2016

Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models

Kory D. Johnson
University of Vienna

Dean P. Foster
University of Pennsylvania

Robert A. Stine
University of Pennsylvania

Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models

Abstract
Fairness aware data mining aims to prevent algorithms from discriminating against protected groups. The literature has come to an impasse as to what constitutes explainable variability as opposed to discrimination. This stems from incomplete discussions of fairness in statistics. We demonstrate that fairness is achieved by ensuring impartiality with respect to sensitive characteristics. As these characteristics are determined outside of the model, the correct description of the statistical task is to ensure impartiality. We provide a framework for impartiality by accounting for different perspectives on the data generating process. This framework yields a set of impartial estimates that are applicable in a wide variety of situations and post-processing tools to correct estimates from arbitrary models. This effectively separates prediction and fairness goals, allowing modelers to focus on generating highly predictive models without incorporating the constraint of fairness.

Keywords
Fairness-Aware Data Mining, Multiple Regression

Disciplines
Physical Sciences and Mathematics

Submitted to Statistical Science

Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models

Kory D. Johnson, Dean P. Foster, and Robert A. Stine

University of Vienna, Amazon, and University of Pennsylvania

Abstract. Fairness aware data mining aims to prevent algorithms from discriminating against protected groups. The literature has come to an impasse as to what constitutes explainable variability as opposed to discrimination. This stems from incomplete discussions of fairness in statistics. We demonstrate that fairness is achieved by ensuring impartiality with respect to sensitive characteristics. As these characteristics are determined outside of the model, the correct description of the statistical task is to ensure impartiality. We provide a framework for impartiality by accounting for different perspectives on the data generating process. This framework yields a set of impartial estimates that are applicable in a wide variety of situations and post-processing tools to correct estimates from arbitrary models. This effectively separates prediction and fairness goals, allowing modelers to focus on generating highly predictive models without incorporating the constraint of fairness.

MSC 2010 subject classifications: Primary 62P25, ; secondary 62J05.
Key words and phrases: Fairness-Aware Data Mining, Multiple Regression.

1. INTRODUCTION

Machine learning has been a boon for improved decision making. The increased volume and variety of data has opened the door to a host of data mining tools for knowledge discovery; however, automated decision making using vast quantities of data needs to be tempered by caution. In 2014, President Obama called for a 90-day review of big data analytics. The review, "Big Data: Seizing Opportunities, Preserving Values," concludes that big data analytics can cause societal harm by perpetuating the disenfranchisement of marginalized groups (House, 2014). Fairness aware data mining (FADM) aims to address this concern.

Broadly speaking, the goal of this project is to allow increasingly complex methods to be used without fear of infringing upon individuals' rights. This will be beneficial in all domains that have the potential for discrimination on the basis on data. Applications abound in both the private sector and academics.

Oskar-Morgenstern-Platz 1, A-1090 Vienna (e-mail: korydjohnson@gmail.com) 7 W 34th St, New York, NY 10001 (e-mail: dean@foster.net) 3730 Walnut St., Philadelphia, PA 19104 (e-mail: stine@wharton.upenn.edu)

Companies will be able to justify the use of partially automated decision making in areas as diverse as loan applications, employment, and college admissions. There will be clear fairness criteria to guide the construction of fair models, thus reducing unintentional discrimination and litigation. A proper understanding of fairness will inform regulatory agencies and policy makers such that they can promote fairness and understand its statistical implications. In legal disputes, a set of models of fairness provides a baseline from which detrimental impact can be assessed.

A simple example clarifies the issue of fairness. Consider a bank that wants to estimate the risk in giving an applicant a loan. The applicant has "legitimate covariates" such as education and credit history, that can be used to determine their risk. They also have "sensitive" or "protected" covariates such as race and gender. Lastly, we introduce a third covariate group of "suspect" or "potentially discriminatory" covariates. The canonical example of a suspect covariate is location information such as the applicant's address. While location is not a protected characteristic such as race, it is often barred from use given the ability to discriminate using it. In order to determine the interest rate of the loan, the bank uses historical data to estimate the credit worthiness of the candidate. FADM asks whether or not the estimates the bank constructs are fair. This is different than asking if the data are fair or if the historical practice of giving loans was fair. It is a question pertaining to the estimates produced by the bank's model. This generates several questions. First, what does fairness even mean for this statistical model? Second, what should the role of the sensitive covariates be in this estimate? Third, how do legitimate and suspect covariates differ? Lastly, how do we constrain the use of the sensitive covariates in black-box algorithms?

Our contributions fall into two main categories: conceptual and algorithmic. First, we provide a *statistical* theory of fairness which revolves around impartiality. The literature lacks a serious discussion of the philosophical components of fairness and how they should be operationalized in statistics. Doing so will require a spectrum of models to be defined, because fairness is a complicated philosophical topic. We address these complications by introducing and demonstrating the importance of the suspect covariate group. Second, after providing this framework, it will be clear how to both construct fair estimates using simple procedures as well as correct black-box estimates to achieve fairness.

It is important to note that these corrections can only be made by using all of the data, including the sensitive covariates. This is intuitively clear because guaranteeing that discrimination has not occurred requires checking the estimates using sensitive characteristics. Having a spectrum of models of fairness also yields two practical applications. First, we can quantify the cost of government programs by considering the different incentives between a profit-maximizing firm and a government-owned one that is constrained to be fair. Second, we can quantify the cost of discrimination, which is an important component of litigation.

