Senior Thesis in Mathematics

Over-Policing and Fairness in Machine Learning

Author:

Justin Weltz

Advisor:

Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment

of the Degree of Bachelor of Arts

May 2, 2019

Contents

Chapter 1

Introduction

Glenn Rodriguez took college classes, trained service dogs and volunteered for a youth program during his 26-year sentence. *Chance Magazine* reported that he was denied parole earlier this year. His COMPAS score, generated by a mysterious proprietary algorithm to estimate his likelihood of recommitting a crime, was cited as evidence against his case for release [7]. How could he prove that an inhuman, computer-generated algorithm had treated him unfairly?

In order for Glenn to contest the court's ruling, a comprehensive dialogue needs to be built around the systemic problems that algorithms develop when optimized for predictive accuracy. ProPublica, an investigative journalism newsroom, recently found that for every white inmate labeled high

risk by COMPAS who didn't recommit a crime upon release, there are two black inmates who are similarly mischaracterized by the algorithm [8]. While this finding screams racial discrimination, the field of machine learning currently lacks the larger philosophical and statistical framework to formalize and resolve algorithmic bias. As machine learning techniques become more pervasive, there is an increasing imperative to evaluate their effect on the populations they model. This thesis is part of an effort in the field of fair machine learning to develop a language with which victims like Glenn can change the ways algorithms affect their lives.

What does it mean for an algorithm to discriminate? It it important to acknowledge that a non-discriminatory model can't exist. Inherently, an algorithm that predicts individual behavior makes decisions based on a series of relevant attributes - it discriminates and generalizes information in order to capture patterns. However, not all attributes should be treated similarly. Over time, the U.S. legal system has come to landmark decisions (many during the civil rights movement) outlawing discrimination based on certain characteristics.

Definition 1.1 Sensitive Attributes: *These are individual characteristics protected by U.S. Law. Discrimination based on these attributes (age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, sexual orientation) is illegal in certain contexts (employment, sentencing, voting etc.).*

Legislation such as the Civil Rights Act, the Fair Housing Act, and the Equal Credit Opportunity Act make discrimination based on sensitive attributes such as color, religion, and sex illegal [2]. My thesis will focus on analyzing algorithmic racial bias, although the question of how to balance a model with respect to multiple sensitive attributes is an interesting line of inquiry with unsatisfying answers. However, diagnosing disparate impact across even one feature is still a complicated, multi-dimensional problem. One might suggest that simply omitting the race variable from a training set would prevent the model from discriminating. In a sense, they would be right. Disparate treatment, a possible definition of discrimination, is defined as explicitly deciding an outcome based on a sensitive attribute. A model that predicts without the race variable would not exhibit disparate treatment. However, there are many ways of defining discrimination.

In Griggs v. Duke Power Company, the US Supreme court ruled that a business's hiring process was problematic even though it did not depend explicitly on a sensitive attribute. Duke Power Co. was banned from using intelligence scores and high school diplomas to determine professional fitness because these measures were correlated with race and consequently caused large differences in hiring outcomes for black and white individuals [9]. This monumental decision began the struggle between concepts of disparate treatment and disparate outcome. The concept of evaluating a model based on "unintended discrimination" (disparate impact) is controversial and often

directly contradicts the implications of disparate treatment. However, the current, heated debate surrounding the concept of affirmative action and white privilege is a testament to its relevance. My thesis will evaluate the effect of over-policing on crime-related predictive algorithms in the context of disparate impact.

Disparate impact is often defined as the violation of outcome equality. The most stringent definition of outcome equality is Independence:

Definition 1.2 (Independence) *If we define R as a binary classifier and A as sensitive attribute with two levels a and b, a model is independent (does not exhibit disparate impact) if it has the following behavior $P(R = 1|A = a) = P(R = 1|A = b)$.*

However, there are also a series of more nuanced definitions of outcome equality. The US Equal Employment Opportunity Commission uses the 80% rule to define outcome equality in hiring decisions:

Definition 1.3 (80% rule) *If we define R as a binary classifier and A as sensitive attribute with two levels a and b, a model follows the* **80%** *rule if it exhibits the following behavior $\frac{P(R=1|A=a)}{P(R=1|A=b)} \geq \epsilon = 0.8$ (assume that $P(R = 1|A = b) > P(R = 1|A = a)$ without loss of generality). Therefore, a model exhibits disparate impact when $\frac{P(R=1|A=a)}{P(R=1|A=b)} < \epsilon = 0.8$*

Both independence and the 80% rule are vulnerable to arguments appealing

to business necessity. A company can claim that avoiding disparate impact in hiring decisions would prevent the acquisition of certain skillsets and hurt business. Variants on the definitions stated above are susceptible to this criticism because they do not account for the accuracy of model decisions with respect to the outcome an institution (or in this scenario, a company) is seeking to maximize. For example, a business hiring process attempts to identify applicants that are most likely to succeed within the company environment. This notion of professional fitness can be evaluated in a variety of ways. In the sales department of a large clothing chain, employee success may be accessed by the revenue of individual sales and the number of new clients acquired. This business may argue that hiring equal numbers of black and white applicants would be impossible for the success of their company because a higher number of white applicants have taken the college marketing classes necessary to succeed. If they implement this hiring policy, they will be violating notions of outcome equality such as Independence and the 80% rule. However, let's take a closer look at the nature of this business' hiring decisions. In this case, the business' hiring model predicts a higher likelihood of success for an individual who has taken more marketing classes. If the company collects information on the sales numbers of black and white applicants and demonstrates that the average difference between the predicted sales (based on the number of marketing classes and possibly modeled using a linear regression) and the actual sales of these groups are the same, they have demonstrated that the **accuracy** and **error** of the model across

these two groups is the same. Therefore, in a sense they have demonstrated that their algorithm treats individuals equally across race without necessarily satisfying a notion of outcome equality. Separation and Sufficiency are two definitions of algorithmic fairness that focus on equal treatment instead of equal outcomes.

Let us define R as a binary classifier, A as the sensitive attribute and Y as the target variable. A target variable is the outcome that the binary classifier is attempting to predict (it is unobservable at the time of the prediction). For example, A is race, R is a high COMPAS score and Y is whether or not the defendant recidivates in the COMPAS case mentioned above.

It is important to note that Separation and Sufficiency can only be verified when the target variable becomes observable (for COMPAS this was two years after the defendant appeared in court), while Independence or the 80% Rule can be evaluated immediately. This is not necessarily a disadvantage of the following definitions, but it is an important distinction to bear in mind.

Definition 1.4 Separation: $P(R = 1|Y = 1, A = a) = P(R = 1|Y = 1, A = b)$ *and* $P(R = 1|Y = 0, A = a) = P(R = 1|Y = 0, A = b)$. *The probability of an outcome's value is independent of the sensitive attribute conditional on being classified positively (or negatively) by the model.*

Definition 1.5 Sufficiency: $P(Y = 1|R = 1, A = a) = P(Y = 1|R = 1, A = b)$ *and* $P(Y = 1|R = 0, A = a) = P(Y = 1|R = 0, A = b)$. *The proba-*

bility of being classified positively (or negatively) by the model is independent of the sensitive attribute conditional on the value of the outcome variable.

Sufficieny implies that there is equal predictive accuracy across the sensitive attribute. In the context of racial bias in the COMPAS algorithm, this means that a high risk black individual is just as likely to have committed a crime as a high risk white individual. Separation is the concept of fairness that we will focus on in this paper. It is consistent with the racial bias found by ProPublica in COMPAS and implies that, in order to achieve equality of outcomes, a black individual who recommits a crime should be just as likely to be labeled high risk as a white individual who recommits a crime. More generally, Separation means that false positive and false negative rates are equal across the sensitive attribute.

