thresholds (shown in blue) can take on a number of values (e.g., $R_i = \hat{R}_i$ or $R_i = u_i$). Likewise, i^{lwr} and i^{upr} can occur at any index i such that $R_i = t^{\mathrm{lwr}}$ or $R_i = t^{\mathrm{upr}}$ is feasible (shown in grey). As a result, the set of risk vectors in* weak *maximization normal form is a very complex search space.*

are almost the same, namely that

$$2 \cdot \mathbf{R}^* - \lambda \cdot \mathbf{1} - \sum_{i=1}^{n} \mu_{0,i} \cdot \mathbf{e}_i + \sum_{i=1}^{n} \mu_{1,i} \cdot \mathbf{e}_i - \sum_{k=1}^{2^n} \nu_k \cdot \mathbf{S}_k = 0 \tag{22}$$

for some λ and $\mu_{0,i}, \mu_{1,i}, \nu_k \geq 0$ satisfying complementary slackness. Virtually exactly as before, it follows from this that Eq. (17) holds, i.e., that

$$R_i^* \in \left\{ \hat{R}_i, u_i, \ell_i, \frac{1}{2} \cdot [\lambda - \Delta], \frac{1}{2} \cdot [\lambda + \Delta] \right\}. \tag{23}$$

Again, as in the case of minimization, we can strengthen Eq. (23). In particular, exactly the same argument used to strengthen the weak conditions in Lemma B.6 yields that if $R_{i_0}^* > \hat{R}_{i_0}$ and $R_{i_0}^* < R_{i_1}^*$, then $R_{i_1}^* = u_{i_1}$; and, conversely, if $R_{i_0}^* < \hat{R}_{i_0}$ and $R_{i_0}^* > R_{i_1}^*$, then $R_{i_1}^* = \ell_{i_1}$. The proof is then otherwise the same as in Lemma B.6.

To see that there can be at most one i such that such that $\ell_i < R_i^* < \hat{R}_i$, suppose that $\ell_{i_0} < R_{i_0}^* < \hat{R}_{i_0}$ and $\ell_{i_1} < R_{i_1}^* < \hat{R}_{i_1}$ for $i_0 \neq i_1$. By Eq. (23), $R_{i_0}^* = R_{i_1}^*$. Choose $\delta > 0$ sufficiently small that $R_{i_0}^* - \delta > \ell_{i_0}$ and $R_{i_1}^* + \delta < \hat{R}_{i_1}$. Then, as in the case of minimization, it follows immediately that $\mathbf{R}^* + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$ is feasible, but increases the objective by $2\delta^2$, contrary to the hypothesis that $\mathbf{R}^*$ was a maximizer. In the same way, it follows that there exists at most one i such that $u_i > R_i^* > \hat{R}_i$.

Combining this with our previous necessary condition on $\mathbf{R}^*$ and letting $t^{\mathrm{lwr}} = \frac{1}{2} \cdot [\lambda - \Delta]$ and $t^{\mathrm{upr}} = \frac{1}{2} \cdot [\lambda + \Delta]$ yields that for all i such that $R_i^* < t^{\mathrm{lwr}}$, $R_i^* = \ell_i$; that for all i such that $R_i^* > t^{\mathrm{upr}}$, $R_i^* = u_i$; that there is at most one index such that $\hat{R}_i > R_i^* > t^{\mathrm{lwr}}$; and, finally, that there is at most one i' such that $\hat{R}_{i'} < R_{i'}^* < t^{\mathrm{upr}}$. □

Lemma B.11 is, by itself, insufficient to form the basis of an effective solution algorithm since it does not determine which indices are i^{lwr} and i^{upr}. Even if these indices were known, they do not pin down the value of R_i for $i \neq i^{\mathrm{lwr}}, i^{\mathrm{upr}}$, which could, almost without restriction, be ℓ_i, $\hat{R}_i$, or u_i. As a result, the search space of vectors in weak maximization normal form is extremely large.

Therefore, to construct an effective maximization algorithm, we must transform vectors merely in weak maximization normal form into vectors that are in full-fledged maximization normal form.

To do so, we will have need of the following elementary fact.

Lemma B.12. *Suppose $a \geq b$ and $c \geq d$. Then*

$$|a - c| + |b - d| \leq |a - d| + |b - c|.$$

Proof. The proof is not complicated, but is simplest to understand if expressed in geometric terms. Consider the points $\mathbf{p}_0 = (a, b)$ and $\mathbf{p}_1 = (c, d)$. Then both $\mathbf{p}_0$ and $\mathbf{p}_1$ lie in the lower half-plane

$$H = \{\, (x, y) : x \geq y\}.$$

Let $T : (x, y) \mapsto (y, x)$ be the linear transformation given by reflection across the line $y = x$. Then, the claim of the lemma can be reframed as follows: for any $\mathbf{p}_0, \mathbf{p}_1 \in H$,

$$\|\mathbf{p}_0 - \mathbf{p}_1\|_1 \leq \|\mathbf{p}_0 - T(\mathbf{p}_1)\|_1. \tag{24}$$

This setup is shown in Figure 7a. We divide into four cases according to whether $a \geq c$ and $b \geq d$, or, equivalently, depending on in which of the four regions shown in Figure 7b the point $\mathbf{p}_1$ lies.

Case (A): $a \geq c$ and $b \geq d$ In this case, we have that

$$\begin{aligned} |a - c| + |b - d| &= a - c + b - d \\ &= a - d + b - c \\ &\leq |(a - d) + (b - c)| \\ &\leq |a - d| + |b - c|. \end{aligned}$$

Therefore, Eq. (24) holds.

Case (B): $a \leq c$ and $b \geq d$ In this case,

$$\begin{aligned} |a - c| + |b - d| &= c - a + b - d \\ &\leq c - a + b - d + 2(a - b) \\ &= c - b + a - d \\ &\leq |(c - b) + (a - d)| \\ &\leq |c - b| + |a - d|, \end{aligned}$$

so Eq. (24) holds in this case as well.

Before completing the proof, we note that T has the useful property that

$$\|T(\mathbf{p})\|_1 = \|\mathbf{p}\|_1. \tag{25}$$

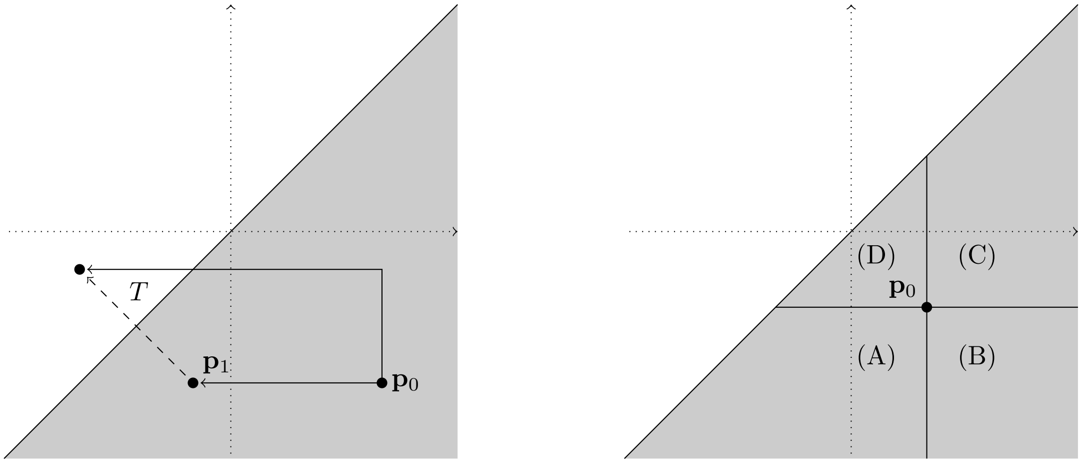
Case (C): $a \leq c$ and $b \leq d$ We note that this case is the same as Case (A) with the roles of $\mathbf{p}_0$ and $\mathbf{p}_1$ reversed. Consequently, by Eq. (24), we have that

$$\|\mathbf{p}_0 - \mathbf{p}_1\|_1 \leq \|T(\mathbf{p}_0) - \mathbf{p}_1\|_1.$$

Applying Eq. (25) and using the fact that T^2 is the identity, we have that

$$\|T(\mathbf{p}_0) - \mathbf{p}_1\|_1 = \|T(T(\mathbf{p}_0) - \mathbf{p}_1)\|_1 = \|\mathbf{p}_0 - T(\mathbf{p}_1)\|_1,$$

which yields Eq. (24).



(a) Setup

(b) Division into cases

Figure 7: *An illustration of the proof of Lemma B.12. The half-plane H is shown in gray, with the dotted lines representing the coordinate axes.* Left: *The claim of the lemma is equivalent to the claim that the path joining $\mathbf{p}_0$ to $\mathbf{p}_1$ is no longer than the path joining $\mathbf{p}_0$ to $T(\mathbf{p}_1)$.* Right: *The division into four cases, depending on where $\mathbf{p}_0$ lies relative to $\mathbf{p}_1$.*

Case (D): $a \geq c$ and $b \leq d$ We argue as in the previous case, noting that this case is equivalent to Case (B) with the roles of $\mathbf{p}_0$ and $\mathbf{p}_1$ reversed, whence

$$\|\mathbf{p}_0 - \mathbf{p}_1\|_1 \leq \|T(\mathbf{p}_0) - \mathbf{p}_1\|_1 = \|\mathbf{p}_0 - T(\mathbf{p}_1)\|_1.$$

□

A further difference from minimization is that the additional assumption of sortability, Definition 4.1, is required to connect maximizers to maximization normal form. If sortability fails, there may be intervals $[\ell_i, u_i]$ that can be non-trivially nested for distinct i, which can cause i_0 and i_1 to "jump between" different indices a potentially exponential number of times in the search for a maximizer. For notational convenience, we assume throughout that sortable risk vectors are, in fact, already sorted.

The first step to deriving an effective maximization algorithm is to remove "trivial" deviations from maximization normal form.

Definition B.13 (Trivial deviations from normality). We say that a risk vector in weak normal form has a *trivial deviation from normality* relative to its *bounds* if there are $i_0 < i_1$ such that $\ell_{i_0} = \ell_{i_1}$, $\ell_{i_0} < R_{i_0}$, and $R_{i_1} = \ell_{i_1}$; or there are $i'_1 > i'_0$ such that $u_{i'_0} = u_{i'_1}$, $R_{i'_1} < u_{i'_1}$, and $R_{i'_0} = u_{i'_0}$.

We say that a risk vector in weak normal form has a *trivial deviation from normality* relative to its *estimates* if there are $i_0 < i_1$ such that $\hat{R}_{i_0} = \hat{R}_{i_1}$ but $R_{i_0} > R_{i_1}$.

An illustration of trivial deviations from normality is shown in Figure 8. As their name suggests, trivial deviations from normality are straightforward to eliminate.

Lemma B.14. *Suppose the constraints are sortable. If a feasible risk vector* $\mathbf{R}$ *is in weak normal form with trivial deviations from normality, then there exists a feasible* $\mathbf{R}'$ *in weak normal form with no trivial deviations from normality such that*

$$\frac{1}{n}\sum_{i=1}^{n} R_i^2 = \frac{1}{n}\sum_{i=1}^{n} (R'_i)^2.$$

Proof. We first define an algorithm for removing trivial deviations from normality, and then show that it must eventually terminate when all trivial deviations have been removed.

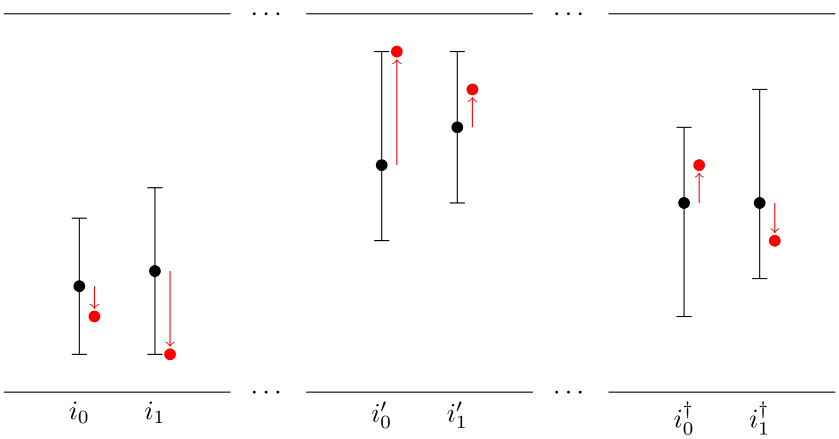


Figure 8: *An illustration of* trivial deviations *from normality. As in Figures 1 and 5, the black dots represent estimated risks $\hat{R}_i$, the error bars represent the allowable range $[\ell_i, u_i]$, and the red dots represent the corresponding values of R_i, with indices increasing from left to right on the x-axis. There is a trivial deviation from normality relative to the lower bound on the left because $\ell_{i_0} = \ell_{i_1}$ but $R_{i_1} = \ell_{i_1}$, and a similar illustration of a trivial deviation from normality relative to the upper bound in the center. On the right there is a trivial deviation from the estimates.*

Algorithm If there were any indices i_0 and i_1 representing a trivial deviation from normality, then we could eliminate them by swapping R_{i_0} and R_{i_1}, i.e., by considering

$$R'_i = R_i \quad (i \neq i_0, i_1), \qquad R'_{i_0} = R_{i_1}, \qquad R'_{i_1} = R_{i_0}.$$

In particular, it follows immediately that $\mathbf{R}'$ is still in weak normal form and that $\frac{1}{n}\sum_{i=1}^n R_i^2 = \frac{1}{n}\sum_{i=1}^n (R'_i)^2$. Because $\mathbf{R}$ is sorted, the only feasibility constraint on $\mathbf{R}'$ that is not satisfied trivially is that $\|\mathbf{R}' - \hat{\mathbf{R}}\|_1 \leq \epsilon$. However, this constraint follows by applying Lemma B.12 with $a = R_{i_0}, b = R_{i_1}, c = \hat{R}_{i_1}$, and $d = \hat{R}_{i_0}$, which yields

$$\|\mathbf{R}' - \hat{\mathbf{R}}\|_1 \leq \|\mathbf{R} - \hat{\mathbf{R}}\|_1 \leq \epsilon.$$

Termination We call a pair of indices (i_0, i_1) *out of order* if $i_0 < i_1$, but $R_{i_0} > R_{i_1}$. We note that any trivial deviation from normality requires that the corresponding indices i_0 and i_1 are out of order. Then, we observe the following facts:

Fact 1. There are only finitely many pairs of out-of-order indices.