The main body of the paper is organized as follows: Section 2 gives notation and defines impartial predictions. Section 3 provides a brief philosophical and legal background of fairness and is motivated by a long history of literature in ethics and political philosophy. Section 4 constructively generates fair estimates using multiple regression. We also compare estimates on an individual level in Section 4.5. The literature lacks such a comparison even though it is crucially important when arguing estimates are fair. Often social discussions revolve around an individual being treated (un)fairly due to their membership in a protected group. Section 5 uses the methods generated in Section 4 to correct estimates from black-box models to achieve fairness. In light of Section 4, this is a straightforward task. We also test our methods on a data example to not only elucidate the conceptual difficulties in the literature, but also to demonstrate that our method achieves superior results.

A few remarks need to be made about the sensitive nature of the topic at hand. These closely follow Holland (2003), and interested readers are referred there. Sensitive covariates such as race and gender are not neutral concepts. It is the plight of the data analyst that these categories are taken as given. We assume that data are provided in which someone else has determined group membership. Our questions are about the types of protection that can be offered given such a data set. Furthermore, this project is descriptive, not normative. Our goal is to provide data-generating models that elucidate philosophical nuances. Each scenario gives rise to a different fair estimate; however, determining which scenario is accurate is outside of the scope of this project.

2. DEFINING IMPARTIAL ESTIMATES

To motivate our definition of impartial predictive modeling, we consider a couple of suggestions which are incorrect. Our goal is to provide impartial estimates of credit risk Y, given legitimate covariates $\mathbf{x}$, sensitive covariates $\mathbf{s}$, and suspect covariates $\mathbf{w}$. These are connected through an unknown, joint probability distribution $\mathbb{P}(Y, \mathbf{x}, \mathbf{s}, \mathbf{w})$. Our data consists of n iid draws from this joint distribution. The standard statistical goal is to estimate the conditional expectation of Y given the covariates:

$$Y = \mathbb{E}[Y|\mathbf{x}, \mathbf{s}, \mathbf{w}] + u$$

where u has mean zero and is uncorrelated with all functions of the covariates. Our goal is to estimate a conditional expectation that is impartial with respect to the sensitive covariates. Intuitively, impartiality requires that the sensitive covariates do not influence estimates. Through the following examples, we demonstrate the need for a more refined definition of impartiality.

Impartial estimates are fair because the sensitive covariates are chosen to be normatively relevant. In the FADM literature, the covariate groups are always assumed to be provided. As such, the statistical task is disjoint from the normative task of identifying covariate groups. This is why FADM techniques can be used in domains such as batch effect analysis as discussed in Calders et al. (2013). When analyzing batch effects, the covariate on which impartiality is desired is merely batch membership.

For the remainder of the paper, we will use the term "impartial" to describe the statistical goal. This is done in order to separate our task from normative complications. That being said, the different covariate groups have normative significance and need to be differentiated. Therefore, when referring to covariate groups and during the background discussion of Section 3 we use the normative language of fairness.

The first conjecture for impartial estimates can be described as "impartiality as indifference." That is, estimates are impartial if changing the value of the

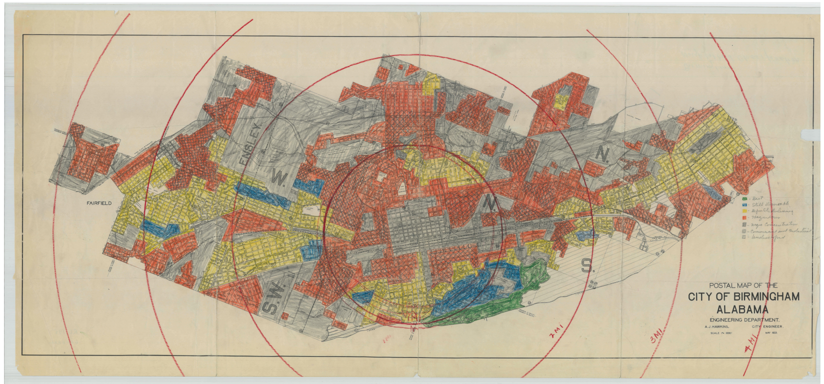


Fig 1: A 1933 map of Birmingham, Alabama marking high and low-risk areas (FHLBB, 1935).

sensitive covariate does not change the estimate. We overload the notation $\hat{Y}$ to be both the estimated credit risk as well as a function that outputs the credit risk given covariates; namely: $\hat{Y} = \hat{Y}(\mathbf{x}, \mathbf{s}, \mathbf{w})$.

DEFINITION 1 (Indifference). *An estimate* $\hat{Y}(\mathbf{x}, \mathbf{s}, \mathbf{w})$ *is indifferent to* $\mathbf{s}$ *if, given two groups of* $\mathbf{s}$, s_+ *and* s_-,

$$\hat{Y}(\mathbf{x}, \mathbf{s}_+, \mathbf{w}) = \hat{Y}(\mathbf{x}, \mathbf{s}_-, \mathbf{w}).$$

There are two problems with this. First, it doesn't allow for affirmative action, which many believe is necessary to be impartial or fair. Our proposal easily accounts for affirmative action policies and clearly indicates the assumptions necessary for their validity. This is done by drawing distinctions between legitimate and suspect covariates, which is motivated in Section 3. Second, indifferent estimates can be accomplished just by removing the sensitive covariates prior to model fitting. Multiple authors have raised doubts that the legal requirement of removing race prior to fitting a model is sufficient to achieve fairness (Kamishima et al., 2012; Kamiran et al., 2013). Due to the relationships between race and other covariates, merely removing race can leave lingering discriminatory effects that permeate the data and potentially perpetuate discrimination. The use of covariates associated with $\mathbf{s}$ to discriminate is called redlining.