For reference, if "TN" is the number of observations correctly classified as negative by the model and "FP" is the number of of observations incorrectly classified as positive by the model then:

False Positive Rate,

$$FPR = \frac{FP}{TN + FP} \tag{1.1}$$

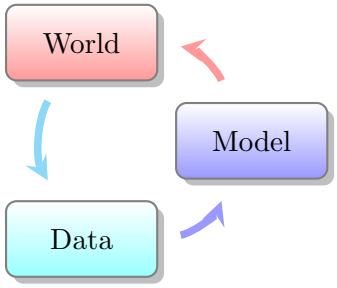
and, if "TP" is the number of observations correctly classified as positive by the model and "FN" is the number of of observations incorrectly classified as negative by the model then:

False Negative Rate,

$$FNR = \frac{FN}{TP + FN} \tag{1.2}$$

Now that we have defined discrimination as disparate impact and decided to measure it using Separation, the next question becomes: what are the mechanisms driving non-Separation disparate impact in algorithms such as COMPAS? If institutions aren't intentionally discriminating, why does algorithmic bias occur across sensitive attributes? In order to pinpoint the root of differential outcomes, it is helpful to imagine models as part of a larger environment.

Figure 1.1: Algorithm Cycle



Each arrow in the diagram above represents different steps in the modeling process. And, at every juncture in the flow of information from one stage to another, an algorithm is susceptible to subjective choices that can result in discrimination and disparate outcomes. We are particularly interested in the blue arrow between the "World" and the "Data" because it is often overlooked. At first glance, establishing data acquisition methodologies seems

like a robotic, mundane task. However, it is clearer on closer inspection that the unique compromise between funding, experimental methodology, privacy laws, and preconceptions of the world drives data collection and almost completely determines the patterns that an algorithm will identify in a training set. In addition, the realities and limitations of data collection often cause data to be far from identically and independently sampled. Without a representative sampling mechanism, many of the assumptions that statistical inference and modeling rely upon fall apart.

1.1 The Thesis

Our hypothesis is that the common practice of over-policing black neighborhoods results in skewed samples that affect models attempting to predict individual crime. This conjecture relates to the COMPAS algorithm, but its scope is not limited to criminal justice. The results can be generalized so that they inform and correct modeling techniques that depend on over-sampled populations in a variety of contexts. Figure 1.1 was included as both an instructive tool and a cautionary tale. The loop will only become stronger and increasingly inaccessible to the public as models become more prevalent and more complex. It is truly scary that a model picking up patterns from biased data can have a substantial impact on the world and consequently create and influence behavior that it then observes - thereby creating a feedback loop

that molds the world in the image of dangerous measurement errors.

In order to evaluate our hypothesis, chapter two will begin by establishing the model being examined, logistic regression. This is a very popular binary classifier and will be the representative algorithm used to evaluate theories of predictive accuracy and disparate impact. Then, it will explain how we simulated the effect of over-policing on two representative populations as well as the basic tendencies of the logistic regressions that modeled this representative dataset. Chapter three will focus on estimating the effect of over-policing on disparate impact contingent on a theory of criminal tendencies. The conclusion will summarize and connect our findings.

Chapter 2

Investigating the Effect of Over-Policing on False Positive Rates

2.1 Logistic Regression

In many ways, my thesis revolves around the behavior of logistic regression. In order to better understand the mechanics of this model, I will use this section to dive into the mathematics that provide the foundation for this prediction algorithm.

A logistic regression, which is a type of generalized linear model, differs from ordinary least squares regression in a small, but fundamental way. Instead of modeling a continuous outcome variable, a logistic regression models a binary outcome variable using a **link function**. The logistic regression link function is called the logit and represents the log odds ratio:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}$$

where i is an index for the subjects or observations and j is an index for the explanatory variables. Solving for π_i gives,

$$\pi_i = \frac{e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}}}$$

The logit is a sigmoid function, dependent on the predictor variables and the beta coefficients and constrained between 0 and 1. There are a couple of important aspects of logistic regression to note at this point (they will come in handy later).

- The maximum likelihood estimates for the beta coefficients do not exist in closed form. Numerical approximation must be used.
- The predicted values from a logistic regression model can be considered

to be probabilities (π_i), where Y_i (the outcome variable) $\sim bern(\pi_i)$. π_i (a number between 0 and 1) is a function of specific beta coefficients and predictor variables for each observation. Consequently, in order to predict on another data set, one needs to establish a **cutoff** point for classifying an observation as 1 or 0 depending on the π_i.

2.2 The Simulation

Based on a hypothetical scenario, our simulation generates a population from which we can sample. Imagine that individuals in a population distinguished by a binary race variable are prone to entering crime. Each individual in the population has a likelihood of entering crime given by a function on a few predictor variables. These variables include age and gender, whose distributions in the population are determined in advance, and past criminal offenses, which is initially set to 0 for every individual. Every year, a crime likelihood is calculated for every individual that is generally higher for men, the youth, and past criminals (as demonstrated in the plot below).

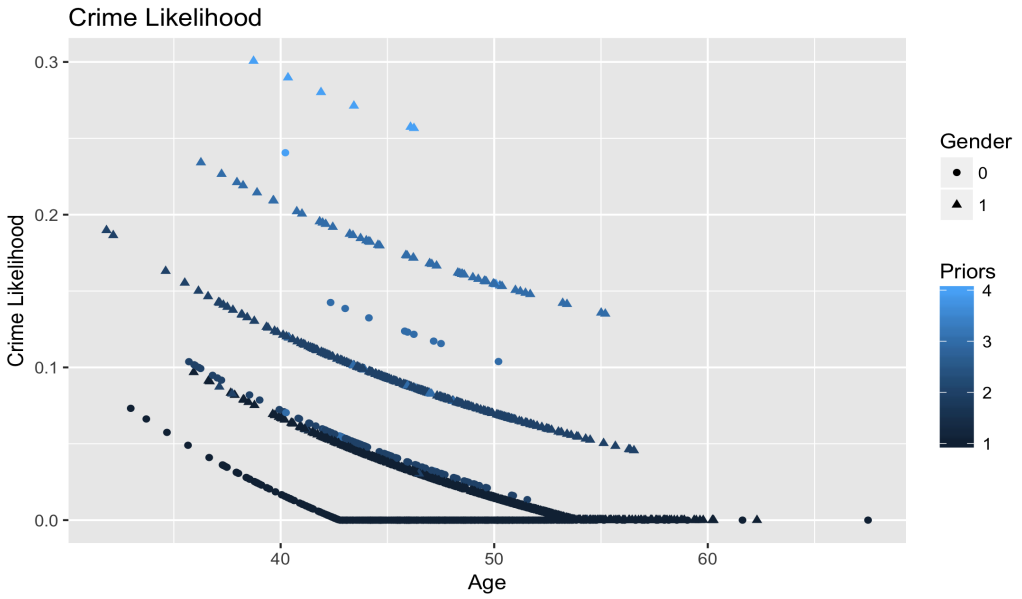


Figure 2.1: A graph of crime likelihood as a function of age, gender, and prior offenses after 10 years

The crime likelihood is defined as the probability of an individual entering crime conditional on their attributes (gender, age, and past priors). In our simulation, we created the crime likelihood function to reflect the general effects of these attributes as seen in the COMPAS data (younger men with more priors are more likely to go into crime), but we cannot claim that the relative effect sizes of the attributes are accurate for any actual population. Hopefully, further applications of this research will attempt to model a realistic crime likelihood function for a given population using social science theory.