Fact 2. As long as there is a trivial deviation of any kind, it is possible to swap indices so as to remove that trivial deviation.

Fact 3. Any swap undertaken as part of the algorithm strictly reduces the number of out-of-order pairs of indices.

We note that it follows immediately from these three facts that a greedy algorithm removing trivial deviations in any order whatsoever must eventually terminate.

Fact 1 is immediate, and Fact 2 follows from the discussion of the algorithm above. To see that Fact 3 holds, it suffices to show that *any* swap of out-of-order pairs of indices strictly reduces the number of out-of-order pairs of indices.

To this end, suppose $i_0 < i_1$ are out of order. As before, we use $\mathbf{R}'$ to denote the risk vector *after* swapping indices i_0 and i_1. First, suppose there is an index i^* such that $i_0 < i^* < i_1$. Then, if (i_0, i^*) is out of order *after* swapping i_0 and i_1, it must have been out of order before swapping, since

$$R_{i_0} > R_{i_1} = R'_{i_0} > R'_{i^*} = R_{i_*}.$$

Likewise, if (i^*, i_1) is out of order *after* swapping i_0 and i_1, it too must have been out of order before swapping, since

$$R_{i^*} = R'_{i^*} > R'_{i_1} = R_{i_0} > R_{i_1}.$$

Therefore, swapping cannot increase the number of out-of-order pairs involving $i_0 < i^* < i_1$.

On the other hand, since $R'_{i_0} < R'_{i_1}$, if $i^* < i_0$, then there are three possibilities:

Case 1. After swapping, neither (i^*, i_0) nor (i^*, i_1) is out of order.

Case 2. After swapping, both (i^*, i_0) and (i^*, i_1) are out of order.

Case 3. After swapping, (i^*, i_0) is out of order, but (i^*, i_1) is not.

In the first case, the number of out-of-order pairs cannot have increased, so we only need to consider the second and third cases. In Case 2, we must have that

$$R_{i^*} = R'_{i^*} > R'_{i_1} = R_{i_0} > R_{i_1},$$

so both (i^*, i_0) and (i^*, i_1) were out of order before swapping. In Case 3, we must have that

$$R_{i_0} = R'_{i_1} \geq R'_{i^*} = R_{i_*} > R'_{i_0} = R_{i_1},$$

so that before swapping (i^*, i_1) was out of order, but (i^*, i_0) was not. Therefore, it follows that swapping again cannot increase the number of out-of-order pairs involving $i^* < i_0$. A similar argument shows that the number of out-of-order pairs involving $i_1 < i^*$ does not increase after swapping. Therefore swapping strictly reduces the number of out-of-order pairs, since it at least eliminates the pair (i_0, i_1), which is Fact 3. □

With these preliminaries in hand, we are prepared to solve the maximization problem.

Lemma B.15. *Suppose the constraints are sortable. Then the single stratum maximization problem in Eq.* (15) *is maximized by a risk vector in maximization normal form and that either exhausts the budget or has* $i^{\text{lwr}} = i^{\text{upr}}$.

Proof. We adopt a strategy similar to that used in the "strengthening" portion of the proof of Lemma B.6. Namely, we show that any point satisfying the necessary conditions of Lemma B.11 but without pivots can be "improved" to a feasible point that achieves a greater objective.

Let $\mathbf{R}^*$ be a maximizer. By Lemma B.11, $\mathbf{R}^*$ is in weak maximization normal form; by Lemma B.14, we may assume that $\mathbf{R}^*$ has no trivial deviations from normality.

To prove the theorem, we first observe that it suffices to prove the following two claims:

1. There do not exist $i_0 < i_1$ such that $R^*_{i_0} > \ell_{i_0}$ and $R^*_{i_1} < \hat{R}_{i_1}$,
2. There do not exist $i_0 < i_1$ such that $R^*_{i_1} < u_{i_1}$ and $R^*_{i_0} > \hat{R}_{i_0}$.

First, we see how the theorem follows from the two claims. Define i^{lwr} and i^{upr} as follows:

$$i^{\text{lwr}} = \min\{\, i : R^*_i > \ell_i \,\}, \qquad i^{\text{upr}} = \max\{\, i : R^*_i < u_i \,\},$$

then, for all $i < i^{\text{lwr}}$, $R^*_i = \ell_i$, and for all $i > i^{\text{upr}}$, $R^*_i = u_i$. Moreover, for all i satisfying $i^{\text{lwr}} < i < i^{\text{upr}}$, by Claim 1, $R^*_i \geq \hat{R}_i$; and, by Claim 2, $R^*_i \leq \hat{R}_i$. Consequently, $R^*_i = \hat{R}_i$ for these indices. Moreover, if $i^{\text{lwr}} \neq i^{\text{upr}}$, by Claim 2, $R^*_{i^{\text{lwr}}} \leq \hat{R}_{i^{\text{lwr}}}$, i.e., $R^*_{i^{\text{lwr}}} \in [\ell_{i^{\text{lwr}}}, \hat{R}_{i^{\text{lwr}}}]$. Similarly, by Claim 1, we conclude that $R^*_{i^{\text{upr}}} \in [\hat{R}_{i^{\text{upr}}}, u_{i^{\text{upr}}}]$.

Therefore it suffices to prove the two claims. Since the proofs are virtually identical, we focus on Claim 1.

The key observation is the following. Suppose there exist indices $R^*_{i_0}$ and $R^*_{i_1}$ with $R^*_{i_1} \geq R^*_{i_0}$, and suppose that there is some δ such that after decreasing $R^*_{i_0}$ and increasing $R^*_{i_1}$ by δ the result is still feasible. Then $\mathbf{R}^*$ cannot be maximal, since, letting $\mathbf{R}' = \mathbf{R}^* + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$, we have that

$$\begin{aligned}
\left[\sum_{i=1}^{n} (R'_i)^2\right] - \left[\sum_{i=1}^{n} (R^*_i)^2\right] &= (R^*_{i_0} - \delta)^2 + (R^*_{i_1} + \delta)^2 - (R^*_{i_0})^2 - (R^*_{i_1})^2 \\
&= 2\delta(\delta + [R^*_{i_1} - R^*_{i_0}]) \qquad (26) \\
&\geq 2\delta^2 \\
&> 0.
\end{aligned}$$

To actually prove the claim, we proceed by contradiction. In particular, suppose that $i_0 < i_1$ represented an exception to Claim 1, so that $R^*_{i_0} > \ell_{i_0}$ and $R^*_{i_1} < \hat{R}_{i_1}$.

If $R^*_{i_0} \leq R^*_{i_1}$, then Eq. (26) immediately applies, since increasing $R^*_{i_1}$ *decreases* the L_1 distance between $\hat{\mathbf{R}}$ and $\mathbf{R}^*$ and decreasing $R^*_{i_0}$, at worst, *increases* it at the same rate. More formally, for $0 < \delta$ sufficiently small,

$$\begin{aligned}
|(R^*_{i_0} - \delta) - \hat{R}_{i_0}| + |(R^*_{i_1} + \delta) - \hat{R}_{i_1}| &= |(R^*_{i_0} - \delta) - \hat{R}_{i_0}| + \hat{R}_{i_1} - (R^*_{i_1} + \delta) \\
&\leq |R^*_{i_0} - \hat{R}_{i_0}| + \delta + \hat{R}_{i_1} - R^*_{i_1} - \delta \\
&= |R^*_{i_0} - \hat{R}_{i_0}| + |R^*_{i_1} - \hat{R}_{i_1}|,
\end{aligned}$$

and so $\mathbf{R}'$ is feasible. Because $R^*_{i_0} \leq R^*_{i_1}$, by Eq. (26), $\mathbf{R}'$ achieves a greater average sum of squares that $\mathbf{R}^*$, contrary to the assumed maximality of $\mathbf{R}^*$.

On the other hand, if $R^*_{i_0} > R^*_{i_1}$, then, by Lemma B.12, the vector $\mathbf{R}^\dagger$ given by switching the indices i_0 and i_1, i.e.,

$$R^\dagger_i = R^*_i \quad (i \neq i_0, i_1), \qquad R^\dagger_{i_0} = R^*_{i_1}, \qquad R^\dagger_{i_1} = R^*_{i_0},$$

satisfies $\|\mathbf{R}^\dagger - \hat{\mathbf{R}}\|_1 \leq \epsilon$. Moreover, applying our sortability hypothesis, we have that

$$u_{i_0} \geq R^*_{i_0} > R^*_{i_1} \geq \ell_{i_1} \geq \ell_{i_0},$$

and

$$u_{i_1} \geq u_{i_0} \geq R^*_{i_0} > R^*_{i_1} \geq \ell_{i_1},$$

we have that for all $i = 1, \ldots, n$, $\ell_i \leq R^\dagger_i \leq u_i$. Therefore $\mathbf{R}^\dagger$ is feasible. However, since $\mathbf{R}^*$ has no trivial deviations from normality, it follows that

$$R^\dagger_{i_0} = R^*_{i_1} \geq \ell_{i_1} > \ell_{i_0}, \qquad R^\dagger_{i_1} = R^*_{i_0} \leq u_{i_0} < u_{i_1}.$$

Since $R^\dagger_{i_0} < R^\dagger_{i_1}$, we would like to apply Eq. (26) to derive a contradiction. To that end, we require that $\mathbf{R}^\dagger + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$ is feasible for sufficiently small $\delta > 0$. If it were the case that

$$\|\mathbf{R}^\dagger + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})\|_1 \leq \|\mathbf{R}^\dagger\|_1, \tag{27}$$

then the result follows immediately. However, Eq. (27) holds except when $R^\dagger_{i_0} \leq \hat{R}_{i_0}$ and $R^\dagger_{i_1} \geq \hat{R}_{i_1}$; or, equivalently, $R^*_{i_1} \leq \hat{R}_{i_0}$ and $R^*_{i_0} \geq \hat{R}_{i_1}$—see Figure 9. However, in this case, we have by sortability that

$$\begin{aligned}
|R^*_{i_0} - \hat{R}_{i_0}| + |R^*_{i_1} - \hat{R}_{i_1}| &= R^*_{i_0} - \hat{R}_{i_0} + \hat{R}_{i_1} - R^*_{i_1} \\
&> (R^*_{i_0} - \hat{R}_{i_0} + \hat{R}_{i_1} - R^*_{i_1}) + 2(\hat{R}_{i_0} - \hat{R}_{i_1}) \\
&= R^*_{i_0} - \hat{R}_{i_1} + \hat{R}_{i_0} - R^*_{i_1} \\
&= |R^\dagger_{i_0} - \hat{R}_{i_0}| + |R^\dagger_{i_1} - \hat{R}_{i_1}|,
\end{aligned}$$

where we have used the fact that, because there are no trivial deviations from normality relative to the estimates, $\hat{R}_{i_1} > \hat{R}_{i_0}$. Therefore $\|\mathbf{R}^\dagger\|_1 < \epsilon$ and so we can choose δ sufficiently small that $\|\mathbf{R}^\dagger + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})\|_1 \leq \epsilon$, completing the proof.

Therefore, no such $i_0 < i_1$ exist, and so $\mathbf{R}^*$ is in maximization normal form. □

Lemma B.15 allows us to efficiently search over the space of risk vectors to find the maximizer. In particular, all that is necessary is to decrease the risk at the lower pivot and increase it at the upper pivot until the L_1 budget is expended. The linear-time algorithm for doing so is given in Algorithm 7. (As in the case of Algorithm 4, there is a potential preprocessing step, given in Algorithm 6.) As we did in the minimization algorithm, we capture the maximization algorithm's value at a variety of L_1 budgets, controlled by some step-size γ—which, we assume, satisfies $k\gamma = \epsilon$ for some $k \in \mathbb{N}$—rather than just the full budget ϵ. We denote this collection with $\boldsymbol{\Sigma}$. The reason for doing this is to allow us to maximize across strata.