The term "redlining" originated in the United States to describe maps that were color-coded to represent areas in which banks would not invest. Figure 1 shows one such map from Birmingham, Alabama in the 1930s. It is marked in different colors to indicate the riskiness of neighborhoods. For example, red indicates hazardous areas and blue indicates good areas for investment. Denying lending to hazardous areas appears facially neutral because it is race blind: the bank need not consider racial information when determining whether to provide a loan. These practices, however, primarily denied loans to black, inner-city neighborhoods. This was used as a way to discriminate against such borrowers without needing to observe race. This clearly demonstrates that merely excluding sensitive information does not remove the possibility for discrimination. Conceptually, the core issue is the *misuse* of available information.

To avoid redlining, previous authors have proposed measures of group discrimination which need to be reduced to zero. For example, if $\mathbf{s}_+$ and $\mathbf{s}_-$ represent two sets of sensitive covariates, the CV-discrimination score (Calders and Verwer, 2010) is

$$DS = \mathbb{E}[\hat{Y}|\mathbf{s}_+] - \mathbb{E}_n[\hat{Y}|\mathbf{s}_-].$$

This methodology is flawed, however, because it presupposes the solution to challenging problems in casual inference. First, defining discrimination in this way implicitly treats $\mathbf{s}$ as a causal variable. While different protected groups can clearly have different experiences and outcomes, treating $\mathbf{s}$ as a casual variable is fraught with difficulties (Holland, 2003). Furthermore, the discrimination score ignores the role of other explanatory variables. In particular, some differences in expected credit worthiness may be explained by legitimate covariates. Such explainable variation needs to be separated from discrimination. Subsequent research has attempted to disentangle explainable variation between sensitive groups and discrimination. This is done either through naive-Bayes models (Calders and Verwer, 2010) or crude propensity score stratification (Calders et al., 2013). As we will introduce in Section 3 and demonstrate in Section 4, these discussions are incomplete and conflate two different effects.

For simplicity, we define impartiality with respect to the linear projection of Y on to a set of covariates $\mathbf{v}$:

$$Y = L(Y|\mathbf{v}) + u,$$

where the error term u satisfies

$$\mathbb{E}[u] = 0, \quad \mathrm{Cov}(\mathbf{v}, u) = \mathbf{0}.$$

This allows core ideas to be fully explained in a familiar framework. The important steps of the process, however, can be accomplished with more robust modeling techniques. We are currently working on a paper to explain the implications of doing so.

Definition 2 (Impartial Estimates and Impartiality Score). *An estimate of Y, $\hat{Y}$, is impartial if its residual, $\hat{u} = Y - \hat{Y}$, and linear projection on $\mathbf{x}$, $L(\hat{Y}|\mathbf{x})$, satisfy*

$$\begin{aligned}
Var(\hat{u})^{-1/2}\mathbb{E}[\hat{u}] &= Cor(\hat{u}, \mathbf{s})\, Cor(\mathbf{s})^{-1} diag(Var(\mathbf{s}))^{-1/2}\mathbb{E}[\mathbf{s}] && (1)\\
Cor(\hat{u}, \mathbf{x}) &= Cor(\hat{u}, \mathbf{s})\, Cor(\mathbf{s})^{-1} Cor(\mathbf{s}, \mathbf{x}) && (2)\\
Cor(\hat{u}, \mathbf{w}) &= Cor(\hat{u}, \mathbf{s})\, Cor(\mathbf{s})^{-1} Cor(\mathbf{s}, \mathbf{w}) && (3)\\
Cor[\eta, \mathbf{s}] &= \mathbf{0} \quad where && (4)\\
\eta &= \hat{Y} - L(\hat{Y}|\mathbf{x}),
\end{aligned}$$

We define the Impartiality Score *(IS) as the sum of the absolute differences between the left and right hand sides of equations (1 - 4), normalized by the total number of covariates. Our models center $\mathbf{s}$, in which case $0 \leq IS \leq 1$.*

While our definition uses linear projections, estimates are not required to be linear. We use this fact in Section 5 to correct "black box" estimates created from a random forest. Similarly, the covariates $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{w}$ need not only include main effects. Interactions within groups as well as between groups are possible. It is only important to treat the resulting interactions as covariates of the appropriate type. Interactions within a covariate group will be of that group, while interactions with suspect variables will be considered suspect and interactions between sensitive and legitimate covariates are considered legitimate. This follows from the interpretations of suspect and legitimate covariates that are provided in Section 4.

Furthermore, the conditions in Definition 2 do not uniquely specify the estimate $\hat{Y}$. For example, consider adding independent, mean-zero noise to $\hat{Y}$. This merely degrades the performance of the estimate. The estimates are still impartial since the noise is independent of the sensitive attributes $\mathbf{s}$.

The criteria in Definition 2 are identified via impartial estimates constructed in Section 4. In the simplest case that has only legitimate and sensitive covariates, coefficients are estimated in the full regression model but predictions are made using only legitimate covariates. Explanatory covariates are centered such that removing the sensitive covariates during prediction does not change the estimated mean. The reverse regression literature in economics uses these estimates as a preprocessing step (Goldberger, 1984). That literature did not justify this as a fair estimate. We do so here and extend the estimates to more philosophically robust settings . Suspect covariates require an additional preprocessing step to ensure impartiality.

Identifying impartial estimates and fairness constructively has both philosophical and empirical support. The philosopher John Rawls discusses fair institutions as the method of achieving fairness (Wenar, 2013). Similarly, Brockner (2006) explains the importance of process fairness as opposed to merely outcome fairness. Process fairness focuses on how people are treated throughout the process of a decision whereas outcome fairness focuses on the results of the decision. The authors identify several examples in which firms attempt to layoff workers in a fair manner. Workers often feel that the decisions are fair when they are consulted frequently and the process is transparent, even if their severance packages are far worse. This points out the importance of fair treatment as fair use of information, not merely a measure of the outcome. This is discussed in the following section.