Once an individual has entered crime, each year they either stay in crime,

leave crime with 0.2 probability (a guess at the rate of this transition), or enter jail. The probability of an individual in crime being jailed is different across the two observed races in our simulation. The table below demonstrates these differences:

Table 2.1: Probabilities of Entering Jail by Race and Prior Offenses

	Black	White
No Prior Offenses	over-policing	0.05
Prior Offenses	over-policing + 0.03	0.08

Once again, the relative differences between the probability of entering jail for whites and for blacks is not meant to model a specific over-policed population. The simulation is simply attempting to investigate how differences in over-policing probabilities affects the population dynamics. In our simulated model, once individuals are in jail, they serve a three year sentence and then are released back into population where they undergo the same modeling process with a prior offense recorded. See Algorithm 1 for more details and pseudo-code for the simulation.

2.3 Running the Simulation

After running the simulation, we have a data set that we can use to train a logistic model. Our binary response variable is whether or not an individual

```
Initialize the population;
Ages are normally distributed around a mean of 30 with a standard deviation of 5;
Sex is equally likely to be male and female;
Race is equally likely to be black and white;
Priors are set to zero for every individual;
Each individual has a 5% chance of starting in crime;
Set the over-policing parameter = θ;
for 10 iterations (10 years) do
    for every individual in the population do
        calculate crime likelihood (γ_i) as a function of the individuals attributes
          (sex, priors and age but not race). It should range from 0 (won't enter
          crime this year) to 1 (will enter crime this year);
        if not in crime or jail then
            Use the γ_i as the probability that the individual will enter crime;
        end
        if in crime but not in jail then
            There is a 20% chance that you leave crime. But, if an individual
              stays in crime;
            if black then
                if no prior offenses then
                    Use θ + 0.03 parameter as the probability that the individual
                      will enter jail;
                end
                if prior offenses then
                    Use θ as the probability that the individual will enter jail;
                end
            end
            if white then
                if no prior offenses then
                    5% chance of going to jail;
                end
                if prior offenses then
                    8% chance of going to jail;
                end
            end
        end
        if in jail then
            An individual is in jail for three years and then reenters the
              population;
        end
    end
    Iterate age;
end
Then return to the first for loop with a population of those who have committed a
  crime in the first 10 years;
If they commit a crime in the second 10 years than they are said to have
  recidivated;
```

Algorithm 1: Running the Simulation

who has committed a crime in the first ten years recommitted a crime in the subsequent ten years, and our predictor variables are age and number of previous offenses at the beginning of the second ten years, gender, and race.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.0445	0.5669	5.37	0.0000
Gender	0.8566	0.1347	6.36	0.0000
Age	-0.1471	0.0143	-10.29	0.0000
White	-0.9045	0.1554	-5.82	0.0000
Priors	0.8205	0.2062	3.98	0.0001

Table 2.2: Logistic Regression for Simulation with Over-Policing Parameter Equal to 0.2

Table 2.2 represents the information contained in the logistic regressions produced by the simulation. Age, gender, race, and priors are all significant variables when predicting ten year recidivism. However, based on the mechanics of our simulation, we knew they would be! From the beginning, we calculated the crime likelihood as a function of age, gender and priors. This means that the likelihood of an individual going into crime and then jail are based on these attributes. The different over-policing parameters acting on the two populations makes the race variable significant as well. Increasing the over-policing parameter increases the rate at which black individuals go to jail and consequently the odds of a black individual having gone to jail during the simulation relative to a white individual. Since the odds of a black individual going to jail relative to a white individual is $e^{B_{White}}$ (See Table 2.2), we know that as we increase the over-policing parameter, the

race variable (B_{White}) will increase in magnitude and the model will predict more blacks to recidivate. However, we don't inherently know how the over-policing parameter will influence the error of the model (the false positive and false negative rates discussed earlier). How will increasing the over-policing parameter change the false positive rate, the false negative rate, and the separation (the fairness) of the model?

2.4 Distinguishing Between False Positive Rates by Over-policing

In order to get a sense for the relationship between the over-policing parameter and the false positive rate we created a meta-simulation in which we ran the simulation in Algorithm 1 for a range of over-policing values between 0.1 and 0.3 over 100 iterations. We then trained logistic regressions on each of the corresponding data sets, predicted whether individuals recidivated during the 10 years of the simulation based on a series of cutoff points (ranging

from 0.1 to 0.9) and calculated the false positive rates (See Algorithm 2).

for *each iteration* **do**

- **for** *each value of the over-policing parameter* **do**
  - I. Run the simulation once.
  - II. Save a data set containing the attributes of everyone who went to jail in the first 10 years and record whether they went to jail in the second 10 years in a recidivism variable.
  - III. Run a logistic regression with the attributes as the independent variables and the recidivism variable as the response variable.
  - IV. Run the simulation a second time and save this data.
  - **for** *each cutoff value* **do**
    - I. Predict whether or not individuals in the second simulation have recidivated using the cutoff values (recall that a cutoff is necessary to predict a binary output since a logistic regression outputs a probability value between 0 and 1).
    - II. Calculate the false positive rate and false negative rate based on these predictions.
  - **end**
- **end**

end

Algorithm 2: Meta-Simulation

The average false positive rates (averaged over the iterations) are plotted in the graph below:

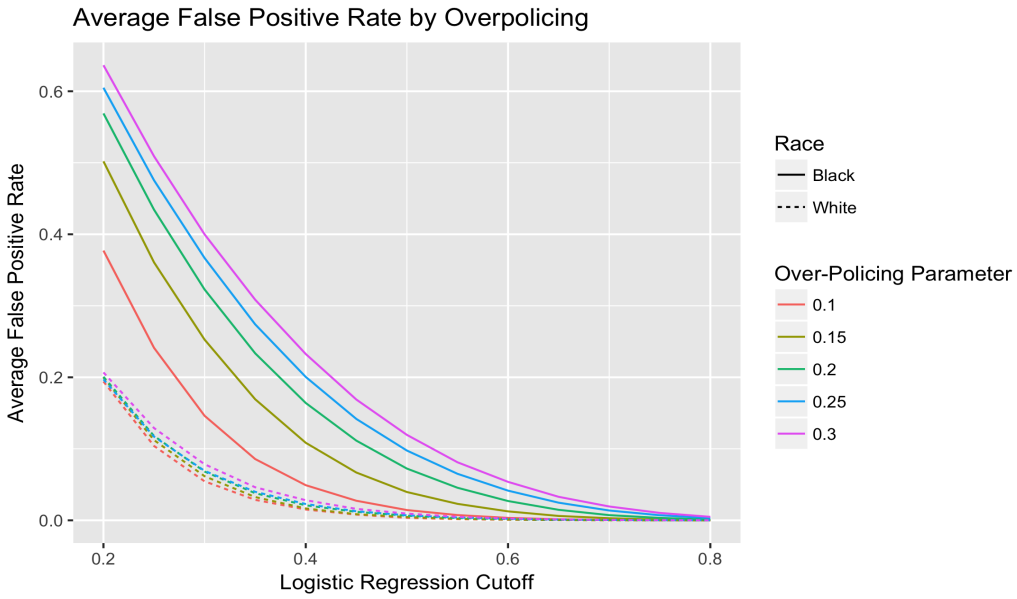


Figure 2.2: False Positive Rates

It seems from the plots that the false positive rate differs across different levels of over-policing and race, but how can we be sure? How do we know these differences aren't the result of the meta-simulation's natural variation? In order to get a sense for whether the false positive rates are significantly different from each other at a given cutoff value, we will take advantage of the fact that the false positive rates are averaged over multiple iterations of the meta-simulation for each over-policing values. Figure 2.3 replots figure 2.2 with standard error bars added from calculations in equations 2.1 and 2.2.