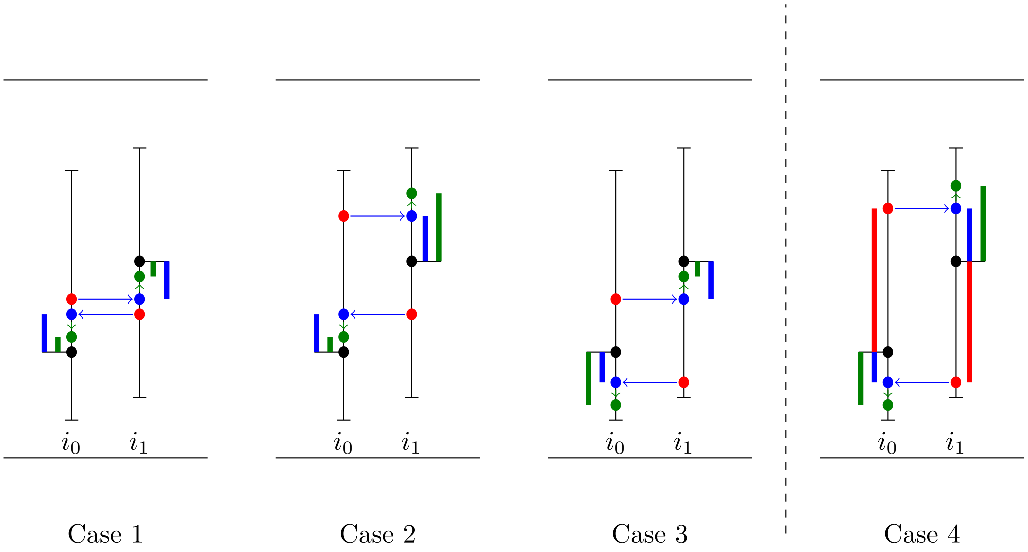


Figure 9: *This diagram shows various possibilities of the relative ordering of* $R^*_{i_0}$, $R^*_{i_1}$, $\hat{R}_{i_0}$, *and* $\hat{R}_{i_1}$ *in the final section of the proof of Lemma B.15. Here, the upper and lower bounds are shown by the error bars, the original estimates (*$\hat{\mathbf{R}}$*) are shown in black, the starting risk vector (*$\mathbf{R}^*$*) as red dots, the risk vector* after *indices* i_0 *and* i_1 *have been exchanged (*$\mathbf{R}^\dagger$*) as blue dots, and the vector with increased objective (*$\mathbf{R}^\dagger + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$*) as green dots. Here we see that in Case 1, where* $\hat{R}_{i_1} \geq R^*_{i_0} \geq R^*_{i_1} \geq \hat{R}_{i_0}$*, perturbing* $\mathbf{R}^\dagger$ *by* $\delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$ *actually* decreases *the* L_1 *distance to* $\hat{\mathbf{R}}$*; the* L_1 *distance between* $\hat{\mathbf{R}}$ *and* $\mathbf{R}^\dagger$ *is shown with the heavy blue lines, and between* $\hat{\mathbf{R}}$ *and* $\mathbf{R}^\dagger + \delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$ *by the heavy green lines. In Case 2, where* $R^*_{i_0} \geq \hat{R}_{i_1} \geq R^*_{i_1} \geq \hat{R}_{i_0}$*, we see that perturbing* $\mathbf{R}^\dagger$ *by* $\delta(\mathbf{e}_{i_1} - \mathbf{e}_{i_0})$ *increases the absolute distance from* $\hat{\mathbf{R}}$ *at* i_1*, but this increase is exactly compensated by a decrease at* i_0*. The situation is identical in Case 3, where* $\hat{R}_{i_1} \geq R^*_{i_0} \geq \hat{R}_{i_0} \geq R^*_{i_1}$*, except with the roles of* i_0 *and* i_1 *reversed. Finally, in Case 4, which is treated separately, we see that the initial transformation from* $\mathbf{R}^*$ *to* $\mathbf{R}^\dagger$ *strictly reduces the* L_1 *distance to* $\hat{\mathbf{R}}$ *when we compare* $\mathbf{R}^*$ *(heavy red lines) to* $\mathbf{R}^\dagger$ *(heavy blue lines), so that the* L_1 *distance between* $\hat{\mathbf{R}}$ *and the perturbed risk vector (heavy green lines) is still less than the constraint.*

B.3.2 Optimizing over all strata

Consider the simplified maximization problem in Eq. (9). Since the objective is separable, the problem would be separable in the different strata, except that the L_1 budget constraint ranges over all of the strata simultaneously. Much as we reduced the base problem to the parameterized problem by introducing additional parameters τ_j, $j = 1, \dots, m$, if the optimal budget allocation to each stratum were known, then we could reduce to the single stratum case, applying Algorithm 7. In particular, we note that the following optimization problem (hereafter **"the separable problem"**) parameterized by the stratum-specific budgets $\epsilon_{1,0}, \dots, \epsilon_{m,0}, \epsilon_{1,1}, \dots, \epsilon_{m,1}$ is separable:

$$\begin{array}{lll} \underset{\mathbf{R}\in\mathbb{R}^n}{\text{Optimize}} & \frac{1}{n}\sum_{i=1}^{n} R_i^2 & \\ \quad\text{s.t. } \frac{1}{|\mathcal{S}_{j,1}|}\sum_{i\in\mathcal{S}_{j,a}}^{n} |R_i - \hat{R}_i| \le \epsilon_{j,a}, & (a = 0, 1;\ j = 1, \dots, m) & \\ \qquad\qquad \sum_{i\in\mathcal{S}_{j,1}} R_i = \rho_j, & (j = 1, \dots, m) & \\ \qquad\qquad \sum_{i\in\mathcal{S}_{j,0}} R_i = \tau_j, & (j = 1, \dots, m) & \\ \qquad\qquad\qquad R_i \le u_i, & (i = 1, \dots, n) & \\ \qquad\qquad\qquad R_i \ge \ell_i. & (i = 1, \dots, n) & \end{array} \tag{28}$$

In particular, the separable problem can be solved stratum by stratum, since the constraints and objective are both separable at the stratum level.

Reducing solving the simplified problem to solving the separable problem, however, is complex; There is no obvious way to determine the optimal L_1 budget allocation other than by sweeping over all possible budgets. However, it is nevertheless possible to efficiently discover an *approximately* optimal budget allocation using a divide-and-conquer approach. In particular, as shown in Algorithm 7, it is only marginally more difficult computationally to output the maximum value of the objective for the entire budget than to output the maximum value of the objective for a range of budgets, i.e., to output $\mathbf{\Sigma}$, which gives the maximum value of the objective for the budgets 0, γ, 2γ, …, ϵ.

Given $\mathbf{\Sigma}^{(0)}$ and $\mathbf{\Sigma}^{(1)}$ corresponding to two different strata when the allocated budget is 0, γ, 2γ, …, ϵ, we can find the approximate maximum value of the objective when the *total* budget allocated to both strata is 0, γ, 2γ, …, ϵ by sweeping over the two dimensional grid of points $\{0, \gamma, \dots, \epsilon\} \times \{0, \gamma, \dots, \epsilon\}$, and, for $k\gamma$, outputting

$$\max\left(\Sigma_0^{(0)} + \Sigma_{k+1}^{(1)}, \Sigma_1^{(0)} + \Sigma_k^{(1)}, \dots, \Sigma_{k+1}^{(0)} + \Sigma_0^{(1)}\right).$$

For completeness, we explicitly give this maximization routine in Algorithm 8.

Lemma B.16. *If $\boldsymbol{\ell}$, $\hat{\mathbf{R}}$ and $\mathbf{u}$ have been sorted, then the approximate solution to the maximization problem in Eq. (9) given by applying Algorithms 7 and 8 is*

$$O\left(m \cdot \left(\frac{\epsilon}{\gamma}\right)^2 + n\right).$$

We note that δ, as used in Section 4 above, equals $2m\gamma$.

Proof. We begin by analyzing Algorithm 6. We note that the number of iterations of the while loop on lines 7 through 12 is capped by the length of the input, since i^{upr} is decremented on each input. Likewise, Algorithm 7 is almost linear in the size of the input: in the while loop on lines 14 though 30, either i^{lwr} is incremented, or the subsequent multiple of γ is reached, meaning that the algorithm as a whole is linear in the size of the input and ϵ/γ. Therefore, running Algorithm 4 over all $2m$ strata requires $O(n + \frac{m\epsilon}{\gamma})$ time.

All of the strata can be combined using $2m - 1$ applications of Algorithm 5. Reviewing the for loop on lines 3 through 11 and the while loop on lines 5 through 8, we see that exactly

$$\frac{\frac{\epsilon}{\gamma} \cdot (\frac{\epsilon}{\gamma} + 1)}{2}$$

iterations are performed. Adding this to the previous runtime obtained and simplifying gives the desired expression. □

B.4 Controlling approximation error

There are two sources of approximation error in our algorithm. The first is the approximation error introduced by solving the maximization problem only approximately. The second is the approximation error introduced by the fact that we cannot sweep over all possible values of $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)$ when solving the simplified problem; instead, we must sweep over a grid where. Characterizing how much error is introduced by each of these approximations is our final task.

We begin with the following simple lemma.

Lemma B.17. *Suppose* $\mathbf{R}^{(0)}$ *and* $\mathbf{R}^{(1)}$ *are risk vectors. Then*

$$\left| \sum_{i=1}^{n} (R_i^{(1)})^2 - \sum_{i=1}^{n} (R_i^{(0)})^2 \right| \leq 2 \cdot \|\mathbf{R}^{(1)} - \mathbf{R}^{(2)}\|. \tag{29}$$

Proof. The claim, which is a version of Hölder's inequality, follows straightforwardly from considering the following difference:

$$\begin{aligned} \left| \sum_{i=1}^{n} (R_i^{(1)})^2 - (R_i^{(0)})^2 \right| &\leq \sum_{i=1}^{n} |(R_i^{(1)})^2 - (R_i^{(0)})^2| \\ &\leq \sum_{i=1}^{n} (R_i^{(1)} + R_i^{(0)}) \cdot |R_i^{(1)} - R_i^{(0)}| \\ &\leq \sum_{i=1}^{n} 2 \cdot |R_i^{(1)} - R_i^{(0)}| \\ &= 2 \cdot \|\mathbf{R}^{(1)} - \mathbf{R}^{(0)}\|_1. \end{aligned}$$

The first inequality follows from the triangle inequality, and the second and third from the fact that $0 \leq R_i^{(0)}, R_i^{(1)} \leq 1$. □

Using Lemma B.17, we obtain the following characterization of the potential error in our maximization algorithm, again recalling that $\delta = 2m\gamma$, where δ is as in Section 4.

Lemma B.18. *The difference between the true maximum of the simplified problem—i.e., Eq. (9)—and the quantity obtained from applying Algorithms 4 and 5 is at most* $2m\gamma$.

Proof. Let $\mathbf{R}^*$ be the true maximizing solution. We will show that Algorithm 8 computes the value of the objective at a "close" point at which we can apply Lemma B.17.

For each stratum $\mathcal{S}_{j,a}$, let $\epsilon_{j,a}$ denote

$$\frac{1}{n} \sum_{i \in \mathcal{S}_{j,a}} |R_i^* - \hat{R}_i|.$$

Then, there exist $\tilde{\epsilon}_{j,a} \in \{0, \gamma, 2\gamma, \dots, \epsilon\}$ such that $\tilde{\epsilon}_{j,a} < \epsilon_{j,a}$ and $\epsilon_{j,a} - \tilde{\epsilon}_{j,a}$ is less than γ.

Let

$$\Delta = \sum_{a=0}^{1} \sum_{j=1}^{m} \epsilon_{j,a} - \tilde{\epsilon}_{j,a} = \epsilon - k\gamma$$

for some k. Since ϵ is a multiple of γ, it follows that $\Delta = k'\gamma$ for some $k \in \mathbb{N}$, where $k' < 2m$, the number of strata.

Set $\epsilon'_{j,a} = \tilde{\epsilon}_{j,a}$ except for k' arbitrarily chosen strata, where we set $\epsilon'_{j,a} = \tilde{\epsilon}_{j,a} + \gamma$ instead. Then, there exists $\mathbf{R}$ that is in (maximization) normal form such that

$$\frac{1}{n} \sum_{i=1}^{n} |R_i - \hat{R}_i| = \epsilon, \qquad \frac{1}{n} \sum_{\substack{A_i = a \\ C_i = c_j}} |R_i - \hat{R}_i| = \epsilon'_{j,a}.$$

The remainder of the proof follows simply by comparing $\mathbf{R}$ and $\mathbf{R}^*$ stratum by stratum. In particular, consider the restrictions of $\mathbf{R}$ and $\mathbf{R}^*$ to a single stratum $\mathcal{S}_{j,a}$. By Eq. (9), we have that

$$\sum_{i \in \mathcal{S}_{j,a}} R_i = \sum_{i \in \mathcal{S}_{j,a}} R_i^*,$$

and, moreover, by Lemma B.15, both $\mathbf{R}_{j,a}$ and $\mathbf{R}_{j,a}^*$ can be assumed to be in maximization normal form. Assume, without loss of generality, that $\epsilon'_{j,a} \le \epsilon_{j,a}$. (The proof is virtually identical if the inequality is reversed.)

Let i_0 and i_1 denote the pivots of $\mathbf{R}$, and i_0^* and i_1^* of $\mathbf{R}^*$, so that $i_0 \le i_0^* \le i_1^* \le i_1$. (To avoid unnecessary notational complication, we write as if all n individuals belong to the stratum $\mathcal{S}_{j,a}$ and the constraints were sorted.) Then, we note that

$$R_i \ge R_i^* \quad (i_0 \le i \le i_0^*), \qquad R_i \le R_i^* \quad (i_1^* \le i \le i_1), \qquad R_i = R_i^* \quad (\text{otherwise}). \tag{30}$$

It follows that

$$\sum_{i=1}^{n} (R_i^*)^2 - R_i^2 = \left[\sum_{i=i_1^*}^{i_1} (R_i^*)^2 - R_i^2\right] - \left[\sum_{i=i_0}^{i_0^*} R_i^2 - (R_i^*)^2\right]. \tag{31}$$

Both terms in Eq. (31) are positive by Eq. (30). Moreover, by Lemma B.17, both terms are less than or equal to

$$2 \cdot \left[\sum_{i=i_1}^{i_1^*} R_i^* - R_i\right] = 2 \cdot \left[\sum_{i=i_0}^{i_0^*} R_i - R_i^*\right] = 2n \cdot \frac{\epsilon_{a,j} - \epsilon'_{a,j}}{2} \le n\gamma.$$

In particular, their difference must also be less than this quantity, and so, summing across strata, we have that

$$\sum_{i=1}^{n} (R_i^*)^2 - R_i^2 \le 2nm\gamma.$$

Dividing through by n gives the result. □

The second element we need is a bound on the sensitivity of the objective of the simplified problem to changes in the parameters.

Lemma B.19. *Let $\mathbf{R}^{(0)}$ and $\mathbf{R}^{(1)}$ be solutions to the simplified problem in Eq. (9) with parameters*

$$\boldsymbol{\tau}^{(0)} = (\tau_1^{(0)}, \dots, \tau_m^{(0)}), \qquad \boldsymbol{\tau}^{(1)} = (\tau_1^{(1)}, \dots, \tau_m^{(1)}),$$

respectively. Then the difference between the objective values of the two solutions is at most $4 \cdot \|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1$.