3. BACKGROUND

Fairness is a vague term and is used broadly; however, it is closely related to equality. This connection provides a rigorous way to discuss fairness. Colloquially, fairness requires that "similar people are treated similarly." This mirrors proportional equality, which is one of several principles of equality. "A form of treatment of others or distribution is proportional or relatively equal when it treats all relevant persons in relation to their due" (Gosepath, 2011). This includes the concept of merit as well as a need to specify a metric for similarity between people and the relationship between merit and outcome. "Merit" here merely means a value which is relevant to the decision at hand. For example, when discussing salary, merit is productivity for a given job. Alternatively, when discussing loans, merit is the riskiness of the borrower. In this paper, estimates are truly proportionally equal because we focus on linear models; however, different models may yield estimates of merit which predict outcomes better. This is left for future work.

The classic alternative to proportional equality is strict equality or equality of outcome. Strict equality makes the perhaps trivial claim that if two things are equal in the relevant respects then they should be treated equally. The issue, of course, is determining what criteria are relevant and how to measure similarity. In some cases, merit is irrelevant and strict equality is desired, such as with civil liberties and basic rights.

FADM addresses areas of justified unequal treatment. In the loan application example, people can receive different interest rates, but the manner in which the estimates differ needs to be controlled. Other considerations that are commonly thought to belong to this group include: "need or differing natural disadvantages (e.g. disabilities); existing rights or claims (e.g. private property); differences in the performance of special services (e.g. desert, efforts, or sacrifices); efficiency; and compensation for direct and indirect or structural discrimination (e.g. affirmative action)" (Gosepath, 2011).

FADM assumes that a list of sensitive characteristics is provided and is motivated by the requirement that sensitive covariates are not relevant measures of similarity in many applications. We will refer to the latter as the fairness assumption:

Definition 3 (Fairness Assumption). *Sensitive covariates are not or ought not be a relevant source of variability or merit.*

Achieving fairness requires sensitive characteristics to not influence the outcome. Due to the possibility of redlining, removing sensitive covariates is insufficient. In order to be more precise, we argue that decisions should be impartial with respect to the protected covariates. Impartiality is easiest to characterize in a negative manner: "an impartial choice is simply one in which a certain sort of consideration (i.e. some property of the individuals being chosen between) has no influence" (Jollimore, 2014).

The need to ignore sensitive information while acknowledging merit differences is addressed by the literature in equality of opportunity. Equality of opportunity is widely appealed to in both legal and philosophical communities as well as being "the frequently vague minimal formula at work in every egalitarian conception of distributive justice" (Gosepath, 2011). Furthermore, the examples that FADM is commonly concerned with are about opportunity, not merely distribution of goods. For example, the opportunity for a home loan or the opportunity to attend university. Equality of opportunity provides a spectrum of restrictions that are concrete enough to be translated into statistical models. Therefore, this is the measure of equality that we will use.

The philosophical literature on equality of opportunity analyses the way in which benefits are allocated in society (Arneson, 2015). A benefit can be anything from a home loan or high salary to college admission and political office. One way of understanding equality of opportunity is formal equality of opportunity (FEO), which requires an open-application for benefits (anyone can apply) and that benefits are given to those of highest merit. Merit will of course be measured differently depending on the scenario or benefit in question.Therefore, the most productive employee receives a high salary, while the least-risky borrower receives a low interest rate loan. There is cause for concern if discrimination exists in either the ability of some individuals to apply for the benefit or in the analysis of merit.

The fairness assumption is that sensitive covariates are not relevant criteria by which to judge merit.

Substantive equality of opportunity (SEO) contains the same strictures as above, but is satisfied only if everyone has a genuine opportunity to be of high merit. In particular, suppose there are social restrictions or benefits that only allow one group to be of high merit. For clarity, consider a rigid caste system, where only the upper caste has the time and financial resources to educate and train their children. Only children born to upper-caste parents will be of high quality and receive future benefits. This can be true even when lower-caste individuals can apply for the benefits and benefits are given based on merit. In this case, proponents of SEO claim that true equality of opportunity has not been achieved. While many countries lack a caste system, some may argue that cycles of poverty and wealth lead to a similar regress in the reasons for the disparity between protected groups, for example.

The difference between FEO and SEO that is relevant to FADM is whether there is benign or prejudicial association between covariates. This lies at the heart of not only legal cases, affirmative action, and social science literature, but also the public debates about fair treatment. The benign association model assumes the relationship between the sensitive covariates $\mathbf{s}$ and legitimate covariates $\mathbf{x}$ is not due to social constraints. In fact, this is why they can be called legitimate. Suppose differences between groups are the result of different motivation via familial socialization. For example, if some communities impart a higher value of education to their children than others, the conditional distributions for educational attainment may be significantly different. These differences, however, appear legitimate. Rejecting this claim in favor of ignoring such merit differences raises questions about whether parents' rights to raise their children take priority over strict adherence to equal treatment among groups (Arneson, 2015; Brighouse and Swift, 2009).

Alternatively, suppose the relationship between sensitive covariates $\mathbf{s}$ and suspect covariates $\mathbf{w}$ is the *result* of either social restrictions or social benefits. This prejudicial association is why the covariates may be called suspect. For example, one group could be historically denied admission to university due to their group membership. This can produce similar observable differences between covariate distributions, in that the favored group has higher educational attainment than the disfavored group.

In FADM, these two cases need to be treated differently; however, determining whether a covariate is legitimate or suspect is in the domain of causal inference, social science, and ethics. Our interest is not in specifying which variables have benign versus prejudicial association, but in constructing impartial estimates once such a determination has been made. As we will see in the simple example of Section 4.5, we must consider the type of association between covariates in order to justify impartial estimates used in FADM. The literature lacks a discussion of these differences, leaving all previous estimates unjustified.