If we label the false positive rate for a given iteration, b, and cutoff value, c,

as FPR_{c_b} and the total number of iterations as n,

$$\overline{FPR_c} = \frac{\sum_{i=b}^{n} FPR_{c_b}}{n} \tag{2.1}$$

$$s_{FPR_c} = \sqrt{\frac{1}{n-1}\sum_{i=b}^{n}(FPR_{c_b} - \overline{FPR_c})^2} \tag{2.2}$$

The 95% confidence interval for each true average false positive rate at a given cutoff value is consequently:

$$\overline{FPR_c} \pm t_{n-1,.975} * \frac{s_{FPR_c}}{\sqrt{n}} \tag{2.3}$$

After adding error bars to the FPR curves, the difference between the rates continues to be statistically significant:

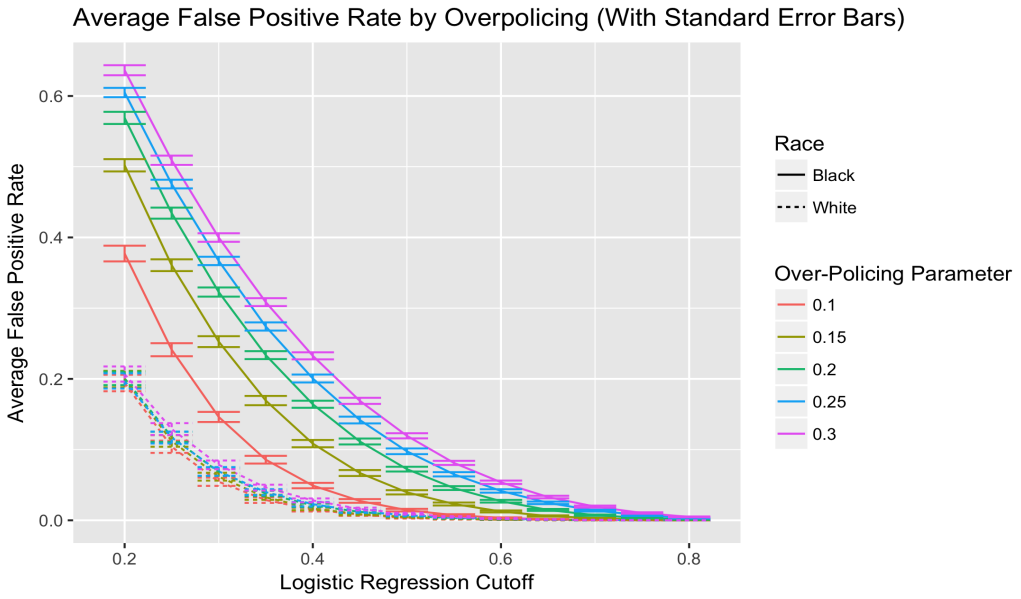


Figure 2.3: False Positive Rates with Standard Error Bars

The chart demonstrates that there is a statistically significant difference between black false positive rates across over-policing levels for most cutoff values. However, it is important to acknowledge that this statistical significance is partly artificial. Since we determine the distance between the over-policing parameters tested, we have some control over whether the false positive rates will be significantly different from each other. Therefore, it is best to treat these statistical tests as more evidence of the general trend indicating that as the prevalence of convicted individuals in the black population increases in response to the increasing over-policing parameter, the false positive rate of the logistic regression for blacks will increase. Lastly, note that the over-policing parameter does not directly affect the white pop-

ulation, so it is encouraging to see that the white false positive rate does not exhibit a similar trend. Let's dive a little deeper into why the relationship between the over-policing parameter and the false positive rate might exist:

When the predictors contain no information relating to the response variable, the logistic regression predicts based on prevalence of the response variable outcomes. This can be seen by manipulating the log-odds formula depicted below:

$$\pi_i = \frac{e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}}}$$

if $B_j = 0 \; \forall j \in \{1, \cdots n\}$, this expression equals:

$$\pi_i = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Therefore, the likelihood for the binary outcome variable y_i with no predictor variables becomes:

$$f(\underline{y}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$$

$$f(\underline{y}) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{y_i} \left(1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{1 - y_i}$$

$$f(\underline{y}) = \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{\sum_{i=1}^{n} y_i} \left(1 - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{n-\sum_{i=1}^{n} y_i}$$

$$ln(f(\underline{y})) = \sum_{i=1}^{n} \beta_0 * y_i - n * ln(1+e^{\beta_0})$$

Differentiating the likelihood with respect to β_0 and setting it equal to zero produces the maximum likelihood estimator of π_i.

$$\frac{dln(f(\underline{y}))}{b\beta_0} = \sum_{i=1}^{n} y_i - n * \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \hat{\pi}_i \tag{2.4}$$

Therefore, it seems that the predictions of a logistic regression are connected to the prevalence of the response variable outcomes and the information contained in the variables. This relationship between predictions, prevalence and information is highly related to the effect of over-policing on false positive rates. At its core, over-policing is affecting the prevalence of the response variable (recidivism) in the black population. By analyzing the connection between prevalence, false positive rates, and variable information at a basic level, we may be able to get a better handle on how over-policing affects false positive rates. In order to examine this relationship, we designed Algorithm 3 to mimic a simple scenario where there is one continuous predictor variable and a binary outcome. During the simulation we will vary the amount of information contained in the predictor variable as well as the prevalence of the

binary outcome, apply logistic regressions to these data sets, and calculate the false positive rates of these models.

```
n is the number of observations;
for each iteration do
    for a range of prevalence (ζ) values between 0 and 1 do
        for a range of entropy (κ) values between 0.55 and 1 do
            Generate the response variable ∼ bin(n, ζ) ;
            Generate the predictor variable values by conditioning on the value of
             the response variable ;
            if the response variable is 1 then
                create the predictor variable ∼ unif(1 − κ, 1)
            end
            if the response variable is 0 then
                create the predictor variable ∼ unif(0, κ)
            end
            Model the data using a logistic regression, predict using a cutoff value
             of 0.5 and calculate the false positive rate ;
        end
    end
end
```

Algorithm 3: Variable Information and Prevalence Simulation

The entropy parameter (κ) in our simulation may be a little confusing on first glance. Basically, it determines how much information the predictor variable contains about the response. If $\kappa = 1$, then the predictor variable $\sim unif(0, 1)$ regardless of the response, and consequently the predictor contains no information about the response. However, as the value of κ decreases,

the uniform distribution that the predictor variable is drawn from is more closely concentrated around the response variable - which means it contains more information about the true value of the response variable. If $\kappa < 0.5$, the uniform distributions for the predictor variables, given the two response variables, have no overlap and the logistic regression can perfectly classify the data.

