# Estimating Racial Disparities in Sentencing Using Knox, Lowe, and Mummolo's Method

Abby Steckel

December 16, 2021

**Advisor: Professor Peter Aronow, Department of Political Science, Yale University**

A major challenge in studies of discrimination in the U.S. criminal legal system is the selection bias present in observational data. In their 2019 paper, "Administrative Records Mask Racially Biased Policing," Dean Knox, Will Lowe, and Jonathan Mummolo demonstrated that prior research on police use of force underestimated racial disparities by failing to account for the fact that whether someone is included in the sample — whether they are stopped by police — is a post-treatment variable.[1] Knox et al introduced a mediating variable representing the proportion of racially discriminatory stops, and used this quantity as a sensitivity parameter to compute bounds on the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT) among the sample population. This paper describes my attempt to use Knox, Lowe, and Mummolo's method to obtain bounds on the proportion of racially discriminatory sentences in a federal sentencing dataset. Although the racism of the criminal justice system affects nonwhite people of different races and ethnicities, this paper focuses on the disparate treatment of white versus Black people. I consider sentencing practices to be discriminatory if there is a difference in the probability that a Black versus a white person will receive a sentence exceeding a particular length, holding other aspects of their case equal. Note, however, that the motivation for applying Knox et al's method is the fact that it is extremely difficult to control for all the aspects of a case that are related to systemic racism in America. I don't expect that this analysis will comprehensively document racial bias in sentencing. Rather, my goal is to compare a naive estimate of the proportion of discriminatory sentences to the result obtained with Knox et al's method.

Consistent with Knox et al's findings for police use of force, I find that for a large sentencing dataset, when we assume that a person's inclusion in the sample is affected by their perceived race, then the estimated racial disparity in sentencing decisions is larger than when we unrealistically assume that there is no discrimination in sample selection (ie no discrimination in the conviction process). In an effort to more realistically meet Knox et al's model assumptions, I also computed treatment effect bounds for a subset of the data involving only immigration cases. The resulting bounds were wide and did not conclusively show that the naive treatment effect underestimates the proportion of discriminatory immigration sentences.

In their October 2013 article analyzing the effects of relaxing federal mandatory sentencing guidelines, Sonja B. Starr and M. Marit Rehavi criticized studies that analyzed disparities in sentencing and controlled for case characteristics at the time of sentencing. Starr and Rehavi observed that "pre-sentencing decision-making can have substantial sentence-disparity consequences." In applying Knox et al's methodology to sentencing analysis, I consider Starr and Rehavi's statement and in particular look at how observed sentencing disparities vary when the pre-sentencing conviction decision ranges from entirely independent of race to entirely discriminatory against Black people.

## Variables

I used data from the U.S. Sentencing Commission's Monitoring of Federal Criminal Sentences series from 2000 to 2008.[2] I chose this dataset because it was used in a 2011 paper about federal sentencing disparities that identified selection bias as a main source of estimation error.[3] To make the size of the dataset more manageable for Knox et al's resampling procedure, I performed the first part of my analysis using only data from years 2006 to 2008. For these three years and for the variables that I considered, there were 165,416 cases. 45,239 of these cases involved Black defendants and 120,177 involved white defendants.

The following is a description of the variables that I used to model sentence outcomes. Please see the appendix for a description of how I prepared this data from the files provided by the U.S. Sentencing Commission.

Response variable:

- SENTENCE: Total prison sentence received, in months. In Knox et al's treatment effect model, the response functions are logit predictions of the probability that the outcome exceeds a certain threshold. I set the thresholds to be the first, second, and third quartiles of the SENTENCE data. For years 2006 to 2008, the first quartile is 15 months, the median is 37 months, and the third quartile is 77 months.

Treatment:

- race: 1 if individual is Black, 0 if they are white.

Covariates:

- AGE: Individual's integer age in years.
- MALE: 1 if individual is male, 0 if female.
- HSGED: 1 if individual's highest level of education is high school or a GED, 0 otherwise.
- SOMEPOSTHS: 1 if individual's highest level of education is one to three years of college, or some trade or vocational school.
- POSTHSDEGREE: 1 if individual's highest level of education is a trade or vocational degree, an associate's degree, a bachelor's degree, or a graduate degree.
- HISPANIC: 1 if individual is Hispanic, 0 if non-Hispanic.
- USCITIZEN: 1 if individual is a US citizen, 0 if not a citizen.
- SWB: 1 if case was adjudicated in the Southwest Border region, 0 otherwise. I included this indicator because Courtney Hagen's study suggested that interactions exist between primary offense type, citizenship, Hispanic ethnicity, and whether a case occured in the Southwest Border region. I followed Hagen's classification of the Southwest Border region as including districts 41, 42, 70, 74, and 84 among the 94 US District Courts.[4]
- CRIMINAL: 1 if individual has prior criminal history, 0 if not.
- CATEGORY2, CATEGORY3, CATEGORY4, CATEGORY5, CATEGORY6: These are indicators of the individual's "final criminal history category," where a higher category indicates that a person is considered to have a more serious criminal history. In the U.S. Sentencing Commission's guidelines, the criminal history category is used along with the present offense level to issue a recommended range of sentence lengths. [5]

- NOCOUNTS: Number of counts of conviction. This is a positive integer.
- POINTS: Non-negative integer count of criminal history points. This quantity is related to the severity of a person's criminal history.
- TRIAL: 1 if conviction was determined by a trial, 0 if it was settled by a plea agreement.
- PRIM_OFFENSE: Variable with 35 possible categories for the individual's primary offense, ranging from traffic violations to murder.
- YR2000, YR2001, YR2002, YR2003, YR2004, YR2005, YR2006, YR2007, YR2008: Indicators of the year in which the individual was sentenced.

Offense Frequencies

| . | Freq | . | Freq |
|---|---|---|---|
| drug trafficking | 66037 | embezzlement | 683 |
| immigration | 41510 | sexual abuse | 648 |
| firearms | 21236 | drug posession | 447 |
| fraud | 12761 | bribery | 392 |
| porn / prostitution | 4041 | arson | 174 |
| bank robbery | 3068 | auto theft | 140 |
| forgery | 2101 | kidnaping | 130 |
| money laundering | 1853 | murder | 123 |
| larceny | 1738 | civil rights offenses | 117 |
| admin of justice | 1715 | environmental | 96 |
| racketeering | 1594 | gambling | 86 |
| traffic and other | 1068 | natl defense | 82 |
| tax offenses | 990 | burglary | 66 |
| drug comm facils | 842 | food and drug | 48 |
| offenses in prisons | 797 | manslaughter | 23 |
| assault | 788 | antitrust violations | 22 |

Summary Statistics, 2006-2008 Cases

| | Mean | Minimum | Maximum | Standard Deviation |
|---|---|---|---|---|
| AGE | 34.577 | 16 | 87 | 10.404 |
| MALE | 0.897 | 0 | 1 | |
| HSGED | 0.287 | 0 | 1 | |
| SOMEPOSTHS | 0.131 | 0 | 1 | |
| POSTHSDEGREE | 0.059 | 0 | 1 | |
| HISPANIC | 0.457 | 0 | 1 | |
| USCITIZEN | 0.618 | 0 | 1 | |
| SWB | 0.322 | 0 | 1 | |
| CRIMINAL | 0.797 | 0 | 1 | |
| CATEGORY2 | 0.121 | 0 | 1 | |
| CATEGORY3 | 0.168 | 0 | 1 | |
| CATEGORY4 | 0.107 | 0 | 1 | |
| CATEGORY5 | 0.066 | 0 | 1 | |
| CATEGORY6 | 0.124 | 0 | 1 | |
| NOCOUNTS | 1.517 | 0 | 247 | 2.693 |
| POINTS | 3.904 | 0 | 91 | 4.985 |
| TRIAL | 0.046 | 0 | 1 | |
| YR2007 | 0.331 | 0 | 1 | |
| YR2006 | 0.329 | 0 | 1 | |
| race1 | 0.273 | 0 | 1 | |

# Summary of Knox, Lowe, and Mummolo

Knox et al define $Y_i$ as an indicator of whether force was used in encounter i. I define $Y_i$ as an indicator of whether the sentence length exceeds a specified threshold. $D_i$ represents the race of the individual in the ith case. Knox et al clarify that the causal inference task is *not* to estimate what a white individual's outcome would be if they were Black, which introduces an impossible counterfactual situation. Instead, the goal is to estimate the difference in the outcome of the ith case if it involved a *different person* with a different racial identity, holding observable covariates constant.