Data sets may differ in terms of which covariates are prejudicially associated with sensitive attributes. To continue the education example, compare and contrast the education systems in Finland and the United States. The Finnish system is predicated on equal educational opportunities instead of the quality of education (Sahlberg and Hargreaves, 2011). Regardless of the community in which students are raised, there is a reasonable expectation that they are provided the same access to education. In the United States, however, there are large differences in school quality. This may require education to be treated differently if one desires fair estimates in Finnish data or United States data.

Many of the relevant distinctions between formal and substantive equality of opportunity are mirrored in the US civil rights tradition by the principles of anti-classification and anti-subordination, respectively. The anti-classification or anti-differentiation principle states that "the government may not classify people either overtly or surreptitiously on the basis" of a sensitive covariate (Balkin and Siegel, 2003). Anti-subordination conceived under the "group disadvantaging principle", on the other hand, contents that laws should not perpetuate "the subordinate status of a specially disadvantaged group" (Fiss, 1976). Both legal theories constrain the use of sensitive covariates, and we formally provide the distinctions between their effects in this paper.

In the United States, legal cases on equality are based on two theories of discrimination outlined under Title VII of the U.S. Civil Rights Act. Disparate treatment is direct discrimination on the basis of a protected trait. It requires justification of the intent to discriminate based on the protected trait. An easy solution to prevent disparate treatment is merely to hide the information. Kamishima et al. (2012) termed this direct prejudice, providing the mathematical definition of its presence as conditional dependence of the response and sensitive covariates given the legitimate covariates.

Disparate impact is discrimination on the basis of another covariate which disproportionately effects a protected class. Under this tenet, a policy is not discriminatory by definition (in that it does not codify treating groups differently) but is discriminatory in practice. Kamishima et al. (2012) called this indirect prejudice, but incorrectly defined its presence as dependence of the response and sensitive covariates. Defining disparate impact requires a more refined notion of impartiality, one that is able to capture the distinction between explainable variability and discrimination. While initially introduced to govern employment, disparate treatment and disparate impact have been expanded to other domains. These concepts also govern legal cases in Europe, Australia, and New Zealand, though by other names such as "discrimination by subterfuge" or "indirect discrimination."

The canonical example of disparate impact is redlining. A bank may treat all individuals equally within each neighborhood; however, by deciding to build offices and provide loans in only select regions, lending practices may be discriminatory. While race is irrelevant in the statement of the policy, the racial homogeneity of many neighborhoods reveals this practice to be potentially discriminatory. The famous Schelling segregation models demonstrate that such homogeneity can arise even without the presence of strong preferences for being around members of a particular group (Schelling, 1971). While redlining occurs increasingly less often, two large cases were settled in Wisconsin and New Jersey in 2015. In Section 4.5, we provide a detailed numerical example that demonstrates redlining and our solution.

The FADM problem is fundamentally about what constitutes *explainable* variation. That is, what differences between groups are explainable due to legitimate covariates, and what differences are due to discrimination. More precisely, there are important distinctions between statistical discrimination and redlining. Statistical discrimination is defined as a sufficiently accurate generalization. In many ways, this is the statistical enterprise. For example, it is a sufficiently accurate generalization that individuals with good repayment history are more likely to repay future loans. Therefore, such applicants are considered to be of lower risk and receive lower interest rates. The Equal Credit Opportunity Acts of 1974 and the amendments in 1975 and 1976 allow such "discrimination" if it is "empirically derived and statistically valid." It is clear that "discrimination" in this case refers to distinguishing good and bad risks.

Statistical discrimination is contrasted with redlining, which is a negative consequence of the ability to estimate sensitive covariates using legitimate ones. This can be used to discriminate against a protected group without having to see group membership. In this case, "discrimination" is used to describe prejudicial treatment. While often clear from context, in the interest of avoiding confusion between a legitimate type of statistical discrimination and redlining, "discrimination" will be used in a normative, prejudicial sense. The exception is in the phrase "statistical discrimination," in which case we will be more precise. This will separate the normative and statistical uses of the term "discrimination."

The distinction between legal and illegal forms of statistical discrimination primarily arises due to which covariates are being used to make generalizations. The concept of a sensitive or protected characteristic prohibits its use for generalizations. For example, in the United States incarceration and race are associated: black males are significantly more likely to have been imprisoned at some point in their lives than white males. Actions based on such heuristics are often illegal, though they may be economically rational. Risse and Zeckhauser (2004) provide a richer account of these cases, addressing concerns surrounding racial profiling. They separate the notion of statistical discrimination from the larger setting of societal discrimination. The debate often centers on what is a "disproportionate" use of sensitive information (Banks, 2001). Both legitimate and suspect covariates may be used for redlining, as the only requirement is that they are correlated with the sensitive attributes. We identify and remove these redlining effects in the following section.

4. MATHEMATICAL MODELS OF IMPARTIALITY

With the above discussion in hand, we can place our work in context of the literature on FADM and discrimination. Afterward, we provide a detailed account of statistical discrimination and redlining.

4.1 Related Work

Other authors have offered answers to various aspects of the FADM problem. The first branch of research focused on how to perform fair classification while respecting group fairness criteria such as the discrimination score. This can be accomplished by penalizing discrimination during estimation (Calders and Verwer, 2010; Calders et al., 2013; Kamishima et al., 2012), modifying the raw data (Pedreschi et al., 2008; Hajian and Domingo-Ferrer, 2013), or modifying objective functions with fairness criteria (Kamiran et al., 2010; Zemel et al., 2013). This literature must separate explainable variability from discrimination (Kamiran et al., 2013). We improve upon their discussions by providing a simple, tractable formulation of impartiality that addresses issues often encountered in real data.