The results of the simulation are depicted in the chart below:

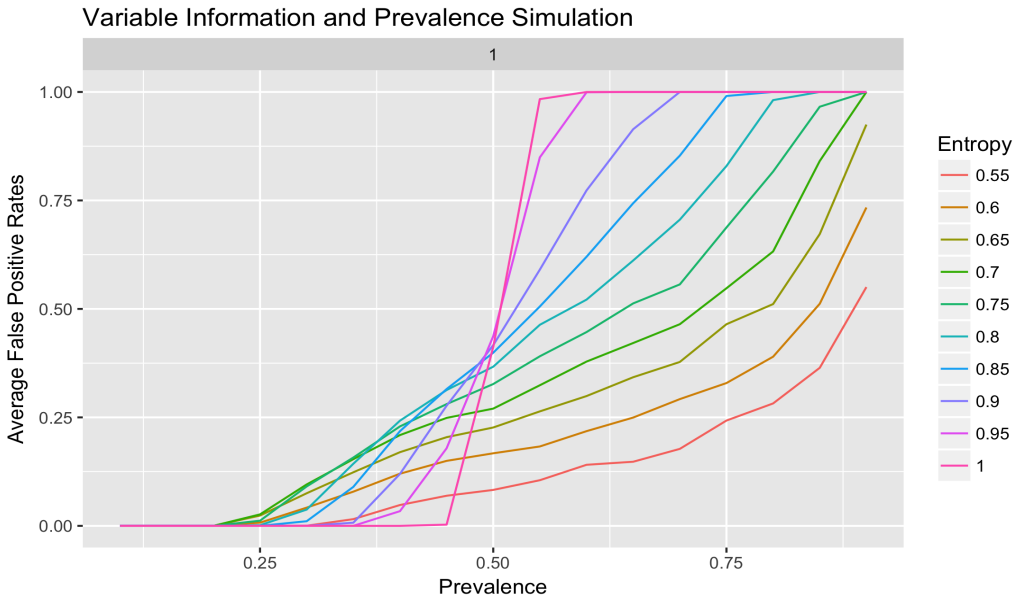


Figure 2.4: False Positive Rates at a Cutoff Value of 0.5

It is clear from this graph that the false positive rate is a function of the prevalence and the information in the model (the latter is highly related to the accuracy of the model). However there are some interesting nuances to

this data. Consider the case when $\kappa = 1$ and the prevalence is less than 0.5. Because the predictor is uninformative, the logistic regression will classify all the observations as ζ (see equation 2.2). At a cutoff of 0.5, no observations will be classified positively, so the false positive rate will be zero (Equation 1.1). Similarly, when $\zeta > 0.5$, all the observations will be classified positively and the false positive rate becomes 1.

The non-linear relationship between the false positive rate and the prevalence for each value of entropy is another intriguing aspect of the graph. It would be very interesting to model the exact mathematical relationship between prevalence, information, and false positive rate in order to understand the idiosyncrasies of this data and better predict the effect of over-policing on false positive rates.

Chapter 3

Estimating the Over-Policing Parameter Conditional on the Crime Likelihood

How can we estimate the over-policing parameter? This is the next major question that this thesis considers. In order to tackle this estimation problem we are going to assume that we have some theory on how the environment and attributes of an individual affect their likelihood of entering crime. Our simulation assumes a very specific relationship between age, gender, priors, and crime likelihood, but we do not believe that these are the only necessary variables or that we modeled the relationship completely.

Papers like "Sex and involvement in deviance/crime: A Quantitative Review of the Empirical Literature" and "The victimful-victimless crime distinction, and seven universal demographic correlates of victimful criminal behavior" [4][5] emphasize the importance of sex and age, but also describe other correlates and interaction effects inherent to criminal behavior that we did not account for. The key to modeling over-policing is assuming that $\hat{\gamma_i} = f(i^{th}$ individual's observable attributes) can be calculated for each individual. The only other major parameter that enters the model (See Algorithm 1) is θ. The value of θ affects the likelihood of an individual being thrown in jail conditional on being in crime. Lastly, the probability of leaving crime without going to jail is set at 0.2 for now, but this should be a parameter estimated from research as well.

Within our framework, there are three states an individual can occupy in any given year - out of jail and not in crime ($state_1$), out of jail and in crime ($state_2$), and in jail ($state_3$). Since the probabilities that determine the ways individuals transition from state to state aren't conditional on time, the process is memoryless and can be modeled using a Markov chain (tailored for every individual because they will all have different crime likelihoods). Our goal is to calculate the probability that an individual being sent to jail, a.k.a that they enter $state_3$, at any point in a 10 year period. If γ_i = the average of the crime likelihood in the first and last years of the simulation and θ = the over-policing parameter, then the transition matrix of this Markov

process describing the i^{th} individual is constructed as:

$$M = \begin{bmatrix} 1-\gamma_i & \gamma_i * (1-\theta) & \gamma_i * \theta \\ 0.2 * (1-\theta) & 0.8 * (1-\theta) & \theta \\ 0 & 0 & 1 \end{bmatrix}$$

This matrix is best understood in the following manner. The entry M_{pq} represents the probability of moving to state q conditional on being in state p. For example, the entry $M_{1,1}$ is $(1-\textit{crime likelihood})$ and represents the probability of staying out of crime conditional on not being in crime; entry $M_{2,1}$, which is $(\textit{probability of leaving crime}) * (1 - \textit{ probability of going to jail})$, is the probability of leaving crime conditional on being in crime. The state vector of an individual, x_ν, is a row vector where the τ^{th} column of the row vector is the probability of an individual being in $state_\tau$ in after ν years.

In order to find the probability of an individual who does not start in crime or in jail of being sent to jail after ν years, we multiply the initial placement vector

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = x_0$$

by the transition matrix in the following manner:

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1-\gamma_i & \gamma_i * (1-\theta) & \gamma_i * \theta \\ 0.2 * (1-\theta) & 0.8 * (1-\theta) & \theta \\ 0 & 0 & 1 \end{bmatrix}^{\nu} = x_\nu$$

When estimating θ, it is important to acknowledge that observational data has information about only two states - whether an individual has been caught by police or not ($state_3$ or $\{state_1$ and $state_2\}$). For a given year, we can calculate the maximum likelihood estimate for θ in the following manner:

n is the number of individuals in the populaiton;
for $individual_i$ *in the population* **do**
  Calculate γ_i (the crime likelihood) for $individual_i$ based on their gender, age and past priors according to equation 3.1;

$$\gamma_i = \frac{10}{age} + \frac{1}{30 - 20 * gender} + \frac{1}{18 - (priors * 2)^2} - \frac{1}{3} \tag{3.1}$$

  Note: The output of equation 3.1 for a specific simulated population is graphed in figure 2.1.;
  Create the transition matrix for $individual_i$ using γ_i and calculate the probability of $individual_i$ being in each state in a given year ν as a function of θ by multiplying the initial state vector with the transition matrix as shown above. Let $\pi_{\tau,i}$ represent the probability of being in $state_\tau$ for $individual_i$;
  Let $y_i = 0$ if $individual_i$ is in $state_1$ or $state_2$ (since we can't observe the difference between the two states), and $y_i = 1$ if $individual_i$ is in $state_3$;
end
Consequently, the likelihood function of θ for a given year ν is

$$L(\theta) = \prod_{i=1}^{n} \pi_{i,1}^{y_i} (\pi_{i,2} + \pi_{i,3})^{1-y_i} \; where \; \pi_{i,\tau} = g(\theta, \gamma_i) \tag{3.2}$$

Where g is some function of θ and γ_i calculated using the transition matrix.

Algorithm 4: Maximum Likelihood Estimation

For example, If we were going to calculate the product above with information form after the first year, we would begin by multiplying:

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1-\gamma_i & \gamma_i * (1-\theta) & \gamma_i * \theta \\ 0.2 * (1-\theta) & 0.8 * (1-\theta) & \theta \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1-\gamma_i & \gamma_i * (1-\theta) & \gamma_i * \theta \end{bmatrix}$$

The likelihood function in equation 3.2 would be:

$$L(\theta) = \prod_{i=1}^{n} (\gamma_i * \theta)^{y_i} (\gamma_i * (1-\theta) + 1 - \gamma_i)^{1-y_i}$$

Where y_i indicates whether or not the individual has been to jail after year 1.