Central to Knox et al's analysis is the mediating variable $M_i$, which indicates whether an encounter is included in the sample. $M_i$ is a function of $D_i$. For Knox et al, $M_i(d)$ indicates "whether encounter i would have resulted in a stop if the civilian were of race $d$" (5). I define $M_i$ as an indicator of whether or not a case was included in the sample, i.e. whether an individual received *any* counts of conviction versus none. So $M_i(d)$ is whether the ith case would have resulted in a conviction if the defendant were of race $d$.

## Naive Estimator

Knox et al criticize standard methods of estimating racial disparities in police use of force. They say that common estimators violate the Stable Unit Treatment Value Assumption (SUTVA) by taking the difference in means of groups whose potential outcomes are not independent of treatment. Knox et al define a naive estimator: $\hat{\Delta} = \overline{Y_i|D_i = 1, M_i = 1} - \overline{Y_i|D_i = 0, M_i = 1}$ (5). They implicitly condition on covariates $X_i$. The naive estimator is the mean outcome for recorded cases involving Black people minus the mean outcome for recorded cases involving white people. The problem is that these groups of cases are not necessarily comparable. In the sentencing context, conditioning on inclusion in the sample of federal cases, potential sentences may systematically differ for people of different races. For example, suppose that white people only get convicted of a firearms charge if they fire a gun, whereas Black people get convicted just for possessing a weapon. Then, conditioning on inclusion in the sample ($M_i$=1), the probability that a Black person's sentence would have exceeded a particular length if they were white is different (lower) than the probability that a white person's sentence will exceed that threshold. This is to say that $Y_i$ is *not* necessarily independent of $D_i|M_i$.

I computed the naive estimators for the difference in probabilities that Black vs white individuals' sentences exceed the 1st, 2nd, and 3rd quartiles for sentence length. I estimated the response probabilities using a logistic regression where explanatory variables are race and all of the covariates described above. I computed the difference in means for a given threshold as the mean of the predicted probabilities when each individual's race is set to "black" minus the mean of the predicted probabilities when each individual's race is set to "white." Note that this computation wrongly assumes that controlling for $X_i$, every case in the sample has the same set of potential outcomes $Y_i(0)$, $Y_i(1)$. As stated in the previous paragraph, this is not necessarily true.

For each threshold, the naive treatment effect is positive. This means that even a downwardly biased estimate indicates that cases involving Black people are more likely than similar cases involving white people to receive a long sentence. The question that I will explore using Knox et al's method is how this naive treatment effect compares to the estimated treatment effect using a plausible value of $\rho$.

Naive Sentencing Disparity, 2006-2008 Cases, Model with All Covariates

| Sentence Threshold (Months) | Mean Difference in Probability of Exceeding Threshold |
|---|---|
| above 77 | 0.0371 |
| above 37 | 0.0496 |
| above 15 | 0.0297 |

## Assumptions

Knox et al define four principal strata based on the different possible combinations of values of $M_i(d = 1)$ and $M_i(d = 0)$. Every case, including those that aren't recorded, falls into one of these strata.

- $M_i(1) = M_i(0) = 1$: Knox et al call this an "always-stop" encounter. I will call it "always-convict." This is a case that results in inclusion in the sample whether the individual is white or Black.
- $M_i(1) = 1, M_i(0) = 0$: anti_Black stop (conviction).
- $M_i(1) = 0, M_i(0) = 1$: anti-white stop (conviction). I assume, like Knox et al, that there are no cases that fall into this stratum.
- $M_i(1) = M_i(0) = 0$: never-stop encounter (never-convict case).

Knox et al identify four assumptions necessary for computing their treatment effect estimators. I will discuss these assumptions and their plausibility with respect to the sentencing data.

**Assumption 1 (Mandatory Reporting):** $Y_i(d, 0) = 0$ for all i and for $d \in \{0, 1\}$.

This assumption says that if an encounter is not included in the sample (ie if $M_i = 0$), then the outcome of interest did not occur. That is, if a case did not result in a conviction and is not included in the federal sentencing dataset, then the case did not result in a federal sentence. This seems like a reasonable assumption.

**Assumption 2 (Mediator Monotonicity):** $M_i(1) \geq M_i(0)$ for all i.

We assume that if a case involving a white person is included in the sample (receives a conviction), then a case with the same characteristics that instead involves a Black person will also result in a conviction. This is equivalent to the assumption that there are no cases in the "anti-white" stratum where $M_i(1) = 0, M_i(0) = 1$. Given the long and well-documented history of anti-Blackness in the American criminal justice system, I feel comfortable assuming that federal prosecutors did not treat Black people with more leniency than white people.

**Assumption 3 (Relative Nonseverity of Racial Stops (Convictions))**:
$\mathbb{E}[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 1, X_i = x] \geq \mathbb{E}[Y_i(d, m)|D_i = d', M_i(1) = 1, M_i(0) = 0, X_i = x]$.

This says that in a case where both a Black and white person would be convicted, the expected sentence length is longer than the expected sentence length for a case that would only result in conviction if the defendant was Black. In other words, assuming that sentence length is positively related to a crime's severity, this says that always-convict cases involve more serious offenses than anti-Black conviction cases.

**Assumption 4 (Treatment Ignorability)**:

- a. With respect to potential mediator: $M_i(d) \perp\!\!\!\perp D_i | X_i$.
- b. With respect to potential outcomes: $Y_i(d, m) \perp\!\!\!\perp D_i | M_i(0) = m', M_i(1) = m'', X_i$.

Generally, Assumption 4(a) asserts that $X_i$ includes all relevant covariates that might be correlated with race and with the conviction indicator $M_i$. Although my model includes information about the severity of the offense and characteristics of the individual involved, it seems unlikely that I've included all the background information necessary for convictions to be uncorrelated with race conditional on $X_i$. For example, my $X_i$ does not include information about the defendant's income. Income influences the neighborhood that a person lives in, which in turn affects their likelihood of being arrested for a particular activity, and subsequently whether they are convicted and included in the sentencing dataset. More broadly, the ubiquity of systemic racism and its impact on people throughout their lives makes it extremely difficult to state that the conviction potential outcomes of Black vs white individuals are the same after controlling for just a handful of covariates. For these reasons, 4(a) probably doesn't hold for this dataset.

4(b) says that given that a case belongs to a particular principal stratum, and controlling for the observable covariates, $Y_i$ is independent of $D_i$. Knox et al say that "conditional on $X_i$, civilian race is 'as good as randomly assigned' to encounters, and officers encounter minority civilians in circumstances that are objectively no different from white encounters" (8).

Unfortunately, I believe that defining $M_i$ as whether or not an individual receives any counts of conviction may results in a violation of Assumption 4(b). This is because unlike Knox et al's binary mediating variable of whether an individual was stopped by police, whether or not a case resulted in a conviction does not fully describe the conviction decision, which is a post-treatment variable. In particular, suppose $M_i(0) = M_i(1) = 1$, so the ith case is in the always-convict group. Suppose that for Black individuals, a particular type of case is prosecuted as a felony, whereas white individuals are only prosecuted for misdemeanors. Then within the always-convict stratum, race is *not* randomly assigned. Treated observations (cases involving Black people) tend to have longer potential sentences.

Under what conditions is Assumption 4(b) plausible for this sentencing data? One possibility is to assume that $X_i$ captures the difference in conviction decisions among the sample population. In particular, controlling for primary conviction offense and the number of counts of conviction should place observations into roughly comparable groups with respect to offense severity. The problem with this is that these conviction-related covariates (TRIAL, NOCOUNTS, and PRIM_OFFENSE) are post-treatment variables. Controlling for post-treatment variables may introduce post-treatment bias. Montgomery, Nyhan, and Torres explain that broadly, when we control for post-treatment variables, we match control observations that have a particular characteristic without treatment with treatment observations that have a particular characteristic despite or because of treatment. These groups have systematically different potential outcomes.[6] For example, suppose a white person's primary offense is drug trafficking, for selling drugs on the street, while a Black person's primary offense is also drug trafficking, for possessing drugs that they allegedly intended to sell. Similar to the example of how the naive estimator could be biased due to differences in potential outcomes when $M_i(1) = 1$ versus when $M_i(0) = 1$, we can see that $Y_i$ may not be independent of $D_i|X_i$.