A second branch of research analyses fairness through the related subject of differential privacy (Dwork et al., 2011). In the language of the current paper, the authors assume that fair estimates of merit are provided and the decision maker must use them to map individuals to randomized decision rules. A similar definition of fairness in multi-armed bandits is offered by (Joseph et al., 2016), in which a decision maker is not allowed to give preferential treatment between groups until one group is known to be of higher merit than another. While this places a constraint on the learning process, the end result only satisfies FEO without incorporating legal constraints, thus failing to capture fairness in a philosophically robust way. Furthermore, depending on how merit is estimated, redlining can still occur.

There is a long history of work in economics on discrimination. A summary is provided in (Fan and Moro, 2011), and the literature can be understood as providing explanations for the presence of discrimination even when FEO is satisfied. It is interesting to note that the data case we consider that allows observable differences in merit between groups is regarded as the trivial case and lacks significant research (Phelps, 1972). Like Joseph et al. (2016), it is assumed that firms can acknowledge merit differences without addressing SEO. Most of the literature focuses on the existence of discrimination in market equilibrium without underlying differences in merit between protected groups (Arrow, 1973; Foster and Vohra, 1992; Coate and Loury, 1993).

4.2 Model Framework

Fairness in modeling will be explained via directed acyclic graphs (DAGs), which are also referred to as Bayesian or Gaussian networks or path diagrams. DAGs will be used to conveniently represent conditional independence assumptions. While often used as a model to measure causal effects, we are explicitly not using them for this purpose. As previously stated, our goal is to create impartial *estimates*, whereas the estimation of a causal effect would attempt to answer whether the historical data are fair.

As such, we do not require the same type of causal interpretation. This stems from a different object of interest: in causal modeling, one cares about a casual parameter or direct effect of the covariate of interest whereas in FADM we care about the estimates produced by the model. Estimating a causal effect requires considering counterfactuals. For example, estimating a treatment effect of a drug requires a comparison of a patient's outcome under both treatment and control, even though only one of these outcomes is observed.

As FADM is concerned with *estimates* from models as opposed to outcomes, counterfactuals are easily computed. This is trivial to accomplish because it only requires producing an estimate for a modified observation. We need not consider the performance of an individual with that set of covariates (or even if it exists). Therefore, we do not need recourse to the interventionist or causal components of standard causal models and can deal only with their predictive components. In short, we only use DAGs to represents the conditional independence assumptions made between variables.

Figure 2 provides an example DAG representing a possible set of variables and relationships in a simplified college admissions process. The variables are

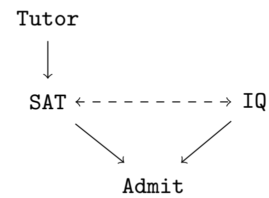


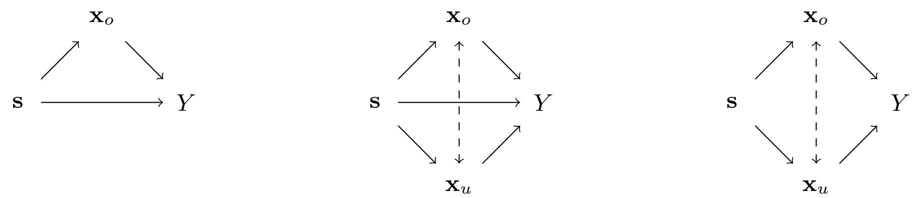
Fig 2: Example Directed Acyclic Graph (DAG)

Tutor, SAT Score, IQ, and Admit. These indicate a student's standardized test score (SAT), intelligence quotient (IQ), whether they received SAT tutoring, and whether they were admitted to a given university. In the graph, the variables are called nodes and are connected via directed edges. This direction captures a causal relationship: the value of Tutor causally relates to the value of SAT score. Dashed edged indicate a latent common cause: an unobserved variable U that causally effects both nodes. We use DAGs to concisely represent conditional independence assumptions. In the language of DAGs, two nodes are called "d-separated" by a set of nodes B if all of the paths (series of edges, regardless of direction) connecting the two nodes are "blocked" by a collection of nodes. The only criterion for blocked paths we will use is the following: a path is blocked by a set of nodes B if it contains a chain $i \to b \to j$ such that b is in B. If two nodes are d-separated given B, then the nodes are conditionally independent given B. For example, Tutor and Admit are d-separated (conditionally independent) given SAT and IQ. For further information on DAGs, see Pearl (2009).

The rest of this section introduces impartial estimates in stages via models in which the fairness assumption is tractable. These models correspond to formal interpretations of what particular notions of equality of opportunity require. We begin by enforcing FEO, which only uses sensitive and legitimate covariates. The goal in FEO is to have a best estimate of merit while satisfying the legal requirements of disparate treatment and disparate impact. Second, we consider a full SEO model (F-SEO), in which there are no legitimate covariates, only sensitive and suspect covariates. This model captures the philosopher John Rawl's theory of fair equality of opportunity. Subsection 4.5 provides a toy example to demonstrate the effect of considering covariates as either legitimate or suspect. The complete data case with sensitive, legitimate, and suspect covariates is considered in subsection 4.6. Lastly, subsection 4.7 discusses ways to use impartial estimates to estimate the cost of discrimination.