Theoretically we could use use software to solve for the value of θ that maximizes the likelihood in equation 3.2 for year 10. In other words, solve for θ that satisfies the following equation:

$$\frac{d}{d\theta} \prod_{i=1}^{n} \pi_{i,1}^{y_i} (\pi_{i,2} + \pi_{i,3})^{1-y_i} = 0$$

However, solving for the MLE of θ using the information contained in year 10 of the simulation analytically would involve solving a polynomial with more than 1000 terms. Instead of attempting this colossal task, we will take

advantage of the fact that θ exists within a relatively small space (from 0 to 1) to grid search for an estimate of the θ that maximizes the likelihood above. We tested this approach on a population that is over-policed with true $\theta = 0.15$. In order to get a sense for the bias and variance of $\hat{\theta}$, we built the sampling distribution by estimating the parameter over many simulations.

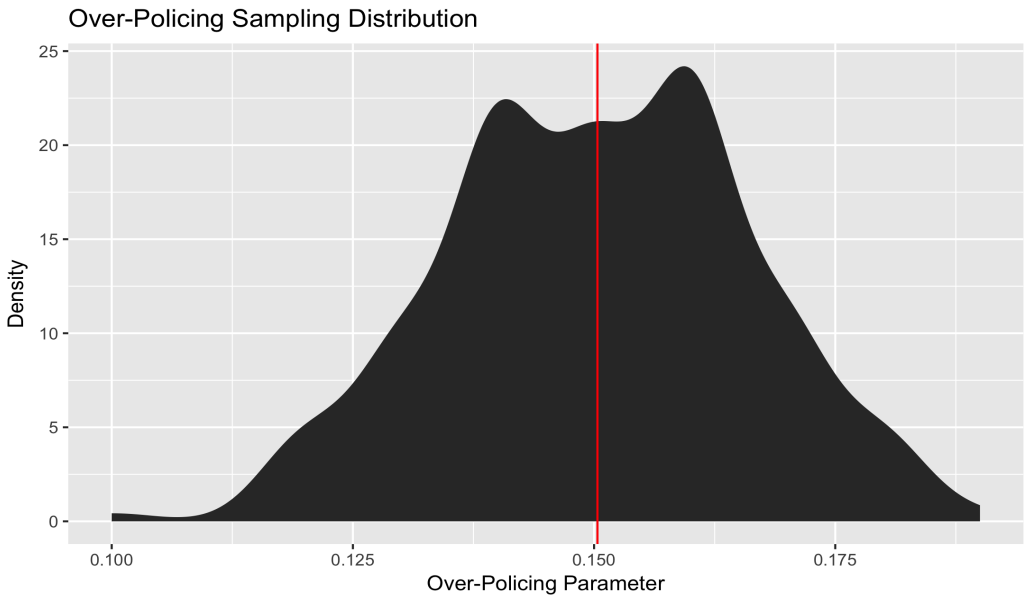


Figure 3.1: Histogram of Estimated θ's (True $\theta = 0.15$)

From this graph, it is clear that the bias and variance are small! The mean of the distribution (the red line above) is 0.15035 and the standard deviation is 0.015. However, if we look at the sampling distribution when the true over-policing parameter equals 0.2:

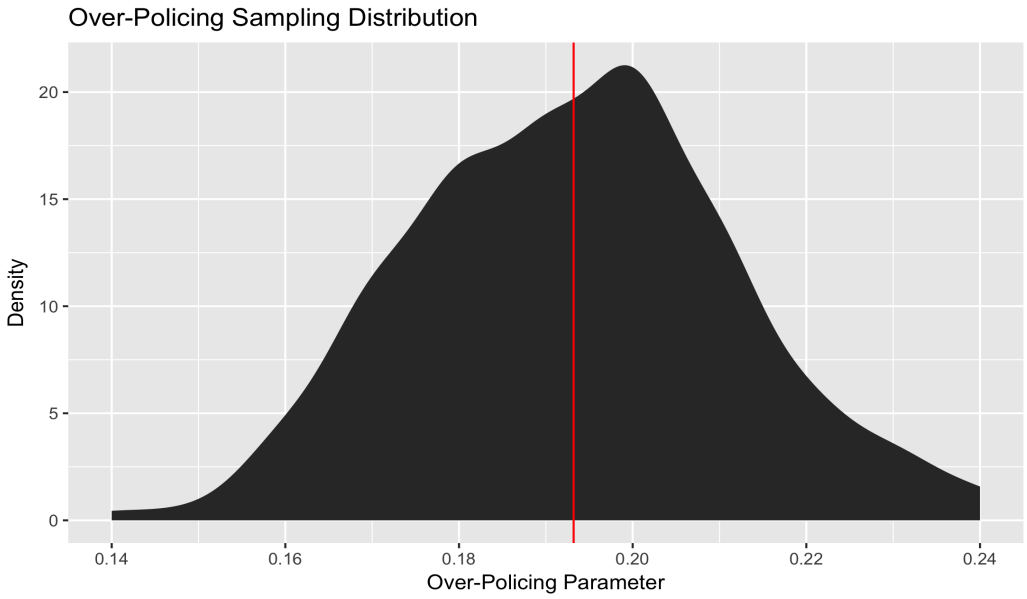


Figure 3.2: Histogram of Estimated θ's (True $\theta = 0.2$)

It appears that both the variance and the bias are higher for a more extreme over-policing parameter! However, in order to confirm this relationship I would need to conduct many more trials. The mean is now 0.1932 and the standard deviation of the θ estimates is 0.0185.

Is it possible to obtain a better estimate of θ? Let's looks at how the standard deviation is related to the size of the population we sample. The sampling distribution of the MLE asymptotically converges to approximately a normal distribution centered around the true parameter with variance $\frac{1}{I_n(\theta)}$, where $I_n(\theta)$ is the Fisher information in a sample of size n with respect to θ. Or in a more mathematical sense, it is the second derivative of the log likelihood

function. For context, an example of the empirical likelihood calculated on data generated from a model with $\theta = 0.2$ is:

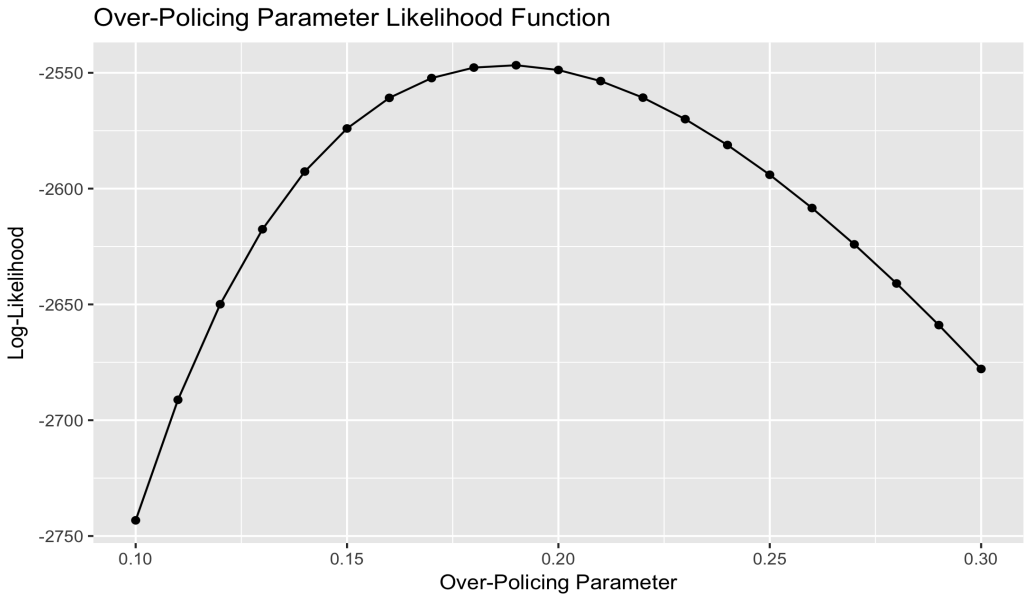


Figure 3.3: An example of an empirical likelihood function generated while creating the sampling distribution (True $\theta = 0.2$)