Proof. The strategy is simple: we will transform $\mathbf{R}^{(0)}$ into a feasible solution of the simplified problem with parameters $\boldsymbol{\tau}^{(1)}$ and *vice versa* for $\mathbf{R}^{(1)}$ and $\boldsymbol{\tau}^{(0)}$. These new solutions lower or upper bound—depending on whether we are considering minimization or maximization—the objective of the new problem; however, by Lemma B.17, we can also bound their difference in objective value from the original solutions, giving a bound on the difference in objective value of the original solutions.

In particular, we see that the lemma immediately follows from Lemma B.17 and if there exist $\mathbf{R}^{(2)}$ and $\mathbf{R}^{(3)}$ such that

$$\|\mathbf{R}^{(0)} - \mathbf{R}^{(2)}\|_1, \|\mathbf{R}^{(1)} - \mathbf{R}^{(3)}\|_1 \le 2 \cdot \|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1$$

and such that $\mathbf{R}^{(2)}$ and $\mathbf{R}^{(3)}$ are feasible for the simplified problem with parameters $\boldsymbol{\tau}^{(1)}$ and $\boldsymbol{\tau}^{(0)}$, respectively. For, in that case, we have that, assuming without loss of generality that we are solving the minimization problem,

$$\sum_{i=1}^{n} (R_i^{(0)})^2 \le \sum_{i=1}^{n} (R_i^{(2)})^2 \le \sum_{i=1}^{n} (R_i^{(1)})^2 + 4 \cdot \|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1,$$

and similarly

$$\sum_{i=1}^{n} (R_i^{(1)})^2 \le \sum_{i=1}^{n} (R_i^{(3)})^2 \le \sum_{i=1}^{n} (R_i^{(0)})^2 + 4 \cdot \|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1.$$

Therefore, it suffices to construct $\mathbf{R}^{(2)}$ and $\mathbf{R}^{(3)}$. The construction involves two steps:

1. Ensure that the sum of the risk vector is correct within each stratum;
2. Ensure that the L_1 budget constraint is satisfied.

The L_1 distance needed to achieve each of these steps is bounded by $\|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1$, and so $\mathbf{R}^{(2)}$ and $\mathbf{R}^{(3)}$ will have the required property.

Assume again that we are solving the minimization problem. Consider the j-th unobserved stratum (i.e., $i \in \mathcal{S}_{j,1}$) with associated upper and lower thresholds t^{lwr} and t^{upr} and suppose without loss of generality that $\tau_j^{(0)} \leq \tau_j^{(1)}$. Then, to achieve the first step, we simply raise t^{upr} until the sum of the risk vector is $\tau_j^{(1)}$ or $t^{\text{upr}} = 1$. If $t^{\text{upr}} = 1$, then we raise t^{lwr} until the sum of the risk vector is $\tau_j^{(1)}$. We note that since our risk vector has strictly increased, the L_1 distance between the original and new risk vectors is exactly the difference in their sums, i.e., $\tau_j^{(1)} - \tau_j^{(0)}$. Call the risk vector that results from performing this operation across all strata $\mathbf{R}'$. Then, it follows that

$$\|\mathbf{R}' - \hat{\mathbf{R}}\|_1 \leq \|\mathbf{R}^{(0)} - \hat{\mathbf{R}}\|_1 + \tau_j^{(1)} - \tau_j^{(0)} \leq \epsilon + \tau_j^{(1)} - \tau_j^{(0)}.$$

To achieve the second step, in strata $\mathcal{S}_{j,a}$ where $t^{\text{upr}} < 1$ and $t^{\text{lwr}} > 0$, we simply raise t^{upr} and lower t^{lwr} so as to ensure that the sum of the risk vector does not change. Doing so reduces the L_1 distance between the risk vector and $\hat{\mathbf{R}}$, so we do so until the L_1 budget constraint is satisfied. In particular, this process can be carried out until $t^{\text{upr}} = 1$ or $t^{\text{lwr}} = 0$ in every stratum. If the budget constraint is still not satisfied, then the simplified problem for $\boldsymbol{\tau}^{(1)}$ is not feasible, contrary to our assumption that there exists a solution $\mathbf{R}^{(1)}$. Therefore, the process halts, and, moreover, requires moving at most $\|\boldsymbol{\tau}^{(0)} - \boldsymbol{\tau}^{(1)}\|_1$ in L_1 distance. Call the resulting risk vector $\mathbf{R}^{(2)}$. Then, it follows that

$$\sum_{i \in \mathcal{S}_{j,1}} R_i^{(2)} = \tau_j^{(1)} \quad (j = 1, \ldots, m), \qquad \|\mathbf{R}^{(2)} - \hat{\mathbf{R}}\|_1 \leq \epsilon.$$

The other constraints are satisfied trivially. We can construct $\mathbf{R}^{(3)}$ similarly, and the proof is complete for minimization. For maximization, the argument is exactly similar, except that we increase and decrease R_{i_0} and R_{i_1} instead of t^{lwr} and t^{upr}. □

We can extend the previous results to bound the error in the calculation of the coefficients themselves.

Lemma B.20. *Let $V(\epsilon)$ denote the value of*

$$\begin{array}{rl} \underset{\mathbf{R} \in \mathbb{R}^n}{\text{Minimize}} & \frac{1}{n} \sum_{j=1}^{m} n_j \cdot \text{VAR}\left((R_i)_{i \in \mathcal{G}_j}\right) \\ \text{s.t.} & \frac{1}{n} \sum_{i=1}^{n} |R_i - \hat{R}_i| \leq \epsilon \\ & \sum_{i \in \mathcal{S}_{j,1}} R_i = \rho_j, \quad (j = 1, \ldots, m) \\ & R_i \leq u_i, \quad (i = 1, \ldots, n) \\ & R_i \geq \ell_i. \quad (i = 1, \ldots, n) \end{array}. \tag{32}$$

To run Algorithm 1 with parameter ϵ, let

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$$

denote the step size of our grid search in each of the m strata on the $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$ scale. Let δ^* denote the true optimum of the base problem, and let $\delta^\dagger$ denote the value of the objective returned by applying Algorithm 1. Then

$$|\delta^* - \delta^\dagger| \leq \frac{6\|\boldsymbol{\eta}\|_1 + 2m\gamma}{V(\epsilon)^2} + \frac{\frac{2}{n_1}\eta_1 + \frac{2}{n_j}\eta_j}{V(\epsilon)} + \frac{2\|\boldsymbol{\eta}\|_1}{V(\epsilon)}.$$

Proof. The proof is a straightforward exercise in bounding the various terms in Eq. (10) in Lemma B.2. In particular, by Eq. (10), $|\delta^* - \delta^\dagger|$ is equivalent to

$$\left| \frac{a_0 \cdot b_0}{c_0 + d_0} - \frac{a_1 \cdot b_1}{c_1 + d_1} \right|, \tag{33}$$

where

$$a_0 = \frac{1}{n}\sum_{j=1}^{m} \sigma_j \cdot \tau_j^* - (1-\sigma_j)\cdot \rho_j, \qquad a_1 = \frac{1}{n}\sum_{j=1}^{m} \sigma_j \cdot \tau_j^\dagger - (1-\sigma_j)\cdot \rho_j,$$

$$b_0 = \frac{\tau_j^* + \rho_j}{n_j} - \frac{\tau_1^* - \rho_j}{n_j}, \qquad b_1 = \frac{\tau_j^\dagger + \rho_j}{n_j} - \frac{\tau_1^\dagger - \rho_1}{n_1},$$

$$c_0 = \frac{1}{n}\sum_{i=1}^{n} (R_i^*)^2, \qquad c_1 = \frac{1}{n}\sum_{i=1}^{n} (R_i^\dagger)^2,$$

$$d_0 = \frac{1}{n}\sum_{j=1}^{m} \left(\frac{\rho_j + \tau_j^*}{n_j}\right)^2, \qquad d_1 = \frac{1}{n}\sum_{j=1}^{m} \left(\frac{\rho_j + \tau_j^\dagger}{n_j}\right)^2.$$

Here, $\mathbf{R}_i^*$ is the optimal risk vector—which, by Corollary B.4, must correspond to an optimum of the simplified problem—and $\mathbf{R}_i^\dagger$ is the risk vector found by Algorithm 1.[24] Since the gap between the optimal and approximate solutions can only increase, we may assume without loss of generality that $\mathbf{R}^\dagger$ is the risk vector found by Algorithm 5 or Algorithm 8 at the gridpoint nearest $(\tau_1^*, \dots, \tau_m^*)$. Finally, since only maximizing the objective involves approximation error, we may assume without loss of generality that $\mathbf{R}^\dagger$ is the risk vector found by Algorithm 8.

Now, it follows by a straightforward algebraic manipulation that

$$\delta^* - \delta^\dagger = a_0 \cdot b_0 \cdot \left(\frac{1}{c_0 + d_0} - \frac{1}{c_1 + d_1}\right) + a_0 \cdot (b_0 - b_1) \cdot \frac{1}{c_1 + d_1} + (a_0 - a_1) \cdot b_1 \cdot \frac{1}{c_1 + d_1}.$$

Therefore, we seek bounds A, B, and V such that

$$|a_0|, |a_1| \le A, \qquad |b_0|, |b_1| \le B, \qquad |c_0 + d_0|, |c_1 + d_1| \ge V$$

and Δ_a, Δ_b, and Δ_v such that

$$|a_0 - a_1| \le \Delta_a, \qquad |b_0 - b_1| \le \Delta_b, \qquad \left|\frac{1}{c_0 + d_0} - \frac{1}{c_1 + d_1}\right| \le \Delta_v.$$

Then, it will follow that

$$|\delta^* - \delta^\dagger| \le AB\Delta_v + \frac{A}{V}\Delta_b + \frac{B}{V}\Delta_a.$$

To find A, B, and V, we note that for any possible t_j and r_j, we have that

$$\sum_{j=1}^{m} |\sigma_j \cdot t_j| \le n, \qquad \sum_{j=1}^{m} |(1-\sigma_j)\cdot r_j| \le n,$$

whence $|a_i| \le 2$ for $i = 0, 1$. Therefore we can take $|A| = 2$. In the same way, we have the bound

$$\left|\frac{\tau_j + \rho_j}{n_j} - \frac{\tau_1 - \rho_j}{n_1}\right| \le 2,$$

so we take $B = 2$ as well. Finally, we have that $|c_0 + d_0| \ge V(\epsilon)$ and $|c_1 + d_1| \ge V(\epsilon)$ by definition, so we take $V = V(\epsilon)$.

Now, note that

$$|a_0 - a_1| = \left|\frac{1}{n}\sum_{j=1}^{m} \sigma_j \cdot (\tau_j^* - \tau_j^\dagger)\right| \le \frac{1}{n}\sum_{j=1}^{m} \left|\tau_j^* - \tau_j^\dagger\right| \le \frac{1}{n}\sum_{j=1}^{m} n_j \eta_j = \|\boldsymbol{\eta}\|_1,$$

[24] We note that Algorithms 5 and 8, for clarity, do not actually return the risk vector $\mathbf{R}$ itself, but could easily be modified to do so. This modification is easily carried through to Algorithm 1.

so we take $\Delta_a = \|\boldsymbol{\eta}\|_1$. Next, since $|\tau_j^* - \tau_j^\dagger| \leq \eta_j$ and similarly for τ_1^* and $\tau_1^\dagger$, we have that

$$|b_0 - b_1| \leq \frac{1}{n}\sum_{j=1}^{m} \left|\frac{\tau_j^* - \tau_j^\dagger}{n_j}\right| + \left|\frac{\tau_1^* - \tau_1^\dagger}{n_j}\right| \leq \frac{2\eta_1}{n_1} + \frac{2\eta_j}{n_j},$$

so we take $\Delta_b = \frac{2\eta_1}{n_1} + \frac{2\eta_j}{n_j}$. Finally, we have that

$$\left|\frac{1}{c_0 + d_0} - \frac{1}{c_1 + d_1}\right| = \left|\frac{c_1 - c_0 + d_1 - d_0}{(c_0 + d_0)(c_1 + d_1)}\right| \leq \frac{|c_0 - c_1| + |d_0 - d_1|}{V^2}.$$

Therefore, it only remains to bound $|c_0 - c_1|$ and $|d_0 - d_1|$. By Lemma B.19, if $\mathbf{R}^\dagger$ were the *true* solution to the simplified problem, then $|c_0 - c_1|$ would be bounded by $4\|\boldsymbol{\eta}\|_1$, since the L_1 distance to the nearest gridpoint is at most $\|\boldsymbol{\eta}\|_1$. However, since $\mathbf{R}^\dagger$ is only an approximate solution, applying Lemma B.18 gives that $|c_0 - c_1|$ is bounded by $2m\gamma + 4\|\boldsymbol{\eta}\|_1$. Similarly, using the fact that

$$\left|\frac{\tau_j^* + \rho_j}{n_j} - \frac{\tau_j^\dagger + \rho_j}{n_j}\right| \leq \eta_j,$$

we can apply Lemma B.17 to obtain that $|d_0 - d_1|$ is less than or equal to $2\|\boldsymbol{\eta}\|_1$. Combining these bounds gives that

$$|\delta^* - \delta^\dagger| \leq \frac{6\|\boldsymbol{\eta}\|_1 + 2m\gamma}{V(\epsilon)^2} + \frac{\frac{2}{n_1}\eta_1 + \frac{2}{n_j}\eta_j}{V(\epsilon)} + \frac{2\|\boldsymbol{\eta}\|_1}{V(\epsilon)},$$

as desired. □

Lemma B.20 gives intuition for how to choose the step sizes appropriately so as to minimize the error in the coefficients for a given amount of computation. In particular, since solving the parameterized problem requires roughly the same number of steps at each grid point, the computation scales like the number of gridpoints, i.e., like

$$\frac{1}{\prod_{j=1}^m \eta_j}.$$

The problem of choosing step sizes $\boldsymbol{\eta}$ so as to maximize accuracy for a given amount of computation is therefore essentially equivalent to the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\eta}\in\mathbb{R}^m}{\text{Minimize}} \quad & \sum_{j=1} \lambda_j \eta_j \\ \text{s.t.} \quad & \eta_j > 0, \quad (j = 1, \ldots, m) \\ & \prod_{j=1}^{m} \eta_j = M. \end{aligned}$$

Applying the first-order necessary KKT conditions yields that

$$\lambda_j = \nu \prod_{k\neq j} \eta_k, \qquad \text{i.e.,} \quad \lambda_j \eta_j = \nu M,$$

whence we have that

$$\eta_j = \frac{\sqrt[m]{M \prod_{j=1}^m \lambda_j}}{\lambda_j}.$$

In particular, the step sizes should be chosen so that they are inversely proportional to their weights in the error bound in Lemma B.20.