4.3 FEO

FEO is not concerned with potentially discriminatory covariates. Consider an idealized population model that includes all possible covariates. For the i'th individual, Y_i is credit risk, $\mathbf{s}_i$ contains the sensitive attributes (race, gender, age, etc), $\mathbf{x}_{o,i}$ contains the observed, legitimate covariates, and $\mathbf{x}_{u,i}$ contains the unobserved, legitimate covariates. Covariates $\mathbf{s}_i$, $\mathbf{x}_{o,i}$, and $\mathbf{x}_{u,i}$ are all bold to indicate they are column vectors, and this convention will be used throughout the paper. Unobserved covariates could be potentially observable such as drug use, or unknowable such as future income during the term of the loan. The data are assumed to have a joint distribution $\mathbb{P}(Y, \mathbf{s}, \mathbf{x}_o, \mathbf{x}_u)$, from which n observations



(a) Observed Data (b) Unrestricted Data (c) Fair Data

Fig 3: Observationally Equivalent Data Generating Models. No restrictions are placed on the dependence structure within the vectors $\mathbf{x}_o$, $\mathbf{x}_u$, and $\mathbf{s}$.

are drawn. The fairness assumption requires that $\mathbf{s}$ is not relevant to credit risk given full information:

$$\mathbb{P}(Y|\mathbf{s}, \mathbf{x}_o, \mathbf{x}_u) = \mathbb{P}(Y|\mathbf{x}_o, \mathbf{x}_u).$$

It is important to posit the existence of both observed and unobserved legitimate covariates to capture the often observed relationship between sensitive covariates and the response. Specifically, observed data often show

$$\mathbb{P}(Y|\mathbf{s}, \mathbf{x}_o) \neq \mathbb{P}(Y|\mathbf{x}_o).$$

This lack of conditional independence violates the fairness assumption that sensitive features are uninformative.

Since assumptions like these will need to be presented many times, they will be succinctly captured using DAGs such as Figure 3. Observed data are often only representable by a fully connected graph which contains no conditional independence properties (Figure 3a). This observed distribution can be generated from multiple full-information models. The first possible representation of the full data is an unrestricted model (Figure 3b). In this case, sensitive covariates are not conditionally independent of the response given full information. Such a model states that there are different risk properties between protected groups even after considering *full* information. The fairness assumption is captured in Figure 3c: Y is d-separated from $\mathbf{s}$ given $\mathbf{x}_u$ and $\mathbf{x}_o$. Stated differently, credit risk is conditionally independent of the sensitive covariates given full information. Under this full-information model, the apparent importance of sensitive information in the observed data is only due to unobserved covariates.

Creating impartial estimates is challenging because it requires estimating Y under an assumption that does not hold in the data; however, the unrestricted model in Figure 3b is equivalent to the fair model in Figure 3c if the direct effect of $\mathbf{s}$ on Y is zero. This provides insight into the manner in which the impartial estimate of Y will be constructed: impartiality requires constraining the sensitive covariates to have "no-direct-effect" on the estimates.

For clarity, this will be described using linear regression models of credit risk. While a general functional form may be preferred, understanding impartiality in linear models provides not only a tractable solution but also insight into how to ensure impartiality in the general case. The insight is gained through properly understanding standard effect decompositions. Non-standard, but conceptually identical decompositions also yield novel connections to both legal and philosophical standards for fairness.

Compare the classical full and restricted regression models. The full regression model includes both the sensitive and legitimate covariates as explanatory variables, while the restricted or marginal regression model only includes legitimate covariates as explanatory variables. Coefficients estimated in these models are given subscripts f and r, respectively. In both cases, and for the rest of the models considered in this paper, ϵ_{model} has mean 0. While the notation is similar to that of multiple and simple regression models, respectively, covariates are potentially vector valued. Since this distinction is clear, we still use the terminology of partial and marginal coefficients for the full and restricted models, respectively. Throughout the remainder of the paper, all estimated coefficients will be given "hats" and double subscripts to indicate the covariate and model to which they belong (e.g. $\hat{\beta}_{0,f}$). Since no parameters are ever given for x_u, the parameters for x_o will be written $\beta_{x,model}$ to improve readability.

	Full Regression	Restricted Regression
Model	$Y = \beta_{0,f} + \beta_{s,f}^\top \mathbf{s} + \beta_{x,f}^\top \mathbf{x}_o + \epsilon_f$	$Y = \beta_{0,r} + \beta_{x,r}^\top \mathbf{x}_o + \epsilon_r$

A standard decomposition demonstrates that the estimated marginal coefficient $\hat{\beta}_{x,r}$ can be represented as a function of the estimated partial coefficients $\hat{\beta}_{x,f}$ and $\hat{\beta}_{s,f}$ (Stine and Foster, 2013). This separates the marginal coefficient into direct and indirect effects:

$$\underbrace{\hat{\beta}_{x,r}}_{\text{marginal}} = \underbrace{\hat{\beta}_{x,f}}_{\text{direct}} + \underbrace{\hat{\Lambda}_x \hat{\beta}_{s,f}}_{\text{indirect}}. \tag{5}$$

where $\hat{\Lambda}_x$ is estimated from the intermediate regression

$$\mathbf{s} = \lambda_0 + \Lambda_x^\top \mathbf{x}_o + \boldsymbol{\epsilon}_s.$$

As alluded to previously, $\hat{\beta}_{s,f} = 0$ is ideal. In this case, estimates are unchanged by excluding $\mathbf{s}$ during model fitting. In general, the same effect can be accomplished by fitting the full regression model to compute coefficient estimates, but making predictions that only use $\hat{\beta}_{0,f}$ and $\hat{\beta}_{x,f}$. This removes the *influence* of $\mathbf{s}$ as desired by impartiality. While only legitimate covariates are used, coefficients must be estimated in the full model, else the relationship between sensitive and legitimate covariates allows discriminatory effects to be included in the coefficient of $\mathbf{x}_o$. The estimates are impartial under FEO since we are not addressing the possibility of discrimination in $\mathbf{x}_o$. The reverse regression literature in economics uses these estimates as a preprocessing step (Goldberger, 1984). That literature did not justify this as a fair estimate. We do so here and extend the estimates to more philosophically robust settings.