Alternatively, we can try to choose a subset of the data so that conviction decisions are more homogenous and potential outcomes more similar, controlling for the set of pre-treatment covariates. I propose that subsetting the 2000-2008 dataset to look only at cases that received immigration convictions makes $M_i$ more compatible with Assumption 4, although there may still be a violation due to unobserved differences between cases. Looking just at immigration cases, $M_i$ indicates whether a case resulted in an immigration conviction. While there are multiple possible immigration-related federal convictions,[7] these convictions are much more similar in severity than the 35 conviction categories in the original dataset. The interquartile range of sentences for immigration offenses ranges from 10 to 37 months, reflecting their relatively similar severity. Compare this to the range of sentences for drug trafficking, which has a first quartile of 30 months and a third quartile of 120 months.

I computed bounds for the following four variations on the full dataset and model:

- $ATE_{M=1}$ for all sentencing data from 2006 to 2008, using all covariates
    - This is the quantity that I originally computed, before considering potential problems with post-treatment covariates and the mediator strata not being fully specified.
- $ATE_{M=1}$ for all sentencing data from 2006 to 2008, using only pre-treatment covariates
    - This model still may violate Assumption 4 because it remains implausible that after controlling for the pre-treatment covariates, race is randomly assigned among cases.
- $ATE_{M=1}$ for immigration sentencing data from 2000 to 2008, using all covariates
- $ATE_{M=1}$ for immigration sentencing data from 2000 to 2008, using only pre-treatment covariates

Note that considering only immigration cases, the total sample size is 96,969, with 4,125 cases involving Black people and 92,844 involving white people. We will see that the bounds and confidence intervals for $ATE_{M=1}$ of the immigration cases are quite wide. Based on Knox et al's bounds equation, the large value of $Pr(D_i = 0|M_i = 1)$ (the probability that a given case in the immigration sample involves a white person) seems to contribute to the wide bounds on $ATE_{M=1}$. Additionally, the small value of $n_{treated}$ probably contributes to the wide confidence intervals for the immigration data.

Below are the variable means and naive ATEs for immigration cases only.

Summary Statistics, Immigration Cases

| | Mean | Minimum | Maximum | Standard Deviation |
|---|---|---|---|---|
| AGE | 33.144 | 16 | 82 | 8.799 |
| MALE | 0.945 | 0 | 0 | |
| HSGED | 0.128 | 0 | 0 | |
| SOMEPOSTHS | 0.039 | 0 | 0 | |
| POSTHSDEGREE | 0.016 | 0 | 0 | |
| HISPANIC | 0.910 | 0 | 0 | |
| USCITIZEN | 0.101 | 0 | 0 | |
| SWB | 0.704 | 0 | 0 | |
| CRIMINAL | 0.868 | 0 | 0 | |
| CATEGORY2 | 0.154 | 0 | 0 | |
| CATEGORY3 | 0.243 | 0 | 0 | |
| CATEGORY4 | 0.168 | 0 | 0 | |
| CATEGORY5 | 0.096 | 0 | 0 | |
| CATEGORY6 | 0.093 | 0 | 0 | |
| NOCOUNTS | 1.086 | 1 | 97 | 0.945 |
| POINTS | 4.428 | 0 | 49 | 4.309 |
| TRIAL | 0.013 | 0 | 0 | |
| YR2007 | 0.138 | 0 | 0 | |
| YR2006 | 0.136 | 0 | 0 | |
| YR2005 | 0.126 | 0 | 0 | |
| YR2004 | 0.111 | 0 | 0 | |
| YR2003 | 0.116 | 0 | 0 | |
| YR2002 | 0.082 | 0 | 0 | |
| YR2001 | 0.066 | 0 | 0 | |
| YR2000 | 0.072 | 0 | 0 | |
| RACE | 0.043 | 0 | 0 | |

Naive Sentencing Disparity, Immigration Cases, Model with All Covariates

| Sentence Threshold (Months) | Mean Difference in Probability of Exceeding Threshold |
|---|---|
| above 37 | 0.06 |
| above 21 | 0.0452 |
| above 10 | 0.0019 |

## Knox et al's Treatment Effect Estimators

Knox et al define $\rho$ as the probability that a case involving a Black person would not have been included in the sample if the person were white (3). They show that given a value for $\rho$, it is possible to compute non-parametric bounds for $ATE_{M=1}$ and a point estimate for $ATT_{M=1}$.

Knox et al define $ATE_{M=1}$ as $E[Y_i(1, M_i(1))|M_i = 1] - E[Y_i(0, M_i(0))|M_i = 1]$. For sentencing data, $ATE_{M=1}$ is the probability of sentence length exceeding the threshold among the sample (convicted) population, considering what would happen if each case involved a Black person, minus the probability of sentence length exceeding the threshold if each case involved a white person. Note that this quantity can't be computed from the data because it involves unobserved events. $Y_i(0, M_i(0))|M_i = 1$ refers to the outcome of the ith case in the sample if it involved a white person. But if this case was an anti-black conviction, then $M_i(0) = 0$ and we wouldn't actually have observed it.

Knox et al provide the following formula for computing nonparametric sharp bounds on $ATE_{M=1}$ for a given $\rho$ (11).

$$\mathbb{E}[\hat{\Delta}] + \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1](1 - Pr(D_i = 0|M_i = 1))$$
$$\leq ATE_{M=1} \leq$$
$$\mathbb{E}[\hat{\Delta}] + \frac{\rho}{1-\rho}(\mathbb{E}[Y_i|D_i = 1, M_i = 1] - max\{0, 1 + \frac{1}{\rho}\mathbb{E}[Y_i|D_i = 1, M_i = 1] - \frac{1}{\rho}\})$$
$$\times Pr(D_i = 0|M_i = 1) + \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1](1 - Pr(D_i = 0|M_i = 1))$$

Knox et al compute these bounds using a Monte Carlo procedure. Briefly, after fitting a logistic regression using all of the data, they simulate noisy coefficients by drawing from a multivariate normal distribution centered at the coefficients from the logistic regression, and with covariance matrix equal to the clustered covariance matrix of the regression coefficients. Then they use the simulated coefficients to predict responses for a subset of the data, and compute $ATE_{M=1}$ for those predictions.

Below is Knox et al's formula for $ATT_M = 1$.

$$ATT_{M=1} = \mathbb{E}[\hat{\Delta}] + \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1]$$

Knox et al compute $ATT_{M=1}$ using a similar Monte Carlo procedure to $ATE_{M=1}$, simulating noisy coefficient observations and averaging treatment effects over chunks of the data. Given a value of $\rho$, we can obtain a point estimate of $ATT_{M=1}$ from the sample.

## Results for All Convictions, 2006-2008

Table 1 and Table 2 compare the naive ATE to the bounds and confidence intervals for $ATE_{M=1}$ when $\rho$ = 0.32. This is the value of $\rho$ that Knox et al consider to be a conservative estimate of the true proportion of racially discriminatory stops in New York City, based on a 2007 study of the city's stop and frisk policy by Gelman, Fagan, and Kiss. I'm assuming that the true proportion of racially discriminatory federal convictions is the same as the true proportion of racially discriminatory stops in New York City. This is unrealistic, but this choice of $\rho$ can still provide a rough idea of the bias of the naive ATE, given that some proportion of cases would not have been included in the sample if the defendant was white.

Table 1: Average Treatment Effect among 2006-2008 Convictions ($ATE_{M=1}$), by Sentence Quartile, Full Model Specification

| | Naive ATE | Lower Bound | Upper Bound | Lower CI | Upper CI |
|---|---|---|---|---|---|
| over 77 months | 0.0372 | 0.0696 | 0.1345 | 0.0656 | 0.1393 |
| over 37 months | 0.0496 | 0.1041 | 0.2165 | 0.0988 | 0.2224 |
| over 15 months | 0.0297 | 0.1040 | 0.2150 | 0.0988 | 0.2196 |

[a] rho = 0.32

Table 1 shows that for each sentence length threshold, the entire 95% confidence interval for $ATE_{M=1}$ exceeds the naive ATE. In other words, the confidence interval that contains the true value of $ATE_{M=1}$ 95% of the time does not include $\hat{\Delta}$, confirming that $\hat{\Delta}$ underestimates the true racial disparity in sentence lengths.