Unfortunately, these weights vary with ϵ. For large ϵ, when $V(\epsilon)$ is close to zero, the weights are dominated by the first term, which is optimized when $\eta_1 = \cdots = \eta_m$. For small ϵ, when $V(\epsilon)$ is large, the weights are mixture of (somewhat larger) terms that would be optimized by $\eta_1 = \cdots = \eta_m$ and (somewhat smaller) terms that would be optimized by $\frac{\eta_1}{n_1} = \cdots = \frac{\eta_m}{n_m}$.

Thus, one reasonable heuristic that is likely to perform well across a range of ϵ is to choose η_j to be equal to some fixed η for all j. We use this heuristic in our experiments and software implementation.

Alternatively, to simplify the error bounds and eliminate the dependence on the data through the n_j, one could choose η_j to be proportional to n_j for all j. Making this choice gives the error bounds in Theorem 4.2, while the runtime bounds are given by Lemmata B.8 and B.16.

Lastly, since it is a convex problem, it is generally practical to compute $V(\epsilon)$. However, to obtain a bound on the error entirely in terms of the input parameters, we can use the following lemma.

Lemma B.21. *Let $V(\epsilon)$ be defined as in Lemma B.20. Then*

$$V(\epsilon) \geq \left[\frac{1}{n}\sum_{j=1}^{m} \textsc{Var}\left((\hat{R}_i)_{i\in\mathcal{S}_{j,1}}\right)\right] - 4\epsilon. \tag{34}$$

Proof. Let $\mathbf{R}^*$ be the solution to the optimization problem defining $V(\epsilon)$. Then, we note that $\|\mathbf{R}^* - \hat{\mathbf{R}}\|_1 \leq \epsilon$, and, moreover, that

$$V(\epsilon) - \frac{1}{n}\sum_{j=1}^{m} \textsc{Var}\left((\hat{R}_i)_{i\in\mathcal{S}_{j,1}}\right) = \frac{1}{n}\left[\left(\sum_{i=1}^{n}(R_i^*)^2 - \hat{R}_i^2\right) + \left(\sum_{j=1}^{m}\left[\frac{\sum_{i\in\mathcal{G}_j} R_i^*}{n_j}\right]^2 - \left[\frac{\sum_{i\in\mathcal{G}_j} \hat{R}_i}{n_j}\right]^2\right)\right].$$

Applying Lemma B.17 to the first term yields a bound of 2ϵ. For the second term, we note that

$$\sum_{j=1}^{m}\left|\frac{\sum_{i\in\mathcal{G}_j} R_i^*}{n_j} - \frac{\sum_{i\in\mathcal{G}_j} \hat{R}_i}{n_j}\right| \leq \sum_{i=1}^{n}|R_i^* - \hat{R}_i| \leq \epsilon,$$

and so we can apply Lemma B.17 to obtain a bound of 2ϵ on the second term as well. Combining these bounds gives the desired result. □

Algorithm 2 Piecewise quadratic function evaluation

Input: The collections $\mathbf{\Delta}$ and $\mathbf{K}$, as well as the L_1 "budget" ϵ.

Output: The value of $\Sigma(\Delta^*)$, where $\epsilon(\Delta^*) = \epsilon$.

```
Set $N \leftarrow$ length($\boldsymbol{\delta}$)
Initialize $i \leftarrow 1$                                        {Pointer to current position in $\boldsymbol{\delta}$}
Initialize $\varepsilon \leftarrow 0$                                        {Budget used so far}
Initialize $t \leftarrow 0$                          {Gap between current and next value of $\mathbf{\Delta}$}
Initialize $\Sigma \leftarrow 0$                                     {Change in sum of squares}
while $\varepsilon + K_i(\Delta_{i+1} - \Delta_i) < \epsilon$ and $i < N$ do
    Set $\varepsilon \leftarrow \varepsilon + K_i(\Delta_{i+1} - \Delta_i)$
    Set $\Sigma \leftarrow \Sigma - K_i \cdot \Delta_i \cdot (\Delta_{i+1} - \Delta_i) + \frac{K}{2} \cdot (\Delta_{i+1} - \Delta_i)^2$
    Set $i \leftarrow i + 1$
end while
if $K_i = 0$ then
    return $\Sigma$ {$K_i = 0$ if and only if the budget was exactly exhausted on the last iteration of the loop}
else if $i = N$ then
    return $\Sigma$             {$i = N$ if and only if we have made it to the end of $\boldsymbol{\delta}$, i.e., to $\delta = 0$}
else
    Set $t \leftarrow (\epsilon - \varepsilon)/K_i$
    Set
    Set $\Sigma \leftarrow \Sigma - K_i \cdot \Delta_i \cdot t + \frac{K}{2} \cdot t^2$
    return $\Sigma$
end if
```

Algorithm 3 Minimization algorithm (sum adjustment)

Input: The bounds $\boldsymbol{\ell}$ and $\mathbf{u}$, the estimates $\hat{\mathbf{R}}$, and the sum μ.
Output: A risk vector $\mathbf{R}$ of the form in Lemma B.6 minimizing $\|\mathbf{R} - \hat{\mathbf{R}}\|_1$ and satisfying $\sum_{i=1}^{n} R_i = \mu$, along with the ending k^{lwr}, k^{upr}, i^{lwr}, i^{upr}, t^{upr}, and t^{lwr}.

```
Set pts to be the concatenation of $\boldsymbol{\ell}$, $\hat{\mathbf{R}}$, and $\mathbf{u}$ in ascending order  {Points at which rates can change}
Set $\mathbf{R} \leftarrow \hat{\mathbf{R}}$  {Risk vector}
Set $n \leftarrow$ length(pts)
Set $i^{\text{lwr}} = 1$, $i^{\text{upr}} = n$  {Indices of next values at which rates might change}
Set $t^{\text{lwr}} \leftarrow$ pts[$i^{\text{lwr}}$], $t^{\text{upr}} \leftarrow$ pts[$i^{\text{upr}}$]  {Thresholds}
Set $k^{\text{lwr}} \leftarrow 0$, $k^{\text{upr}} \leftarrow 0$  {Number of active indices}
$D \leftarrow \mu - \sum_{i=1}^{n} R_i$  {Difference between required sum and actual sum}
if $D > 0$ then
    $t^{\text{nxt}} \leftarrow$ pts[$i^{\text{lwr}} + 1$]
    $d \leftarrow 0$  {Change in sum from moving from $t^{\text{lwr}}$ to $t^{\text{nxt}}$}
    while $D > d$ and $i^{\text{lwr}} < n$ do
        $D \leftarrow D - d$
        $i^{\text{lwr}} \leftarrow i^{\text{lwr}} + 1$
        $t^{\text{lwr}} \leftarrow t^{\text{nxt}}$
        $t^{\text{nxt}} \leftarrow$ pts[$i^{\text{lwr}} + 1$]
        if $t^{\text{nxt}}$ corresponds to an element of $\hat{\mathbf{R}}$ then
            $k^{\text{lwr}} \leftarrow k^{\text{lwr}} + 1$
        else if $t^{\text{nxt}}$ corresponds to an element of $\mathbf{u}$ then
            $k^{\text{lwr}} \leftarrow k^{\text{lwr}} - 1$
        end if
        $d \leftarrow k^{\text{lwr}}(t^{\text{nxt}} - t^{\text{lwr}})$
    end while
    if $i^{\text{lwr}} < n$ then
        $t^{\text{lwr}} \leftarrow t^{\text{lwr}} + \frac{D}{k^{\text{lwr}}}$
    end if
else if $D < 0$ then
    {Similar steps to the $D > 0$ case but adapted for the upper threshold}
end if

return $\mathbf{R}$, $k^{\text{lwr}}$, $k^{\text{upr}}$, $i^{\text{lwr}}$, $i^{\text{upr}}$, $t^{\text{upr}}$, and $t^{\text{lwr}}$
```

Algorithm 4 Minimization algorithm (single stratum)

```
Input: The bounds $\boldsymbol{\ell}$ and $\mathbf{u}$, the estimates $\hat{\mathbf{R}}$, and the sum $\mu$.
Output: The collections $\boldsymbol{\Delta}$ and $\mathbf{K}$.
Initialize pts, $\mathbf{R}$, $t^{\text{lwr}}$, $t^{\text{upr}}$, $k^{\text{lwr}}$, $k^{\text{upr}}$, $i^{\text{lwr}}$, and $i^{\text{upr}}$ as in Algorithm 3
Set $n \leftarrow$ length(pts)
Set $\varepsilon \leftarrow 0$    {Amount of budget expended so far}
Set $\Delta \leftarrow t^{\text{upr}} - t^{\text{lwr}}$    {Gap between thresholds}
Using Algorithm 3, update $\mathbf{R}$, $k^{\text{lwr}}$, $k^{\text{upr}}$, $i^{\text{lwr}}$, $i^{\text{upr}}$, $t^{\text{upr}}$, and $t^{\text{lwr}}$    {Ensure that sum is correct}
Set $\epsilon \leftarrow \epsilon - \|\mathbf{R} - \hat{\mathbf{R}}\|_1$, returning that the problem is infeasible if the result is negative
if $n = 1$ then
    return $\boldsymbol{\Delta} = (\Delta_0)$ and $\mathbf{K} = 0$
end if
$D^{\text{lwr}} \leftarrow k^{\text{lwr}}(\texttt{pts}[i^{\text{lwr}} + 1] - t^{\text{lwr}})$    {Cost of moving from lower threshold to next change point}
$D^{\text{upr}} \leftarrow k^{\text{upr}}(t^{\text{upr}} - \texttt{pts}[i^{\text{upr}} - 1])$    {Cost of moving from upper threshold and next change point}
$D \leftarrow \min(D^{\text{lwr}}, D^{\text{upr}})$    {Smaller of the costs}
$K \leftarrow 0$
while $i^{\text{upr}} - i^{\text{lwr}} > 1$ do
    Append $\Delta$ to $\boldsymbol{\Delta}$ and $K$ to $\mathbf{K}$
    if $D^{\text{lwr}} = D$ then
        $i^{\text{lwr}} \leftarrow i^{\text{lwr}} + 1$    {Increment lower active index}
        $t^{\text{lwr}} \leftarrow \texttt{dps}[i^{\text{lwr}}]$    {Update the lower threshold}
        if $t^{\text{lwr}}$ came from $\hat{\mathbf{R}}$ then
            $k^{\text{lwr}} \leftarrow k^{\text{lwr}} + 1$    {Increment $k^{\text{lwr}}$ if an index is activated}
        else if $t^{\text{lwr}}$ came from $\mathbf{u}$ then
            $k^{\text{lwr}} \leftarrow k^{\text{lwr}} - 1$    {Decrement $k^{\text{lwr}}$ if an index is deactivated}
        end if
        $D^{\text{upr}} \leftarrow D^{\text{upr}} - D^{\text{lwr}}$    {Calculate new upper active gap}
        $t^{\text{upr}} \leftarrow t^{\text{upr}} - \frac{D^{\text{lwr}}}{k^{\text{upr}}}$    {Calculate new upper threshold}
        $D^{\text{lwr}} \leftarrow k^{\text{lwr}}(\texttt{dps}[i^{\text{lwr}} + 1] - t^{\text{lwr}})$    {Calculate new lower active gap}
    else
        {Similar steps as in previous branch, adapted to the upper threshold}
    end if
    $D \leftarrow \min(D^{\text{lwr}}, D^{\text{upr}})$
    $\Delta \leftarrow t^{\text{upr}} - t^{\text{lwr}}$
    $K \leftarrow \frac{k^{\text{lwr}} \cdot k^{\text{upr}}}{k^{\text{lwr}} + k^{\text{upr}}}$
end while
Append 0 to $\boldsymbol{\Delta}$ and $\mathbf{K}$
return $\boldsymbol{\Delta}, \mathbf{K}$
```

Algorithm 5 Minimization algorithm (combining strata)

Input: The collections $\mathbf{\Delta}^{(0)}$, $\mathbf{K}^{(0)}$, $\mathbf{\Delta}^{(1)}$, and $\mathbf{K}^{(1)}$. **Output:** A pair $(\mathbf{\Delta}^{(2)}, \mathbf{K}^{(2)})$ representing the sum of functions $\epsilon^{(0)}(\Delta) + \epsilon^{(1)}(\Delta)$ and $\Sigma^{(0)}(\Delta) + \Sigma^{(1)}(\Delta)$.

```
Initialize $i_0 \leftarrow 0$, $i_1 \leftarrow 1$          {*Pointers to smallest indices not yet combined*}
Initialize $K_0 \leftarrow 0$, $K_1 \leftarrow 0$, $\Delta_0 \leftarrow 1$, $\Delta_1 \leftarrow 1$          {*Current values of the parameters*}
**while** $i_0 \leq$ `length`$(\mathbf{\Delta}_0)$ **and** $i_1 \leq$ `length`$(\mathbf{\Delta}_1)$ **do**
  **if** $\Delta^{(0)}_{i_0} > \Delta^{(1)}_{i_1}$ **then**
    Set $i_0 \leftarrow i_0 + 1$, $K_0 \leftarrow K^{(0)}_{i_0}$, and $\Delta_0 \leftarrow \Delta^{(0)}_{i_0}$
    Append $\Delta_0$ to $\mathbf{\Delta}^{(2)}$ and $K_0 + K_1$ to $\mathbf{K}^{(2)}$
  **else**
    Set $i_1 \leftarrow i_1 + 1$, $K_1 \leftarrow K^{(1)}_{i_1}$, and $\Delta_1 \leftarrow \Delta^{(1)}_{i_1}$
    Append $\Delta_1$ to $\mathbf{\Delta}^{(2)}$ and $K_0 + K_1$ to $\mathbf{K}^{(2)}$
  **end if**
**end while**
**return** $\mathbf{\Delta}^{(2)}$ and $\mathbf{K}^{(2)}$
```