But, how do we find the Fisher information of θ explicitly? Finding this value usually involves taking the expected value of the likelihood function's 2^{nd} derivative. However, given the incredibly complicated nature of the likelihood we are considering, it will be very hard if not impossible to solve this problem analytically. Therefore, we will again appeal to numerical methods.

Since the grid of possible θ values we use to estimate the MLE is evenly spaced, we can use an approximation of $\frac{d^2 f(\theta)}{d\theta^2}$ (where $f(\theta)$ is the likelihood of θ) based on central difference coefficients. In order to derive the coefficients

for this estimation we will begin with the Taylor expansions of $f(x \pm h)$ and $f(x \pm 2h)$. The Taylor series will be truncated at the 4th terms in order to achieve a 4th order accuracy approximation [3].

$$f(x+h) \approx f(x) + hf'(x) + \frac{h^2 f''(x)}{2!} + \frac{h^3 f^{(3)}(x)}{3!} + \frac{h^4 f^{(4)}(x)}{4!} \tag{3.3}$$

$$f(x-h) \approx f(x) - hf'(x) + \frac{h^2 f''(x)}{2!} - \frac{h^3 f^{(3)}(x)}{3!} + \frac{h^4 f^{(4)}(x)}{4!} \tag{3.4}$$

$$f(x+2h) \approx f(x) + 2hf'(x) + \frac{4h^2 f''(x)}{2!} + \frac{8h^3 f^{(3)}(x)}{3!} + \frac{16h^4 f^{(4)}(x)}{4!} \tag{3.5}$$

$$f(x-2h) \approx f(x) - hf'(x) + \frac{4h^2 f''(x)}{2!} - \frac{8h^3 f^{(3)}(x)}{3!} + \frac{16h^4 f^{(4)}(x)}{4!} \tag{3.6}$$

We can eliminate the odd terms by adding equations 3.3 to 3.4 and equations 3.5 to 3.6:

$$f(x-h) + f(x-h) \approx 2f(x) + 2\frac{h^2 f''(x)}{2!} + 2\frac{h^4 f^{(4)}(x)}{4!} \tag{3.7}$$

$$f(x-2h)+f(x-2h) \approx 2f(x)+8\frac{h^2 f''(x)}{2!}+32\frac{h^4 f^{(4)}(x)}{4!} \tag{3.8}$$

If we multiply equation 3.7 by 16 and subtract 3.8 in order to eliminate the 4th derivative term, we are left with:

$$16f(x-h)+16f(x-h)-f(x-2h)-f(x-2h) \approx 30f(x)+24\frac{h^2 f''(x)}{2!} \tag{3.9}$$

Solving for $f''(x)$:

$$f''(x) \approx \frac{16f(x-h)+16f(x-h)-f(x-2h)-f(x-2h)-30f(x)}{12h^2} \tag{3.10}$$

If $f(\theta_{MLE})$ is the likelihood function evaluated at the MLE, $f(\theta_{MLE} \pm i\Delta x)$ is the likelihood function evaluated at the i^{th} index of the grid above or below the index of the MLE, where Δx is the uniform distance between the grid indexes, Equation 3.10 becomes:

$$f''(\theta_{MLE}) \approx \frac{1}{(\Delta x)^2}\Bigg(-\frac{1}{12}f(\theta_{MLE}-2\Delta x)+\frac{4}{3}f(\theta_{MLE}-\Delta x)-\frac{5}{2}f(\theta_{MLE}) \\ +\frac{4}{3}f(\theta_{MLE}+\Delta x)-\frac{1}{12}f(\theta_{MLE}+2\Delta x)\Bigg) \tag{3.11}$$

Over ten simulated samples, the average estimated Fisher information contained in a sample of one thousand observations is 2891.67.

The MLE is asymptotically distributed:

$$(\theta_{MLE}-\theta)\sqrt{I_n(\theta)} \sim N(0,1)$$

For high values of n, the sampling distribution is approximately (as stated above):

$$\theta_{MLE} \overset{approx}{\sim} N\left(\theta, \frac{1}{I_n(\theta)}\right)$$

The approximate normal distribution (graphed in black) based on the estimated Fisher information together with the estimated sampling distribution, depicted as a smoothed histogram using a kernel density estimator (in red),

can be seen in Figure 3.4:

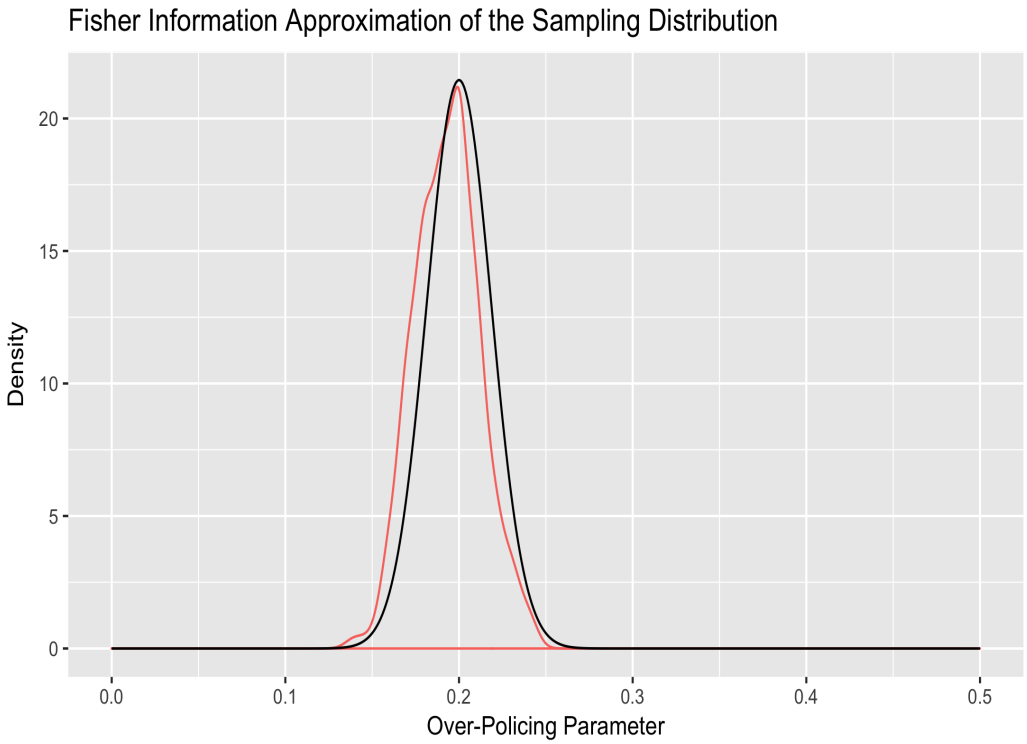


Figure 3.4: Fisher Normal Approximation of the Sampling Distribution

The normal distribution is a great approximation of the sampling distribution! These results have two main implications. First, it means that the standard deviation of the numerically estimated MLE depends on the sample size and the Fisher information calculated from our grid estimated likelihood in the ways we would expect. Considering the number of approximations involved in this process, the similarity between the two distributions represented above is pretty amazing. This is strong evidence indicating that if we

increased the size of our sample, we would decrease the standard error of our estimator. Second, it was very computationally expensive to calculate the standard deviation of the MLE estimate using an empirical sampling distribution (we would bootstrap to find this sampling distribution if we had real data). Being able to quickly compute the standard error of the over-policing estimator based on the Fisher information in the sample could be very useful in further applications.