Table 2: Average Treatment Effect among 2006-2008 Convictions ($ATE_{M=1}$), by Sentence Quartile, Pretreatment Covariates Only

| | Naive ATE | Lower Bound | Upper Bound | Lower CI | Upper CI |
|---|---|---|---|---|---|
| over 77 months | 0.0685 | 0.0978 | 0.1794 | 0.0931 | 0.1857 |
| over 37 months | 0.0854 | 0.1368 | 0.2904 | 0.1311 | 0.2979 |
| over 15 months | 0.0544 | 0.1267 | 0.2706 | 0.1212 | 0.2748 |

[a] rho = 0.32

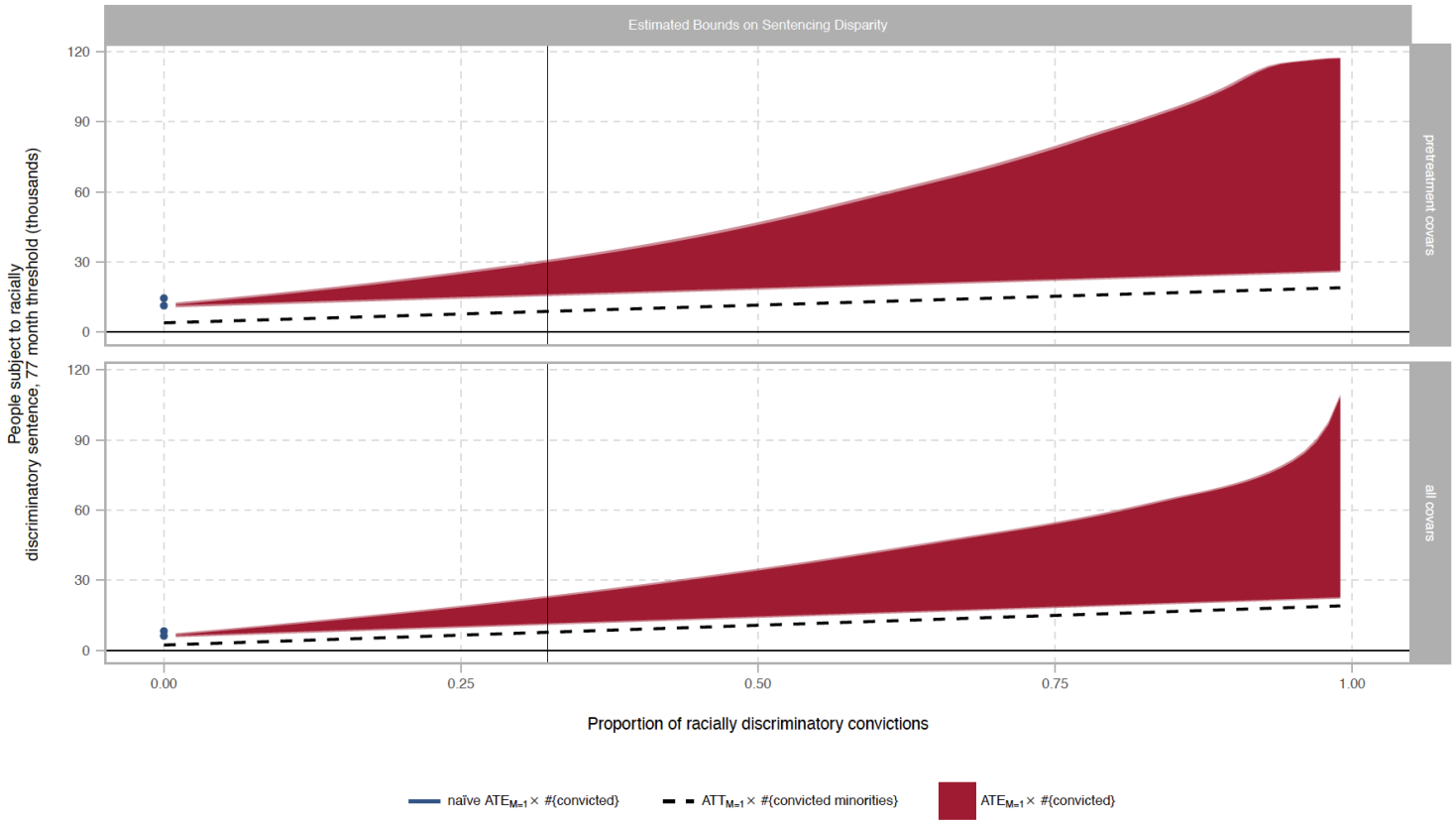
Table 2 compares the naive ATE to the $ATE_{M=1}$ when the response functions involve only the pre-treatment covariates, so we leave out information about number of counts of conviction, whether the case went to trial, and the primary conviction offense. Without adjusting for these post-treatment covariates, every value of the naive ATE and the $ATE_{M=1}$ intervals is larger than in Table 1. As shown in Table 3, race is positively correlated with the post-treatment covariates TRIAL and NOCOUNTS. Because PRIM_OFFENSE is a categorical variable, I defined an indicator of whether a charge would be considered violent. This indicator is not based on a statutory definition; it's just my rough attempt to evaluate the correlation of the post-treatment covariates with race. Table 3 shows that the prevalence of these more serious convictions does differ significantly by the perceived race of the defendant. To be clear: these tables do *not* indicate that there is a difference between the so-called "criminality" of white and Black people. Rather, the correlation between race and conviction severity reflects the impact of systemic racism prior to and during the conviction stage of a case, which is not captured by the mediating variable. The result of this correlation is that when we adjust for post-treatment covariates, the estimated treatment effect is smaller relative to the model with only pre-treatment covariates.

Table 3: Race is Predictive of Post-Treatment Covariates

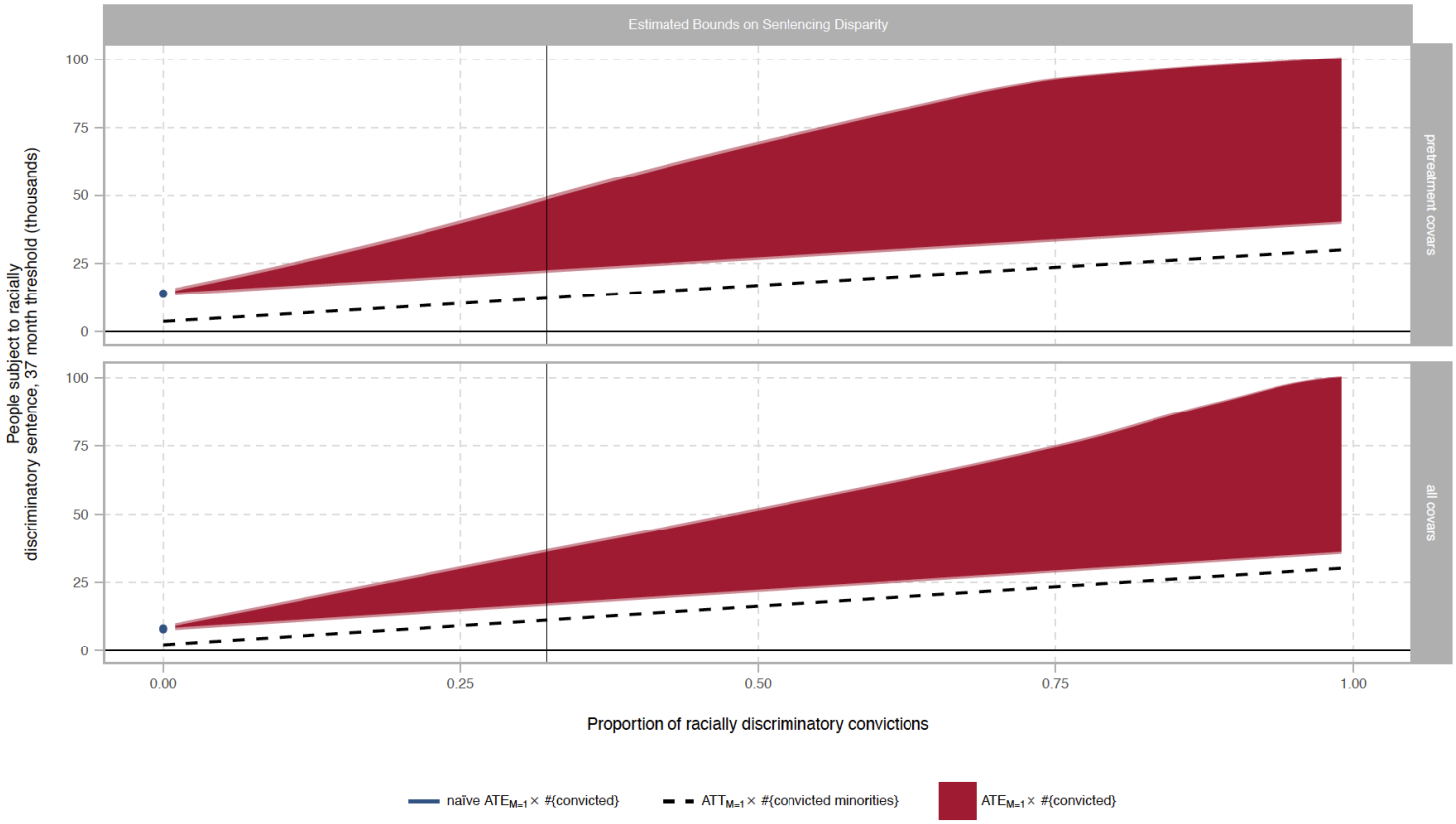
| | $Mean_{D_i=1}$ | $Mean_{D_i=0}$ | P Value |
|---|---|---|---|
| TRIAL | 0.075 | 0.035 | 0 |
| NOCOUNTS | 1.691 | 1.451 | 0 |