Algorithm 6 Maximization algorithm (sum adjustment)

Input: The bounds $\boldsymbol{\ell}$ and $\mathbf{u}$, the estimates $\hat{\mathbf{R}}$, and the sum μ.
Output: A risk vector $\mathbf{R}$ of the form in Lemma B.11 minimizing $\|\mathbf{R} - \hat{\mathbf{R}}\|_1$ and satisfying $\sum_{i=1}^{n} R_i = \mu$, along with i^{lwr}, i^{upr}.

```
Set $\mathbf{R} \leftarrow \hat{\mathbf{R}}$          {*Risk vector*}
Set $n \leftarrow$ `length`$(\mathbf{R})$
Set $i^{\text{lwr}} = 1$, $i^{\text{upr}} = n$          {*Pivots*}
$D \leftarrow \mu - \sum_{i=1}^{n} R_i$          {*Difference between required sum and actual sum*}
**if** $D > 0$ **then**
  $d \leftarrow u_{i^{\text{upr}}} - R_{i^{\text{upr}}}$
  **while** $D > d$ **and** $i^{\text{upr}} > 0$ **do**
    $R_{i^{\text{upr}}} \leftarrow u_i$
    $D \leftarrow D - d$
    $i^{\text{upr}} \leftarrow i^{\text{upr}} - 1$
    $d \leftarrow u_{i^{\text{upr}}} - R_{i^{\text{upr}}}$
  **end while**
  $R_{i^{\text{upr}}} \leftarrow R_{i^{\text{upr}}} + D$
**else if** $D < 0$ **then**
  {*Similar steps to the $D > 0$ case but adapted for the lower pivot*}
**end if**

**return** $\mathbf{R}$, $i^{\text{lwr}}$, and $i^{\text{upr}}$.
```

Algorithm 7 Maximization algorithm (single stratum)

Input: The bounds $\boldsymbol{\ell}$ and $\mathbf{u}$, the estimates $\hat{\mathbf{R}}$, the sum μ, and the step size γ.
Output: The collection $\boldsymbol{\Sigma}$.

```
Initialize $\mathbf{R}$, $i^{\mathrm{lwr}}$, and $i^{\mathrm{upr}}$ as in Algorithm 6
Set $n \leftarrow \texttt{length}(R)$
Set $\varepsilon \leftarrow 0$    {Amount of budget expended so far}
Using Algorithm 6, update $\mathbf{R}$, $i^{\mathrm{lwr}}$, and $i^{\mathrm{upr}}$, $t^{\mathrm{upr}}$    {Ensure that sum is correct}
Set $\varepsilon \leftarrow \|\mathbf{R} - \hat{\mathbf{R}}\|_1$ and $\epsilon \leftarrow \epsilon - \varepsilon$, returning that the problem is infeasible if $\varepsilon > \epsilon$
while $\varepsilon \geq \gamma$ do
  Append $\infty$ to $\boldsymbol{\Sigma}$    {Sentinel value to indicate infeasibility}
  $\varepsilon \leftarrow \varepsilon - \gamma$
end while
$\Delta^{\mathrm{lwr}} \leftarrow R_{i^{\mathrm{lwr}}} - \ell_{i^{\mathrm{lwr}}}$
$\Delta^{\mathrm{upr}} \leftarrow u_{i^{\mathrm{upr}}} - R_{i^{\mathrm{upr}}}$
$\Delta \leftarrow \min(\Delta^{\mathrm{lwr}}, \Delta^{\mathrm{upr}})$
$\Sigma \leftarrow \sum_{i=1}^{n} R_i^2$    {Sum of squares}
while $i^{\mathrm{upr}} < i^{\mathrm{lwr}}$ do
  if $\Delta \geq \gamma - \varepsilon$ then
    $\Sigma \leftarrow \Sigma + 2\gamma \cdot (\gamma + [R_{i^{\mathrm{upr}}} - R_{i^{\mathrm{lwr}}}])$    {If next step would exceed step size, record change}
    Append $\Sigma$ to $\boldsymbol{\Sigma}$
    $R_{i^{\mathrm{lwr}}} \leftarrow R_{i^{\mathrm{lwr}}} + \gamma$, $R_{i^{\mathrm{upr}}} \leftarrow R_{i^{\mathrm{upr}}} + \gamma$    {Update risk vector at pivots}
    $\Delta \leftarrow \Delta - \gamma$, $\Delta^{\min} \leftarrow \Delta^{\min} - \gamma$, $\Delta^{\max} \leftarrow \Delta^{\max} - \gamma$    {Update gaps}
    $\varepsilon \leftarrow 0$    {Only adjust step size on first iteration}
  else
    if $\Delta^{\mathrm{lwr}} = \Delta$ then
      $\Sigma \leftarrow \Sigma + 2\Delta \cdot (\Delta + [R^{i^{\mathrm{upr}}} - R^{i^{\mathrm{lwr}}}])$
      $R_{i^{\mathrm{lwr}}} \leftarrow \ell_{i^{\mathrm{lwr}}}$, $R_{i^{\mathrm{upr}}} \leftarrow R_{i^{\mathrm{upr}}} + \Delta$
      $i^{\mathrm{lwr}} \leftarrow i^{\mathrm{lwr}} + 1$    {Increment lower active index}
    else
      {Similar steps as in previous branch, adapted to the upper pivot}
    end if
  end if
end while
Append $\Sigma$ to $\boldsymbol{\Sigma}$ until it has the appropriate length
return $\boldsymbol{\Sigma}$
```

Algorithm 8 Maximization algorithm (combining strata)

Input: The collections $\boldsymbol{\Sigma}^{(0)}$ and $\boldsymbol{\Sigma}^{(1)}$. **Output:** A collection $\boldsymbol{\Sigma}^{(2)}$ representing the (approximate) maximum objective across both collections.

```
Initialize $i_0 \leftarrow 0$, $i_1 \leftarrow 1$    {Pointers to indices currently being combined}
Initialize $M \leftarrow -\infty$    {Current maximum}
for $i = 0, \ldots, \epsilon/\gamma$ do
  $i_0 \leftarrow 0$, $i_1 \leftarrow i$
  while $i_0 < i$ do
    $M \leftarrow \Sigma^{(0)}_{i_0} + \Sigma^{(1)}_{i_1}$    {Find the maximum across gridpoints whose total budget is $i\gamma$}
    $i_0 \leftarrow i_0 + 1$, $i_1 \leftarrow i_1 - 1$
  end while
  Append $M$ to $\boldsymbol{\Sigma}^{(2)}$
  $M \leftarrow -\infty$
end for
return $\boldsymbol{\Sigma}^{(2)}$
```

C Estimation of the Non-Parametric Estimand by Linear Regression

To assess the quality of risk-adjusted regression as an estimator of the non-parametric estimand in (4), we construct synthetic datasets based on the NYPD data. On this synthetic dataset, we compare the true value of the risk adjusted disparities—as defined in (4)—to estimates from a risk-adjusted regression.

To construct the synthetic datasets, we first use the real NYPD data to estimate the probability of frisk, $\Pr(A = 1 \mid \tilde{X})$, conditional on the same observed pre-frisk covariates $\tilde{X}$ used to estimate risk in the main analysis, excluding race, suspected crime, and precinct.[25] We similarly estimate the probability a weapon is recovered on a frisked individual, $\Pr(W = 1 \mid \tilde{X}, A = 1)$, again conditional on the same pre-frisk covariates. In both cases, we use logistic regression to estimate the probabilities. The frisk model is trained on all data from 2008 and 2009, and the weapon-recovery model is trained on the subset of stops of the same in which an individual was frisked.

With the resulting covariate estimates $\hat{\boldsymbol{\beta}}_{\text{weapon}}$ and $\hat{\boldsymbol{\beta}}_{\text{frisk}}$, for the k-th iteration of the $m = 100$ simulation instances, we generate a population of $n =$ 1,000,000 observations as follows:

1. Sample noise terms $\boldsymbol{\epsilon}_{\text{weapon}} \sim \mathcal{N}(0, \frac{1}{50} \cdot I)$ and $\boldsymbol{\epsilon}_{\text{frisk}} \sim \mathcal{N}(0, \frac{1}{50} \cdot I)$, resulting in new weapon possession and frisk model coefficients $\boldsymbol{\beta}_{k,\text{weapon}} = \hat{\boldsymbol{\beta}}_{\text{weapon}} + \boldsymbol{\epsilon}_{\text{weapon}}$ and $\boldsymbol{\beta}_{k,\text{frisk}} = \hat{\boldsymbol{\beta}}_{\text{frisk}} + \boldsymbol{\epsilon}_{\text{frisk}}$.

2. Sample, with replacement, n covariate vector and race pairs $(\tilde{X}_i, C_i)_{i=1,\ldots,n}$ from the original NYPD dataset covering the years 2010 and 2011.

3. For all $i = 1, \ldots, n$:

   (a) Define the probability the i-th individual is frisked: $p_i = \text{logit}^{-1}(\boldsymbol{\beta}_{k,\text{frisk}}^\top \tilde{X}_i)$. Then sample $A_i \sim \text{Bernoulli}(p_i)$.

   (b) Define the probability the i-th individual has a weapon: $R_i = \text{logit}^{-1}(\boldsymbol{\beta}_{k,\text{weapon}}^\top \tilde{X}_i)$. Then, sample $W_i \sim \text{Bernoulli}(R_i)$.

The above procedure produces a set of tuples, $\Omega^j = \{(\tilde{X}_i, C_i, A_i, R_i, W_i)\}_{i=1,\ldots,n}$, each of which is a synthetic population. On this population, we compute the ground-truth risk-adjusted disparities via Eq. (4).[26] These disparities are typically non-zero; note, though, that there is no disparate *treatment* in this example since, by construction, the probability p_i of frisking an individual does not explicitly depend on race. The disparate impact we measure in this scenario thus arises from decisions that are simply not appropriately tailored to risk, as in *Griggs*.

We next evaluate the ability of our own risk-adjusted regression to recover the ground-truth disparate impact. To do so, we first divide Ω^j into two random sets of equal size, Ω_1^j and Ω_2^j. On Ω_1^j, we use logistic regression to estimate the probability that frisked individuals are found to have a weapon,[27] based on the pre-frisk covariates $\tilde{X}$. This model is trained on the subset of stops in Ω_1^j in which a frisk was carried out. We then use this fitted model to predict *ex ante* risk $\hat{R}_i$ for every stop in Ω_2^j. Lastly, we fit a linear risk-adjusted regression on Ω_2^j to estimate the Black-white and Hispanic-white disparities.

We note that to simulate real-world settings, we have ensured that, by construction, the linear probability model used to estimate disparities is misspecified. Nevertheless, it is a reasonably robust estimator of the true disparity, as can be seen in Figure 10. For both Black and Hispanic pedestrians in the synthetic data, our risk-adjusted estimates of disparate impact are very close to the ground-truth estimands, with model misspecification leading to slight overestimates in some instances, and slight underestimates in others.

[25] We exclude race to demonstrate that disparate impact can occur in the absence of disparate treatment. We exclude suspected crime and precinct because the large number of strata for these covariates leads to sampling difficulties.

[26] The definition in Eq. (4) involves conditioning on risk, which is continuous. This is problematic on a finite population, as there may be only one individual with any given risk. Binning risks avoids this issue, and, in this case, we get nearly identical values for any appropriately small bin size.

[27] To avoid risk estimates not being well defined when certain feature levels appear in Ω_2^j but not Ω_2^j, the models are fit using penalized maximum likelihood with the `glmnet` package and an elastic net penalty of $\alpha = \frac{1}{10}$.

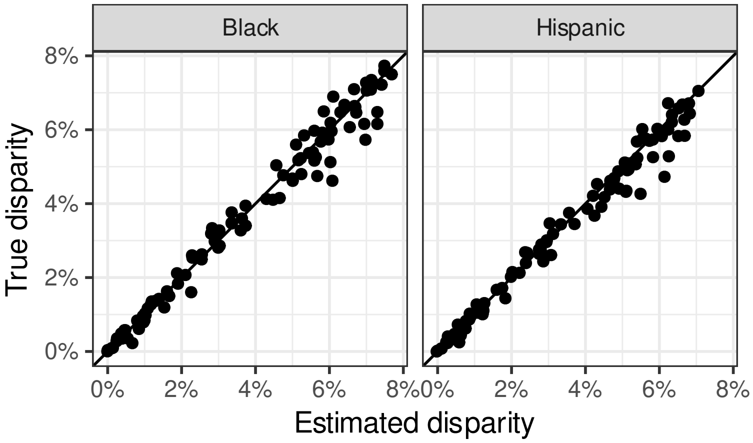


Figure 10: *Estimates of disparate impact using risk-adjusted regressions on* $m = 100$ *synthetic datasets, each consisting of* $n = 1{,}000{,}000$ *simulated stops. Despite model misspecification, the risk-adjusted regressions coefficient is a reasonably robust estimator of the true disparity.*

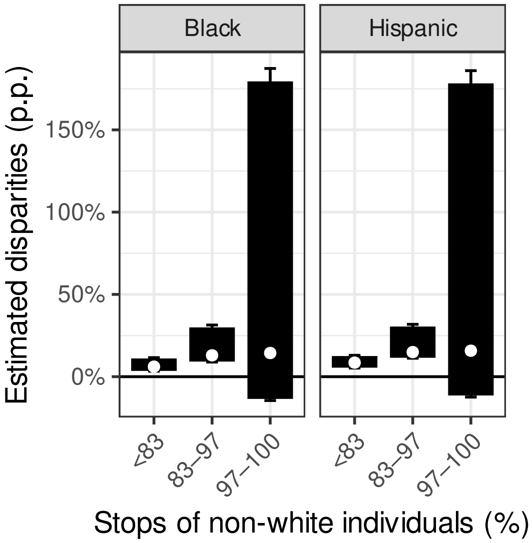


Figure 11: *Estimates of disparate impact across precincts stratified by racial composition of stops, with the black bands showing the range of possible estimates if the odds ratio of estimated risk and true risk differ by at most* $\epsilon = 0.7$ p.p.