Chapter 4

Future Directions

Although our simulation models general processes with an element of randomness, it lacks an explicit connection to reality. In particular, the percentage of the population that goes to jail and the recidivism rate is not based on crime records. Further applications of this thesis could closely model a current population in order to get a better sense for how well our estimation technique handles realistic over-policing parameters.

It is also important to note that the COMPAS algorithm predictions contained another layer of complexity. While the false positive rate of these predictions was higher for black individuals, the overall model accuracy was surprisingly equal across race. Our model does not produce this nuance. Figure 4.1 demonstrates that the overall predictive accuracy is very different

across race and over-policing values. Future work on this topic will hopefully test the robustness of our results for different accuracy relationships, including a model that better reflects the realities of the COMPAS algorithm predictions. Additionally, it seems that the logistic regression model is not capturing the structure of the data (i.e. isn't predicting well). This problem will need to be resolved before continuing this research.

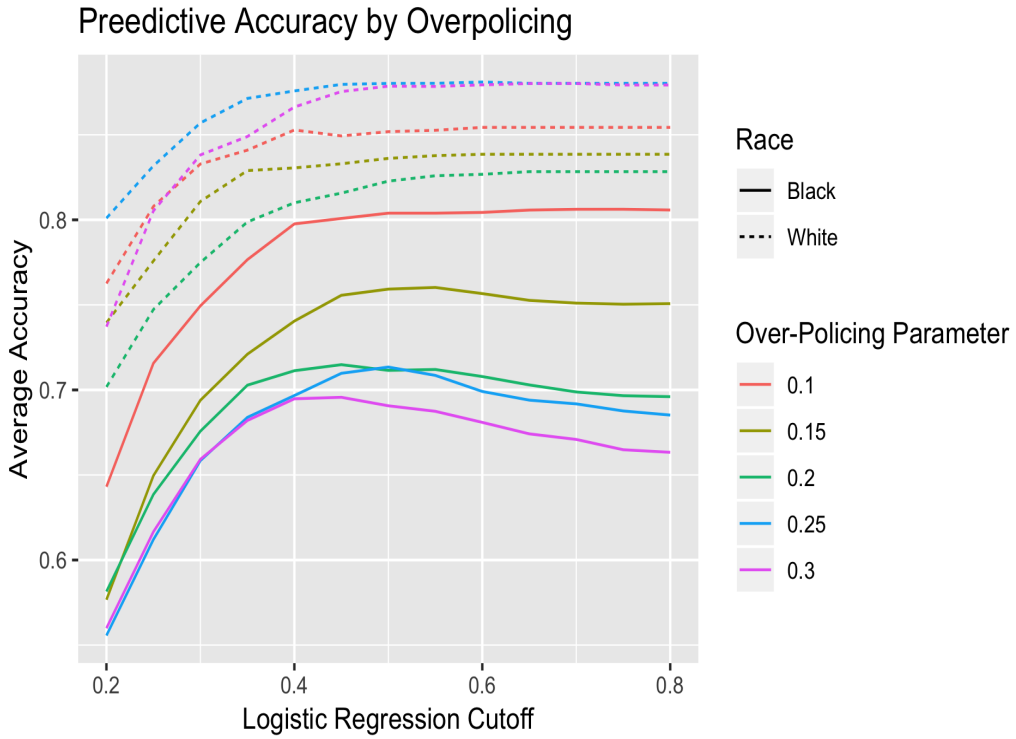


Figure 4.1: Predictive Accuracy Across Race and Over-Policing Values

.

Chapter 5

Conclusion

Consider the implications of applying the steps in this work to an actual population. If we assume that a positive identification by the algorithm (an individual is predicted to recidivate) factors into whether the individual is sent to jail, then the logistic regression's high false positive rates will augment the effect of over-policing (which, in turn, results in more false positives), creating a vicious feedback loop that only ends when every member of the population is incarcerated.

As demonstrated by others and confirmed by this thesis, overall model accuracy is not an adequate measure of algorithmic validity. As models assume a larger role in decisions about criminal justice and other important domains, it is necessary to delve into the intricacies of different error types and their

context-dependent impacts on lives. This thesis was not an attempt to accurately model the impact of over-policing on disadvantaged populations in specific communities. Rather, it operated on a powerful hypothetical. What if you had two identical populations separated by a biased sampling mechanism? What if a false positive meant additional years of jail time? How could we begin to examine the possible mechanisms that lead to different false positive rates? How do classification models like the logistic regression balance different types of information inherent to skewed data? And, finally, once we identify possible reasons for differential errors, how do we estimate the effect of unfair sampling techniques (over-policing) when there are a series of complex variables and relationships at play? This thesis begins to scratch the surface of these questions. Figure 2.3 tells an important story. It demonstrates that over-policing is a possible mechanism for explaining the large number of black defendants falsely labeled high risk by an algorithm like COMPAS. By empirically modeling the relationship between over-policing and false positive rates, the thesis helps to conceptualize the way over-policing may directly affect high false positive rates. Through a statistical lens, over-policing is the implementation of a biased sampling mechanism, and the thesis suggests possible methods for estimating the degree of the bias (maximum likelihood estimation of the over-policing parameter). There are many other rabbit holes to explore, from different definitions of fairness to different classification algorithms to the different context-relevant impacts of error relationships. It is imperative that progress be made on

these topics in order to ensure that individuals are treated equitably in an algorithm-dominated world.

Bibliography

[1] Kuiper and Sklar; *Practicing Statistics* Pearson Education, Boston-Massachussets, 2015.

[2] Barocas, Hardt, and Narayanan; "*Fairness and Machine Learning*", fairmlbook.org, 2018.

[3] Fink and Mathews; "*Numerical Methods Using MATLAB, 4th Edition*",Prentice-Hall Inc., 2004.

[4] Smith, Douglas A., and Christy A. Visher; "*Sex and involvement in deviance/crime: A Quantitative Review of the Empirical Literature*", American Sociological Review, vol. 45, no. 4, 1980, pp. 691?701. JSTOR, www.jstor.org/stable/2095016.

[5] Ellis, Lee "*The victimful-victimless crime distinction, and seven universal demographic correlates of victimful criminal behavior*", Personality and Individual Differences, Volume 9, Issue 3, 1998, pp. 525-548.

[6] Espino, Luis; "*Racism without a Face: Predictive Statistics in the Criminal Justice System*", 2018.

[7] Wexler, Rebecca; "*Code of Silence*", CHANCE Magazine, 2018.

[8] Angwin, Larson, Mattu, Kirchner; "*Machine Bias*", ProPublica, 2016.

[9] "*Griggs v. Duke Power Company*", 401 U.S. 424, 1970.