| | | | |
|---|---|---|---|
| SERIOUS CHARGE | 0.256 | 0.095 | 0 |

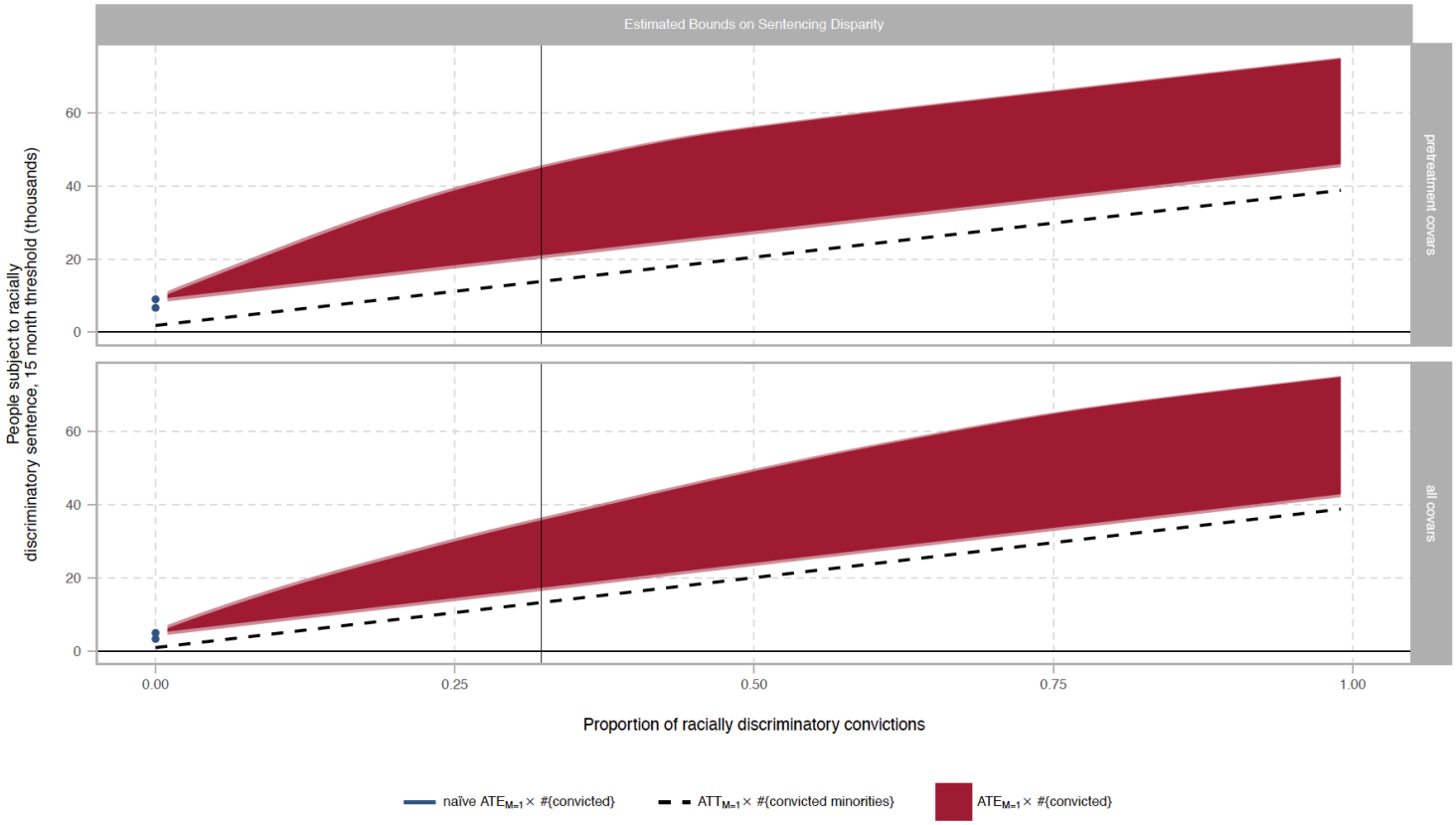
## Estimated Discriminatory Sentences Over 77 Months, 2006-08 Cases



## Estimated Discriminatory Sentences Over 37 Months, 2006-08 Cases



**Estimated Discriminatory Sentences Over 15 Months, 2006-08 Cases**




The above plots show the bounds and confidence intervals for $ATE_{M=1}$ and $ATT_{M=1}$ for a sequence of possible values of $\rho$, represented by the x-axis. The three different plots consider three different codings of the response variable: whether the sentence length exceeds the third quartile length, the median, or the first quartile, respectively. Note that while Tables 1 through 4 show the treatment effect as a proportion of individuals who received discriminatory sentences, the y-axis in the plots is scaled by the number of cases in the sample. So y-values indicate the number of cases that would have had different outcomes (would have been on the opposite side of the sentence threshold) if the defendant was of a different race. The opaque red shape represents the bounds for $ATE_{M=1} \times n_{total}$. The light red shading covers the confidence interval for the bounds of $ATE_{M=1} \times n_{total}$. We see that the confidence interval is very near the bounds, which is what we expect with a large sample size. The dashed line represents the point estimates for $ATT_{M=1} \times n_{cases\ involving\ black\ people}$ Note that $ATE_{M=1}$ and $ATT_{M=1}$ are scaled by different numbers of cases.

Using the model with only pretreatment covariates and $\rho = 0.32$, the confidence interval for the number of cases where the decision to sentence above or below 77 months was discriminatory is between 15,400 and 30,718 people. For the decision to sentence above or below 37 months, somewhere between 21,686 and 49,277 cases are labeled as discriminatory. For the 15 month sentence length threshold, between 20,048 and 45,456 of cases are estimated to be discriminatory.

The shape and size of the bounds differs depending on where we set the sentence length threshold for the response variable. For each threshold, the bounds for higher values of $\rho$ are wider. Knox et al's plots also show the bound width increasing with $\rho$. For large values of $\rho$, there appears to be greater uncertainty about $ATE_{M=1}$ for the highest sentence threshold than for the lowest threshold. But for smaller values of $\rho$,

including Knox et al's conservative estimate of $\rho = 0.32$, the bounds for the highest threshold are the narrowest. I think that the smaller treatment effect estimate for longer sentences may be related to the correlation between measures of conviction severity and perceived race. If conviction severity is correlated with race, then the bound procedure may determine more cases involving Black people to "correctly" fall above the high sentence threshold. Again, this emphasizes how difficult it is to control for the impact of systemic racism throughout the criminal legal process. Overall, I am not sure how to explain why changing the coding of the response variable affects the magnitude and range of the estimated $ATE_{M=1}$ in the ways we observe. This is something I would be interested to research in the future.