D Accounting for factors beyond risk

Our analysis in the main text implicitly assumed that risk of weapon possession is the only legitimate consideration for carrying out a frisk—an assumption motivated by the fact that a frisk is legally allowed only to ensure officer safety. Officers, however are not required to frisk every individual who they are legally allowed to frisk (i.e., the law only sets a lower bar), leaving open the possibility that risk is not the only factor officers consider when making frisk decisions.[28] For example, in theory, resource-constrained officers might choose to frisk only the riskiest individuals they stop, effectively setting the bar to frisk individuals higher than the law demands. If resources differ across neighborhoods, which in turn correlate with racial composition, then the risk-adjusted disparities we see may accordingly have a policy-relevant justification.

In Figure 11, we aim to account for this possibility by repeating our analysis on three subsets of stops stratified by geography. Specifically, we split the 76 police precincts in our data into three bins based on the racial composition of stopped individuals. Across strata, we find qualitatively similar estimates of disparate impact, corroborating our main results. These within-strata estimates are also reasonably robust to mismeasurement of risk, as indicated by the black bands, although less so than our main result—a 0.7 p.p. average absolute difference in risk would potentially be sufficient to change the sign of the disparity in the most non-white neighborhoods, or to mask a disparity roughly ten times larger than our estimate.

That factors beyond risk may justifiably inform frisk decisions can be viewed as a form of omitted-variable bias, but one distinct in kind from that typically considered in studies of discrimination. Our disparate impact analysis is predicated on the understanding that risk, appropriately defined and estimated, captures nearly all policy-relevant considerations. Although it can be important in practice to accommodate exceptions when there is a clearly articulated rationale—as we have done above—care must be taken not to blindly adjust for every available factor, lest one re-introduce included-variable bias.

[28] In *Floyd*, the court found that officers at times frisked individuals even in the absence of safety concerns, in violation of the Fourth Amendment of the U.S. Constitution (Goel et al., 2016b).

E Additional figures

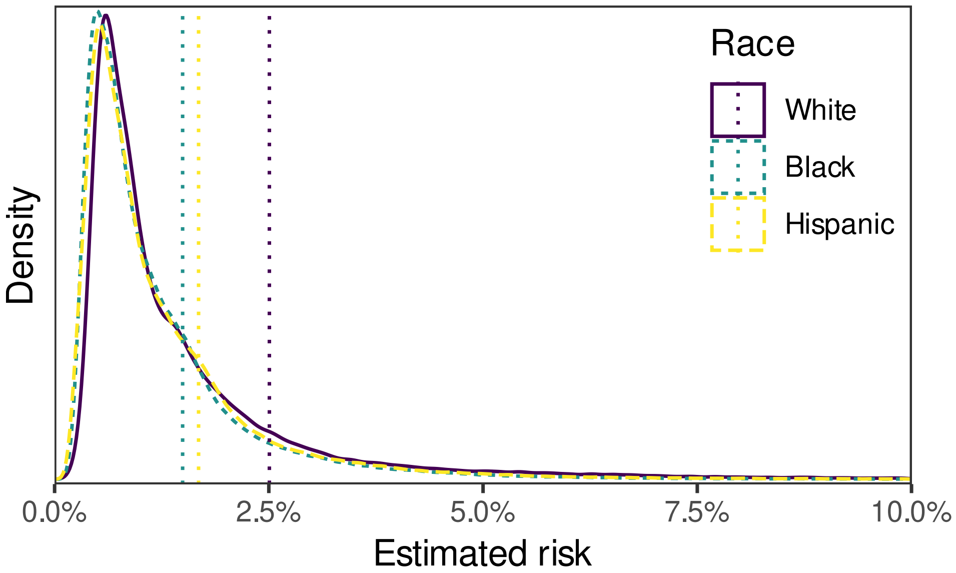


Figure 12: *Distributions of the estimated* ex ante *risk of probability of carrying a weapon for stopped pedestrians, on a log scale. The vertical lines indicate each group's average risk (i.e., the estimated rate at which stopped members of the group carry weapons): 2.7% for white pedestrians, 1.5% for Black pedestrians, and 1.7% for Hispanic pedestrians.*

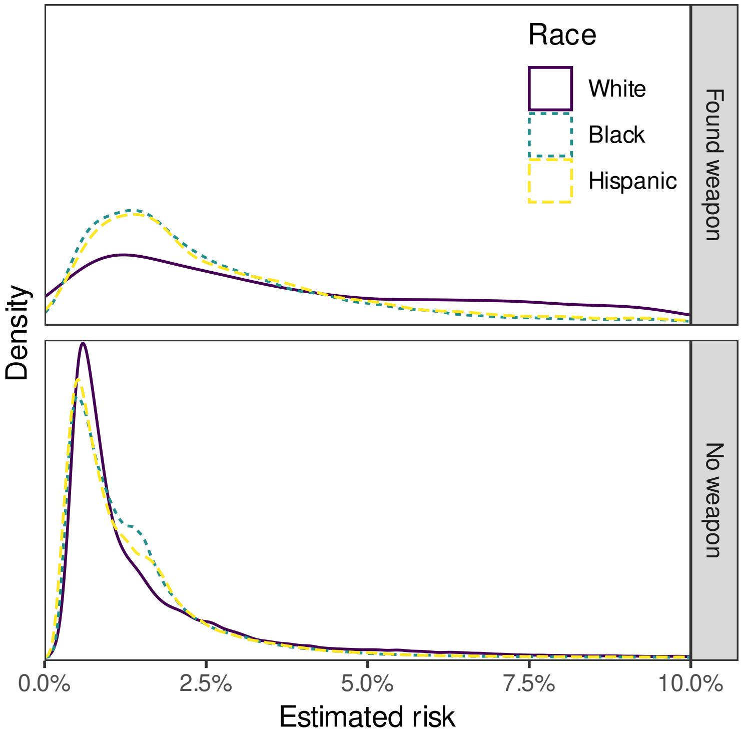
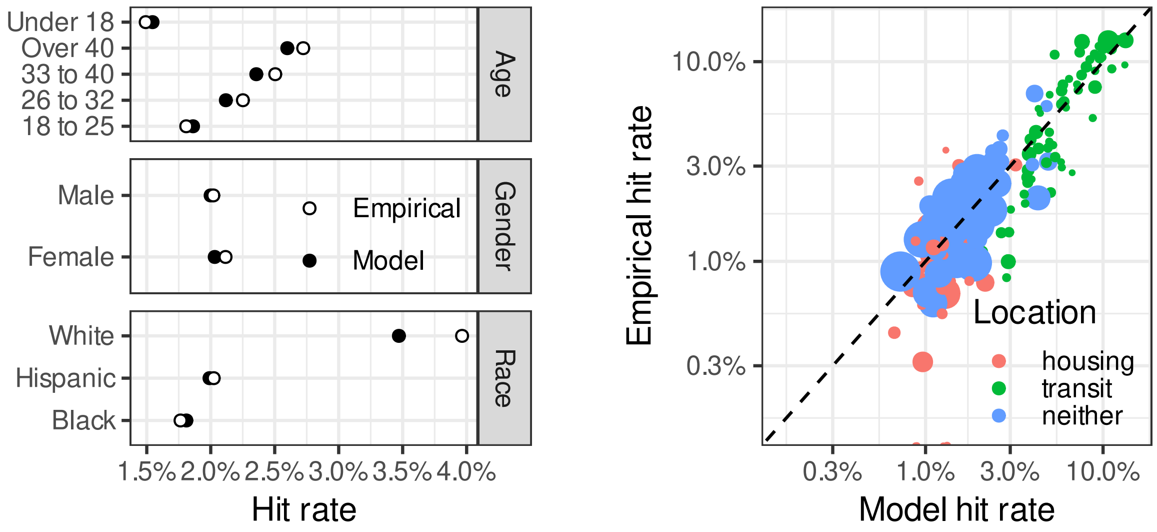


Figure 13: *Distributions of the estimated* ex ante *risk of probability of carrying a weapon for stopped pedestrians, on a log scale among frisked individuals, faceted by whether or not a weapon was found.*



(a) Demographic groups

(b) Locations

Figure 14: *Comparison of model-predicted versus empirical weapon recovery rate ("hit rate"). Figure 14a shows that the model-predicted hit rates are close to their empirical counterparts, conditional on values of age, race, and gender. In Figure 14b, stops are binned by precinct and stop location. Points are plotted for each bin with more than 100 stops, sized by the number of stops, with colors representing the stop location type: transit, housing, or other. The plotted points are near the diagonal, suggesting that the outcome model is well-calibrated and predicts well over the full range of hit rates and frisk rates, respectively. The model itself achieves an AUC of 81% on the second half of the data.*

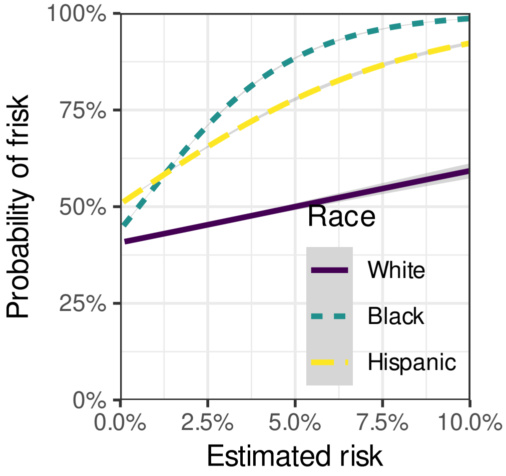


Figure 15: *Frisk rates vs. risk (where risk has been estimated without race as a covariate), as estimated via logistic regression curves fit separately for each race group. Across risk levels, stopped Black and Hispanic pedestrians are frisked substantially more frequently than comparably risk white individuals, indicative of disparate impact.*

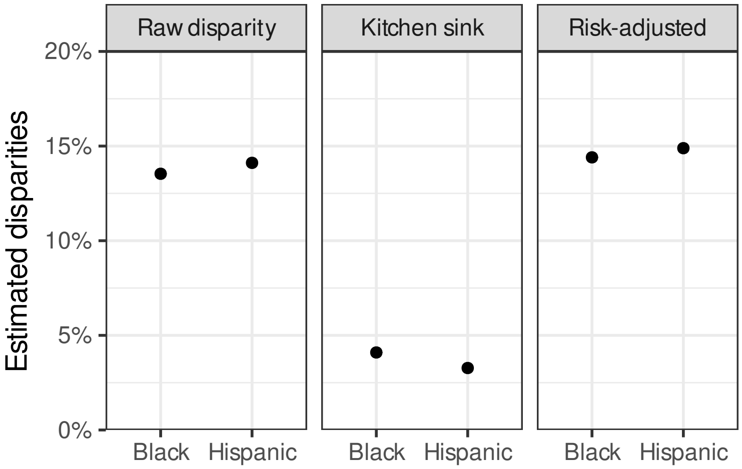


Figure 16: *Racial gaps in frisk rates adjusting for different sets of covariates, where the y-axis shows the percentage point difference relative to stopped white individuals. The left panel shows the raw disparities in frisk rates. As a measure of discrimination, raw disparities suffer from omitted-variable bias: there may, in theory, be legitimate reasons why Black and Hispanic pedestrians are more likely to be frisked. The middle panel shows the estimated race effects in a kitchen-sink regression, adjusting for all pre-frisk covariates. These estimates suffer from included-variable bias because they adjust for features that are correlated with race but unrelated to risk. The right panel shows the results of our risk-adjusted regression—where risk has been estimated without race as a covariate—adjusting exclusively for estimated risk of weapon possession. In all cases, estimated standard errors are less than 0.2 percentage points, and so are not visible in the plot.*

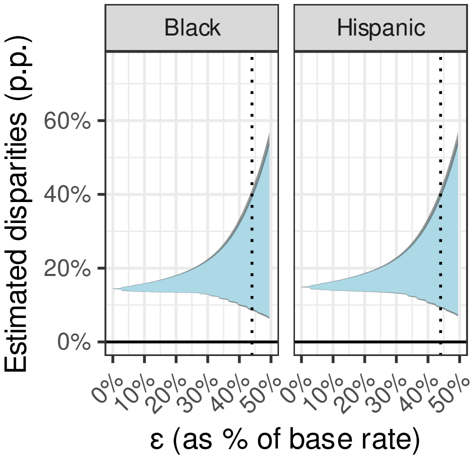


Figure 17: *Sensitivity of the risk-adjusted disparities in frisk decisions to mismeasurement of risk, where risk has been estimated without race as a covariate. The blue bands bound our estimates of disparate impact as a function of the average absolute difference between the true and estimated risks ϵ, relative to the base rate (1.7%). The dotted line at 44% ($\epsilon = 0.7$ p.p.) corresponds to a simulated situation with severe confounding. The grey bands represent 95% percentile bootstrapped confidence intervals ($N = 1000$; Zhao et al., 2019). The step size for the grid search over group-level total risks was 0.05 p.p. for all groups, and $\gamma = 0.01$ p.p. was the approximation parameter for the maximization routine.*

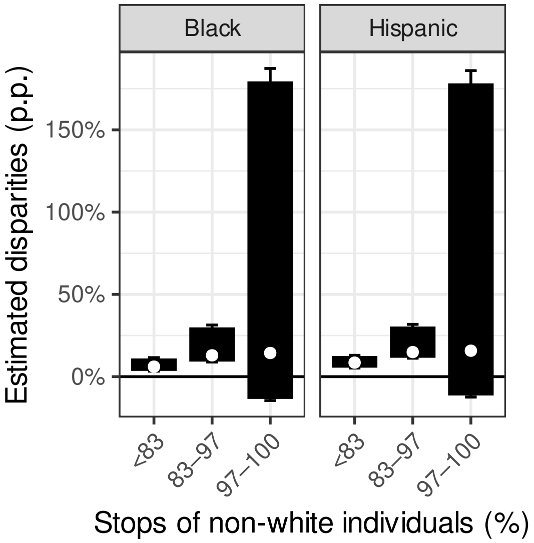


Figure 18: *Estimates of disparate impact across precincts stratified by racial composition of stops, with the black bands showing the range of possible estimates if the odds ratio of estimated risk and true risk differ by at most* $\epsilon = 0.7$ p.p. *Here, risk is estimated without race as a covariate.*

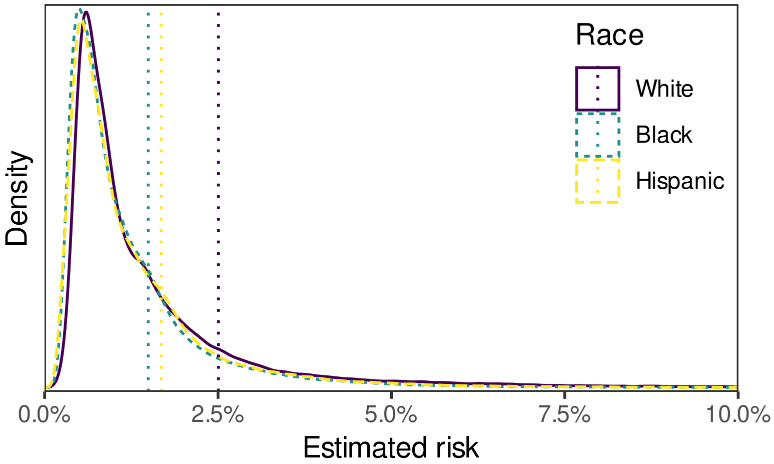


Figure 19: *Distributions of the estimated* ex ante *risk of carrying a weapon for stopped pedestrians, on a log scale and estimated without race as a covariate. The vertical lines indicate each group's average risk (i.e., the estimated rate at which stopped members of the group carry weapons): 2.7% for white pedestrians, 1.5% for Black pedestrians, and 1.7% for Hispanic pedestrians.*

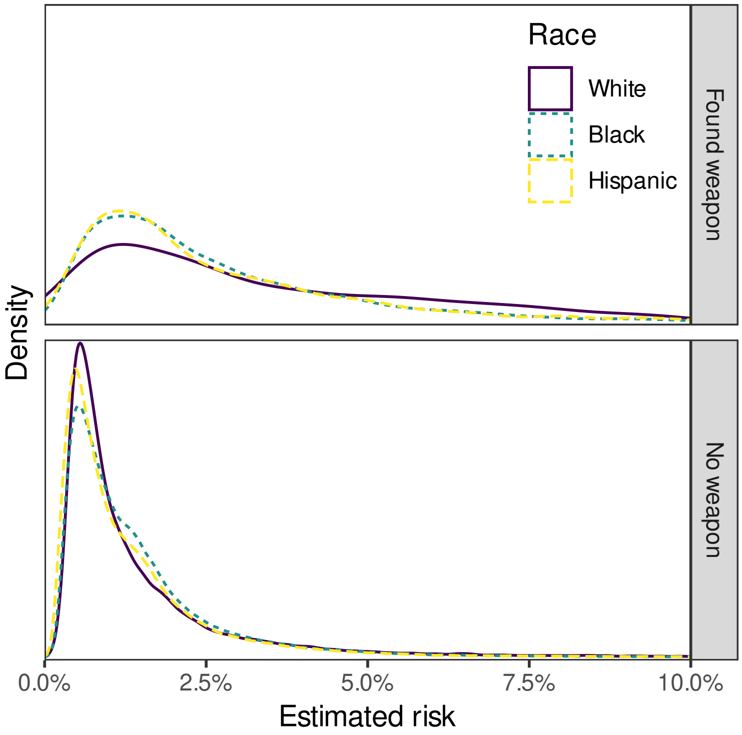
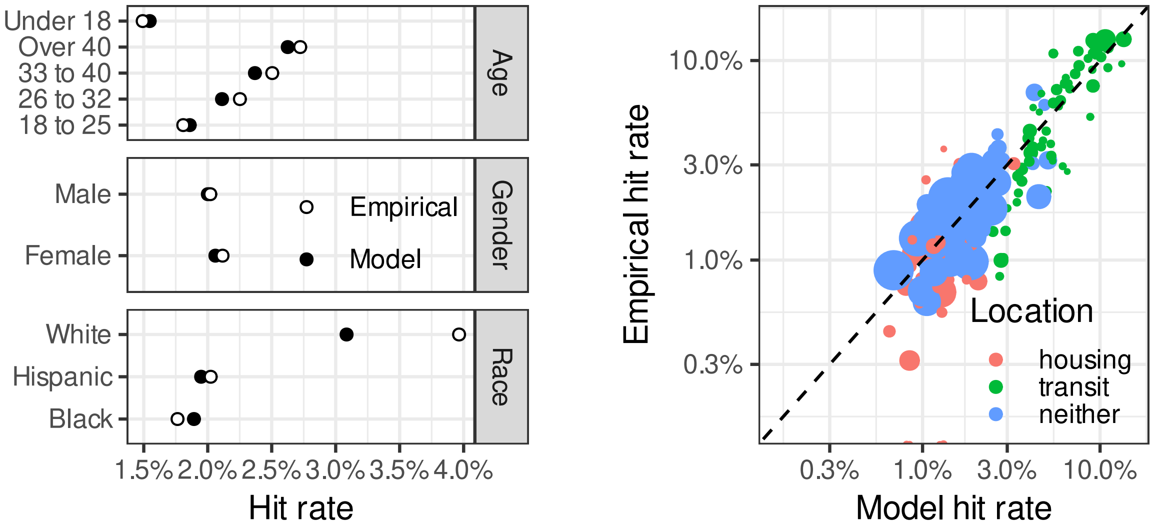


Figure 20: *Distributions of the estimated* ex ante *risk of probability of carrying a weapon—estimated without race as a covariate—for stopped pedestrians, on a log scale among frisked individuals, faceted by whether or not a weapon was found.*



(a) Demographic groups

(b) Locations

Figure 21: *Comparison of model-predicted versus empirical weapon recovery rate ("hit rate"), where the risk model does not include race as a covariate. Figure 21a shows that the model-predicted hit rates are close to their empirical counterparts, conditional on values of age, race, and gender. In Figure 21b, stops are binned by precinct and stop location. Points are plotted for each bin with more than 100 stops, sized by the number of stops, with colors representing the stop location type: transit, housing, or other. The plotted points are near the diagonal, suggesting that the outcome model is well-calibrated and predicts well over the full range of hit rates and frisk rates, respectively. The model itself achieves an AUC of 81% on the second half of the data.*

References

David S. Abrams. The law and economics of stop-and-frisk. *Loyola University Chicago Law Journal*, 46: 369–381, 2014.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

David Arnold, Will S Dobbie, and Peter Hull. Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research, 2020.

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in algorithms. In *AEA Papers and Proceedings*, volume 111, pages 49–54, 2021.

Kenneth Arrow. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.

Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Ian Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in Biology and Medicine*, 48(1):68–S87, 2005.

Ian Ayres. Testing for discrimination and the problem of "included variable bias". *Working paper*, 2010. Available at http://islandia.law.yale.edu/ayers/ayresincludedvariablebias.pdf.

Ana I Balsa, Thomas G McGuire, and Lisa S Meredith. Testing for statistical discrimination in health care. *Health Services Research*, 40(1):227–252, 2005.

Gary S Becker. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, 101 (3):385–409, 1993.

James A Berkovec, Glenn B Canner, Stuart A Gabriel, and Timothy H Hannan. Mortgage discrimination and FHA loan performance. In *Mortgage Lending, Racial Discrimination, and Federal Policy*, pages 289–305. Routledge, 2018.

J Aislinn Bohren, Peter Hull, and Alex Imas. Systemic discrimination: Theory and measurement. Technical report, National Bureau of Economic Research, 2022.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Shubham Chaudhuri and Rajiv Sethi. Statistical Discrimination with Peer Effects: Can Integration Eliminate Negative Stereotypes? *The Review of Economic Studies*, 75(2):579–596, 04 2008. ISSN 0034-6527. doi: 10.1111/j.1467-937X.2008.00468.x. URL https://doi.org/10.1111/j.1467-937X.2008.00468.x.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 82(1):39–67, 2020.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023. URL http://jmlr.org/papers/v24/22-1511.html.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593, 2020.

Will Dobbie, Jacob Goldin, and Crystal S Yang. The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2): 201–240, 2018.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, pages 1–13, 2022.

Hadi Elzayn, Evelyn Smith, Thomas Hertz, Arun Ramesh, Jacob Goldin, Daniel E Ho, and Robin Fisher. *Measuring and mitigating racial disparities in tax audits*. Stanford Institute for Economic Policy Research (SIEPR), 2023.

Thomas J Espenshade, Chang Y Chung, and Joan L Walling. Admission preferences for minority students, athletes, and legacies at elite universities. *Social Science Quarterly*, 85(5):1422–1446, 2004.

Jeffrey Fagan. Report of Jeffrey Fagan in the case of Floyd v. the City of New York, 2010. Available at https://www.law.columbia.edu/sites/default/files/microsites/policing-litigation-conference/files/Fagan%20Report%20with%20Technical%20Appendices.pdf.

Floyd v. City of New York. 959 F. Supp. 2d 540, S.D.N.Y, 2013.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 9:26–48, 2022.

Nikhil Garg, Hannah Li, and Faidra Monachou. Standardized tests and affirmative action: The role of bias and variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 261–261, 2021.

Andrew Gelman, Jeffrey Fagan, and Alex Kiss. An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479):813–823, 2007.

Sharad Goel, Justin M Rao, and Ravi Shroff. Personalized risk assessments in the criminal justice system. *The American Economic Review*, 106(5):119–123, 2016a.

Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 10, 2016b.

Sharad Goel, Maya Perelman, Ravi Shroff, and David Sklansky. Combatting police discrimination in the age of big data. *New Criminal Law Review*, 20, 2017.

Sharad Goel, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research Handbook on Big Data Law*. Edward Elgar Publishing, 2021.

D James Greiner and Donald B Rubin. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785, 2011.

Joshua Grossman, Julian Nyarko, and Sharad Goel. Reconciling legal and empirical conceptions of disparate impact, 2023a. Available at https://5harad.com/papers/disparate-impact.pdf.

Joshua Grossman, Sabina Tomkins, Lindsay Page, and Sharad Goel. The disparate impacts of college admissions policies on asian american applicants. *Working paper*, 2023b.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Melody Huang and Samuel D Pimentel. Variance-based sensitivity analysis for weighting estimators result in more informative bounds. *arXiv preprint arXiv:2208.01691*, 2022.

Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3): 771–800, 2020.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1), 2001.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages pp 481–492. University of California Press, Berkeley-Los Angeles, Calif., 1951.

John MacDonald and Steven Raphael. The effect of scaling back punishment in racial disparities in criminal case outcomes. *University of California Berkeley Working Paper*, 2019.

Sandra G Mayson. Bias in, bias out. *Yale Law Journal*, 128:2218, 2018.

Cory McCartan, Jacob Goldin, Daniel E Ho, and Kosuke Imai. Estimating racial disparities when race is not observed. *arXiv preprint arXiv:2303.02580*, 2023.

John Monahan and Jennifer L Skeem. Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513, 2016.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

Edmund S Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4): 659–661, 1972.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4(7):736–745, 2020.

Solomon W Polachek. Earnings over the life cycle: The Mincer earnings function and its applications. *Foundations and Trends in Microeconomics*, 4(3):165–272, 2008.

Manish Raghavan. What should we do when our ideas of fairness conflict? *Communications of the ACM*, 67(1):88–97, 2023.

M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal charging and its sentencing consequences. *U of Michigan Law & Econ, Empirical Legal Studies Center Paper*, (12-002), 2012.

Paul R Rosenbaum. *Observational studies*. Springer, 2002.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45:212–218, 1983a.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.

Sartaj Sahni. Computationally related problems. *SIAM Journal on computing*, 3(4):262–279, 1974.

SFFA v. Harvard. Students for Fair Admissions, Inc., Petitioner, v. President and Fellows of Harvard College. Students for Fair Admissions, Inc., Petitioner, v. University of North Carolina, et al., 2023. https://www.supremecourt.gov/opinions/22pdf/20-1199_l6gn.pdf.

Ravi Shroff. Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 5(3):189–196, 2017.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Scott Smart and Joel Waldfogel. A citation-based test for discrimination at economics and finance journals. Technical report, National Bureau of Economic Research, 1996.

Tyler J VanderWeele and Whitney R Robinson. Confounding and mediating variables. *Epidemiology*, 25(4): 473–484, 2014.

Washington v. Davis. 426 U.S. 229, 1976.

Yao Zhang and Qingyuan Zhao. Bounds and semiparametric inference in l^∞- and l^2-sensitivity analysis for observational studies. *arXiv preprint arXiv:2211.04697*, 2022.

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.