One general observation is that the estimated $ATT_{M=1} \times n_{cases\ involving\ black\ people}$ is less than the lower bound $ATE_{M=1} \times n_{total}$. Broadly, this seems reasonable because the total sample population is substantially larger than the number of Black people in the sample. The motivation for the different scale is that $ATT_{M=1}$ quantifies the number of cases involving a Black person that would not have resulted in a sentence above the threshold if the defendant were white, while $ATE_{M=1}$ also counts the number of cases involving a white person that would have resulted in a conviction if they were Black. This labeling of shorter sentences for white people as discriminatory is why it's possible for $ATE_{M=1}$ to exceed the proportion of Black people in the sample (about 0.2735), as we observe in the estimated ATE bounds.

Specifically, the lower bound of $ATE_{M=1}$ is greater than $P(D_i = 1|M_i = 1)$ when:

$$\mathbb{E}[\hat{\Delta}] + \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1](1 - P(D_i = 0|M_i = 1)) > P(D_i = 1|M_i = 1)$$
$$\mathbb{E}[\hat{\Delta}] + \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1](P(D_i = 1|M_i = 1)) > P(D_i = 1|M_i = 1)$$
$$\mathbb{E}[\hat{\Delta}] > (1 - \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1])P(D_i = 1|M_i = 1)$$
$$\frac{\mathbb{E}[\hat{\Delta}]}{1 - \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1]} > P(D_i = 1|M_i = 1)$$
$$\frac{E[Y_i|D_i = 1, M_i = 1] - E[Y_i|D_i = 0, M_i = 1]}{1 - \rho\mathbb{E}[Y_i|D_i = 0, M_i = 1]} > P(D_i = 1|M_i = 1)$$

The form of the inequality indicates that all else equal, larger values of $\rho$ make it more likely that we'll see the lower bound of $ATE_{M=1}$ exceed the proportion of the Black people in the sample, which is consistent with the results shown in the plots.

## Results for Immigration Convictions

For $\rho = 0.32$, the lower bound on $ATE_{M=1}$ is very slightly greater than the naive ATE, but the lower bound of the confidence interval for $ATE_{M=1}$ is less than the naive ATE. This is true for all three sentence length thresholds. There isn't much of a difference between the bounds for the different sentence length thresholds, except perhaps for the low bound for the 1st quartile threshold. Overall, the bounds computed for the immigration data do not provide evidence that the unbiased treatment effect is greater than the naive estimate of the proportion of discriminatory immigration sentences.

Table 3: Average Treatment Effect among Immigration Convictions ($ATE_{M=1}$), by Sentence Quartile, Full Model Specification

| | Naive ATE | Lower Bound | Upper Bound | Lower CI | Upper CI |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| over 37 months | 0.0599 | 0.0634 | 0.1884 | 0.0481 | 0.2073 |
| over 21 months | 0.0452 | 0.0517 | 0.2196 | 0.0366 | 0.2365 |
| over 10 months | 0.0018 | 0.0117 | 0.1834 | -0.0019 | 0.1956 |

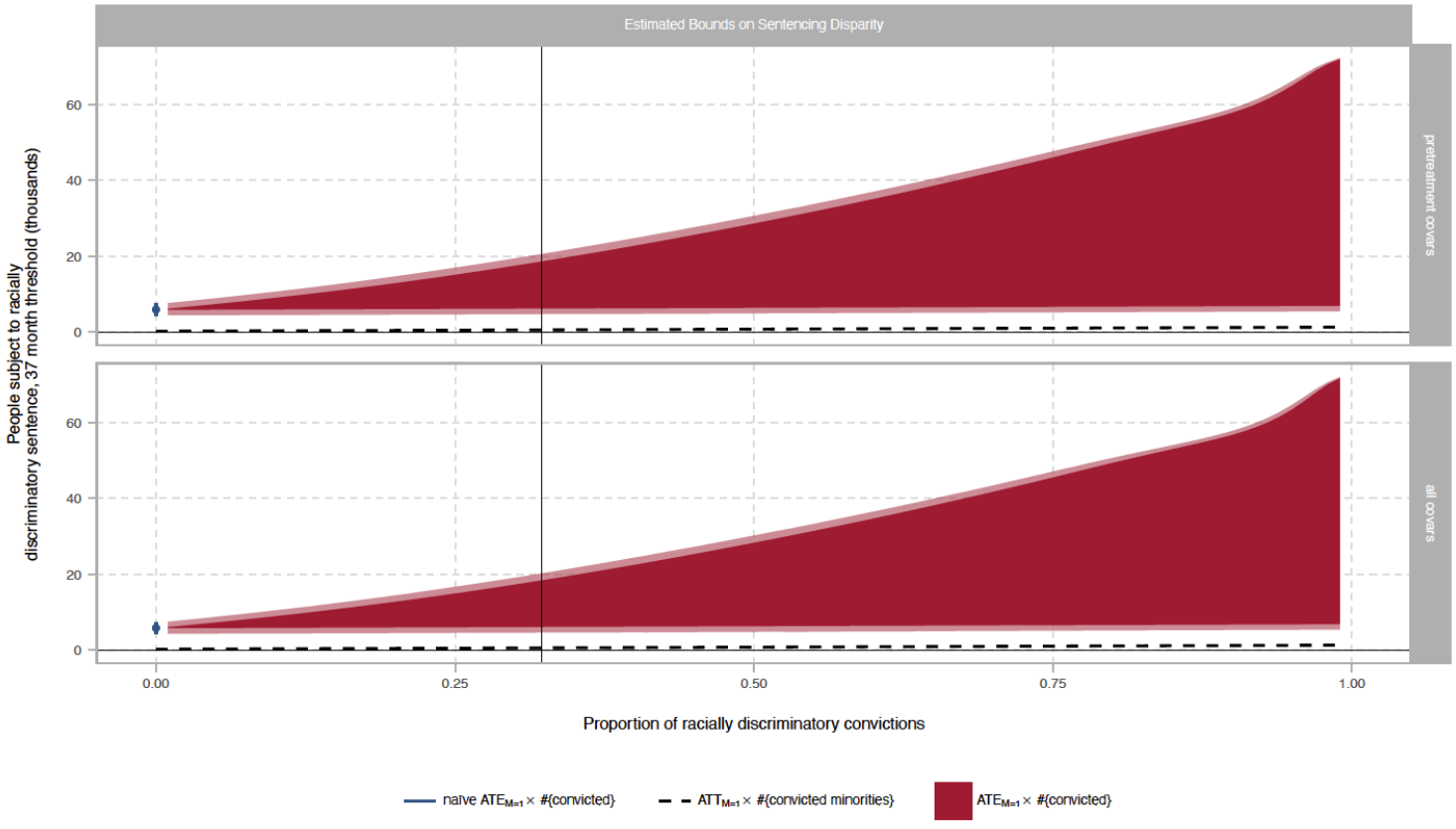
[a] rho = 0.32

Table 4: Average Treatment Effect among Immigration Convictions ($ATE_{M=1}$), by Sentence Quartile, Pretreatment Covariates Only

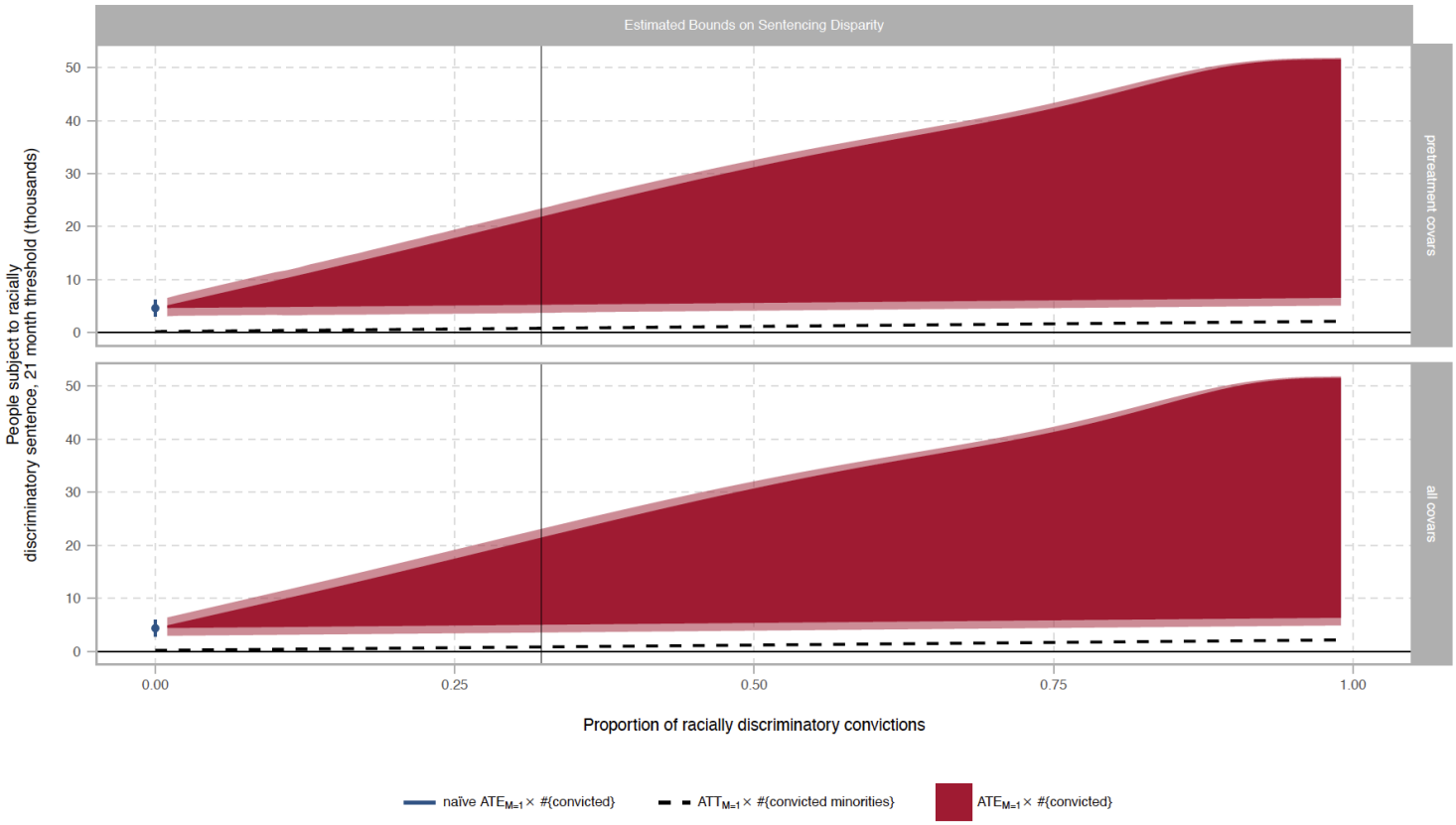
| | Naive ATE | Lower Bound | Upper Bound | Lower CI | Upper CI |
|---|---|---|---|---|---|
| over 37 months | 0.0609 | 0.0644 | 0.1908 | 0.0495 | 0.2109 |
| over 21 months | 0.0479 | 0.0544 | 0.2240 | 0.0386 | 0.2401 |
| over 10 months | 0.0050 | 0.0148 | 0.1882 | 0.0010 | 0.2004 |

[a] rho = 0.32

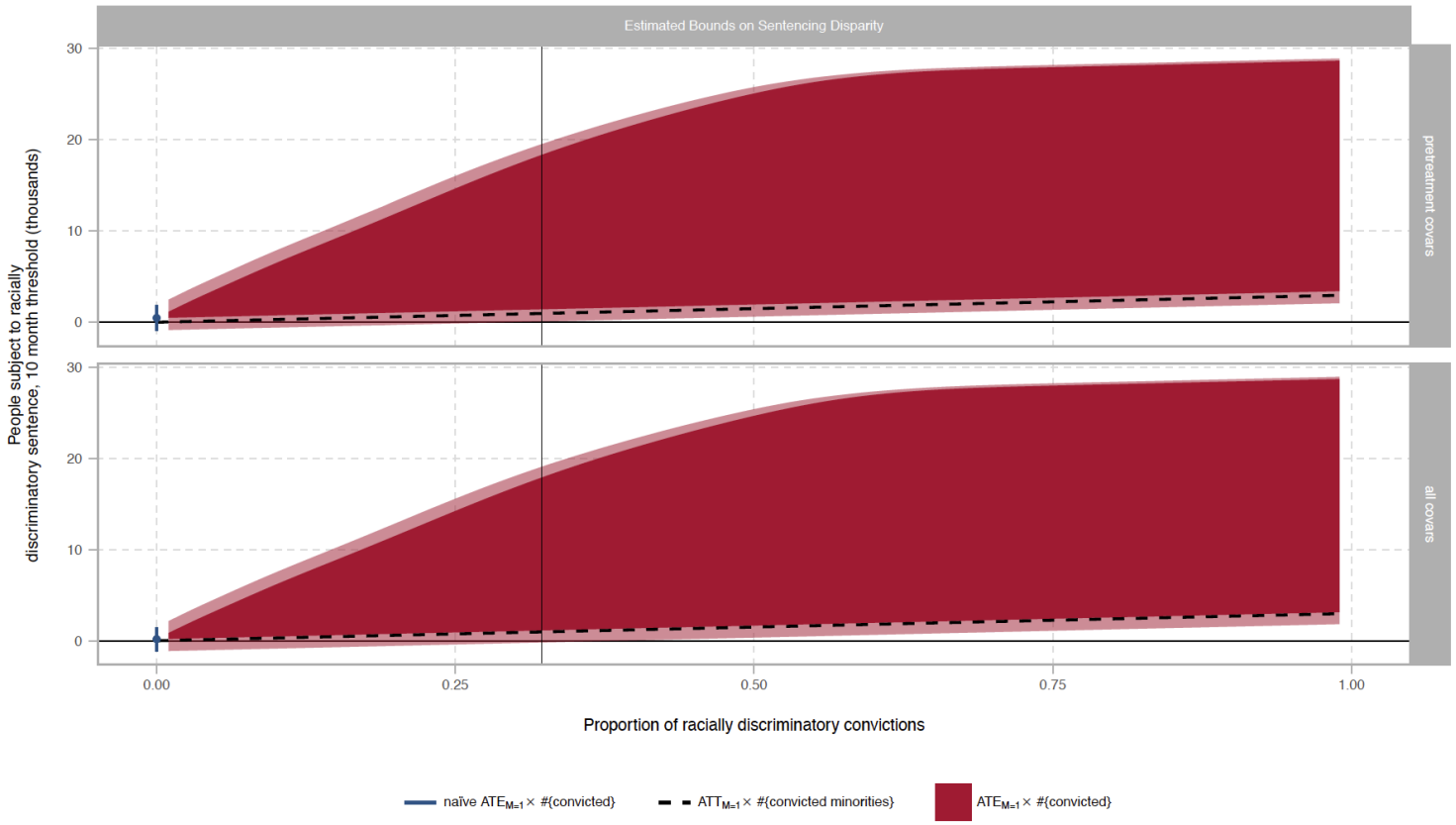
**Estimated Discriminatory Sentences Over 37 Months, Immigration Cases**



## Estimated Discriminatory Sentences Over 21 Months, Immigration Cases



**Estimated Discriminatory Sentences Over 10 Months, Immigration Cases**




For the immigration data, there is barely any difference between the bounds computed using the model with the post-treatment TRIAL and NOCOUNTS covariates, versus the bounds using only the pre-treatment covariates. This is reflective of the immigration cases being more homogenous than the full set of cases. It seems more plausible (although it is still not a trivial assumption) to assert that among immigration cases, race is randomly assigned conditional on pre-treatment covariates. Unfortunately, although Knox et al's assumptions seem more reasonable for the immigration data, the bounds on $ATE_{M=1}$ aren't very informative. Particularly for the 1st quartile sentence threshold, the bounds and confidence intervals are very wide. For the first quartile sentence threshold, the $ATE_{M=1}$ bounds dip below 0, which is clearly incorrect under the assumption that there is not an anti-white bias. As stated above, the large value of $P(D_i = 0|M_i = 1)$ (proportion of immigration cases involving white people) contributes to this particularly low bound. Comparing Tables 2 and 4, the bounds for the immigration $ATE_{M=1}$ are wider than the bounds for the 2006-08 $ATE_{M=1}$, even when modeling the 2006-08 data with only pre-treatment covariates.

For $\rho = 0.32$, whether or not we include post-treatment covariates, there is a significant disparity between the probability of similar Black and white cases receiving immigration sentences above the median (21 months) and above the third quartile (37 months). According to the confidence interval for the bounds using the pre-treatment model, between 4,800 and 20,451 decisions about whether to give an immigration sentence longer than 37 months would have been different if the defendant was of another race. For decisions about immigration sentences longer than 21 months, between 3,743 and 23,282 decisions out of 96,969 are estimated to be discriminatory.

# Conclusion

In summary, for a dataset of 165,416 federal sentences from 2006 to 2008, I found that the naive ATE significantly underestimates $ATE_{M=1}$. However, I observed that the assumption of treatment ignorability conditional on mediator strata and covariates (Knox et al Assumption 4) may not hold for this dataset, potentially introducing bias into the estimation of $ATE_{M=1}$. In an attempt to meet the conditions of Assumption 4, I computed $ATE_{M=1}$ bounds for immigration cases from 2000 to 2008. This set of cases is more homogenous than the variety of convictions from 2006-2008, making treatment ignorability more plausible. The resulting bounds were very uncertain due at least in part to the fact that only about 4% of immigration cases involved Black defendants. Notwithstanding the imprecision and the violated assumptions, the analysis reaffirmed that federal courts discriminate against Black people, sentencing some proportion of cases to more severe punishments than they would receive if the defendant was white. In closing, I want to acknowledge that anti-Blackness is build into the American criminal legal system. Although it is a worthwhile effort to improve statistical methods for describing disparities, it doesn't take a perfect statistical analysis to know that systemic racism exists and needs to be addressed.

## Questions for Future Research

As mentioned above, I used the value of $\rho$ that Knox et al identified as a conservative estimate of the proportion of discriminatory police stops in New York. I would like to identify a plausible estimate of $\rho$ in the sentencing context based on studies of federal convictions. More importantly, I would also like to improve my model of sentencing outcomes and my understanding of the resulting $ATE_{M=1}$ bounds. I would like to more carefully select a set of pre-treatment covariates thought to be predictive of $Y_i$. I would also like to gain some intuition for where the uncertainty of the bounds comes from and how the shape of the bounds changes depending on the sentence length threshold. I think that part of this understanding would come from closer inspection of the bounds formula, but it would also be helpful to simulate the bounds computation for different model specifications and different sentence thresholds.

# Appendix

## Data Cleaning Process

The data from the U.S. Sentencing Commission is available to download in ASCII, SAS, or SPSS formats. Note that each dataset has two parts: Part 1 - Main Data, and Part 2 - Supplementary Data. I only used Part 1 data. I chose to use the data formatted for SPSS because I found R packages for this conversion. One difficulty was that the SPSS files for 2000-2005 are different from those for 2006-2008. For the earlier years, the data comes in a .txt file with a .sps setup file. For the later years, the data and metadata comes in one .sav file. For 2000-2005, I used "asciiSetupReader" to parse the .sps setup file for the SPSS-formatted data and then read the .txt data file according to that format. For 2006 onward, I used the "foreign" package, which has an easy-to-use read.spss() function for files with the .sav extension. To verify that the .sav files were loaded properly, I imported the data into Python and checked that it matched the corresponding R dataframe. I used the read_spss function from the pandas module. I obtained the same number of observations, the same number of NAs, and the same summary statistics for the 2006 data in R and Python. I did not find a Python package for importing the old SPSS data with .sps and .txt files. To check my work in loading these files into R, I compared the number of observations and the mean value of each variable to the values in Hagen's paper. There are a few small discrepancies between my and Hagen's summary statistics, which I believe come from data cleaning choices including whether to code certain observations as 0 or NA. I also ended up with 65 more observations than Hagen, but in a dataset with over 500,000 observations, I hope that this is an inconsequential difference.

The Sentencing Commission data is provided by year, so I had to combine several files to create a multi-year dataset. Variable names varied slightly from year-to-year, so I checked each year's codebook to figure out which variable names represented the same quantity. I then renamed the variables to have consistent names. I created indicator variables for each year from 2000 to 2008 to identify which dataset each observation came from. Each dataset includes cases sentenced from October of the previous year through September of the indicator variable year. Variables also had different formats from year to year. The different packages that I used to import the data and the different ways that values were recorded by the Sentencing Commission meant that observations for a single variable might be numeric, factor, or character data. I chose an appropriate data type for each variable and converted all observations to that type, making sure not to create NAs in the process. After this, I combined the data from all years.

I re-coded several variables to make for a more easily interpretable model. The Sentencing Commission provides an education variable with many categories for an individual's highest level of educational attainment. Following Hagen's paper, I created indicators of whether someone obtained a high school GED, whether they received some post-secondary education, and whether they obtained a post-secondary degree. For the sentence length variable, I re-coded sentences of "990 months or more" as 990. I re-coded life sentences as 945.6 months or 78.8 years, which was the average life expectancy for the total U.S. population in 2019.[8]. Hagen found differences in immigration case outcomes depending on whether cases were adjudicated in the Southwest Border region of the U.S., so I used the judicial districts that Hagen identified to create a Southwest Border indicator. I also followed Hagen in creating six indicators of whether an individual's criminal history was designated as Category 1 through 6, 6 being the most serious. I also recoded text

responses so that responses that had the same meaning had the same value (e.g. I made sure that abbreviations were consistent). Finally, I re-coded the primary offense variable to provide concise descriptions of the conviction. Many of the convictions were originally described by numeric codes.

I wrote the cleaned dataframe to a compressed .csv.gz file. I verified that when I loaded and unzipped the file in R, the data had the expected format and summary statistics. For subsequent analysis, I began with this clean data file.

## Github Repository

This repository contains data used, a data cleaning script, scripts for computing bounds and creating plots (modified from Knox et al's replication code), and results.

https://github.com/absteck/senior-project (https://github.com/absteck/senior-project)

---


1. Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." American Political Science Review. Volume 114 , Issue 3 , August 2020 , pp. 619 - 637. https://doi.org/10.1017/S0003055420000039 (https://doi.org/10.1017/S0003055420000039)↩
2. "Monitoring of Federal Criminal Sentences Series." Inter-university Consortium for Political and Social Research. University of Michigan. Accessed December 15, 2021. https://www.icpsr.umich.edu/web/ICPSR/series/83 (https://www.icpsr.umich.edu/web/ICPSR/series/83).↩
3. Hagen, Courtney 2011, 'Bias in the federal judicial system: do sentencing disparities exist in the Southwest Border region of the United States?', MPP thesis, Georgetown University, Washington, D.C.↩
4. "About the U.S. Courts of Appeals." United States Courts. Administrative Office of the US Courts. Accessed December 15, 2021. https://www.uscourts.gov/about-federal-courts/court-role-and-structure/about-us-courts-appeals (https://www.uscourts.gov/about-federal-courts/court-role-and-structure/about-us-courts-appeals).↩
5. "2018 Chapter Five - Determining the Sentence." United States Sentencing Commission, April 5, 2019. https://www.ussc.gov/guidelines/2018-guidelines-manual/2018-chapter-5 (https://www.ussc.gov/guidelines/2018-guidelines-manual/2018-chapter-5).↩
6. Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." American Journal of Political Science 62, no. 3 (2018): 760–75. https://doi.org/10.1111/ajps.12357 (https://doi.org/10.1111/ajps.12357).↩
7. "Immigration-Related Criminal Offenses." Congressional Research Service, January 21, 2020. https://sgp.fas.org/crs/homesec/IF11410.pdf (https://sgp.fas.org/crs/homesec/IF11410.pdf).↩
8. Elizabeth, Arias, Tejada-Vera Betzaida, and Ahmad Farida. "Vital Statistics Rapid Release, Report No. 010." cdc.gov. CDC, February 2021. https://www.cdc.gov/nchs/data/vsrr/VSRR10-508.pdf (https://www.cdc.gov/nchs/data/vsrr/VSRR10-508.pdf).↩