Definition 4 (Impartial Estimate: Formal Equality of Opportunity). *Using a linear regression model, $\hat{Y}$ is impartial if:*

$$\hat{Y} = \hat{\beta}_{0,f} + \hat{\beta}_{x,f}^\top \mathbf{x}_o,$$

where the coefficients are estimated in the model

$$Y = \beta_{0,f} + \beta_{s,f}^\top \mathbf{s} + \beta_{x,f}^\top \mathbf{x}_o + \epsilon_f.$$

The standard decomposition in equation (5) can be presented in a non-standard way to yield additional insight. Collect the observations into matrices $\mathbf{Y}$, $\mathbf{S}$, and $\mathbf{X}$ and consider writing the *estimated* response from the full regression. By decomposing this expression we can identify components which are of philosophical and legal interest. Separate the sensitive covariates into the component which is orthogonal to the legitimate covariates and that which is correlated with them. We will refer to these as the "unique" and "shared" components, respectively. It is important to note that the coefficient is computed only from the unique component, which follows from writing the multiple regression coefficient as the solution to a simple regression problem (Hastie et al., 2009). This decomposition can be done by considering the projection matrix on the column space of $\mathbf{X}_o$. For a full-rank matrix $\mathbf{M}$, the projection or hat matrix is $\mathbf{H}_\mathbf{M} = \mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'$.

$$\hat{\mathbf{Y}} = \hat{\beta}_{0,f} + \mathbf{S}\hat{\beta}_{s,f} + \mathbf{X}_o\hat{\beta}_{x,f}$$

$$\hat{\mathbf{Y}} = \hat{\beta}_{0,f} + \overbrace{\underbrace{\mathbf{H}_{\mathbf{X}_o}\mathbf{S}}_{di}\hat{\beta}_{s,f} + \underbrace{(\mathbf{I} - \mathbf{H}_{\mathbf{X}_o})\mathbf{S}}_{dt}\hat{\beta}_{s,f}} + \mathbf{X}_o\hat{\beta}_{x,f} \tag{6}$$

The resulting terms are identified in equation (6) as di and dt, to indicate their legal significance. The term dt captures the disparate treatment effect: it is the component of the estimate which is due to the unique variability of $\mathbf{S}$. Given the fair model in Figure 3c, we know the apparent importance of $\mathbf{S}$ (signified by the magnitude of $\hat{\beta}_{s,f}$) is due to excluded covariates; however, it is identified by $\mathbf{S}$ in the observed data. While this may be a "sufficiently accurate generalization," in that the coefficient may be statistically significant for example, this is illegal statistical discrimination. The term "statistical discrimination" is commonly used in social science to refer to these cases where sensitive covariates are used to estimate unobserved but possibly economically relevant characteristics (Blank et al., 2004).

The term di captures the disparate impact effect. We refer to it as the *informative* redlining effect in order to contrast it with an effect identified later. Intuitively, it is the misuse of a legitimately informative variable and is the result of the ability to estimate $\mathbf{S}$ with other covariates. It is an adjustment to the influence of $\mathbf{X}_o$ that accounts for different performance between groups of $\mathbf{S}$. It is important that the adjustment is identified by variability in $\mathbf{S}$ instead of $\mathbf{X}_o$, as seen in equation (5). Identifying a disparate impact effect may be challenging because it is in the space spanned by the legitimate covariate, $\mathbf{X}_o$. The current legal solution merely removes the sensitive features from the analysis which allows for redlining via the term dt.

4.4 Full SEO

One objection to this model is the assumption that all $\mathbf{x}$ covariates are legitimate. Thus, while credit risk may be explained in terms of $\mathbf{x}$ without recourse to $\mathbf{s}$, that is only because the covariates $\mathbf{x}$ are the result of structural discrimination. This critique stems from concerns over SEO: different $\mathbf{s}$ groups may not have the same possibility of being of high merit as measured by $\mathbf{x}$. If this is driven by societal constraints such as a class hierarchy or a cycle of poverty, these covariates may be suspect, and their use could perpetuate the disenfranchisement of historically marginalized groups. Such seemingly legitimate variables which are prejudicially associated with sensitive covariates need to be considered differently, as they are simultaneously potentially informative and discriminatory. This class of "potentially illegitimate" or "suspect" covariates can be used to estimate merit, but only in such a way that does not distinguish between groups in $\mathbf{s}$.

Fairness requires average group differences to be removed because averages differ due to discrimination. Other papers have advocated a regression approach where all variables are considered suspect variables (Calders et al., 2013). Without a proper understanding of the implications of this viewpoint, however, the results are highly unsatisfactory. This will be discussed in detail via example in Section 4.5.

Ideally, one would prefer to estimate the causal effect of the sensitive covariates such that the effect can be removed. A conservative assumption mirrors the fairness assumption for sensitive covariates: suspect covariates are assumed to have no direct effect on the outcome in the full information model. As such, they are merely used as proxy variables for missing information.

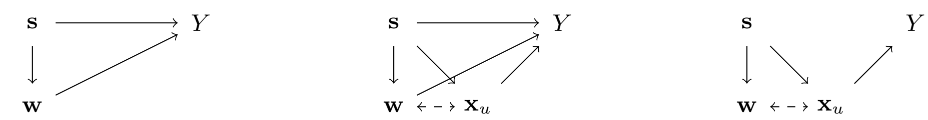
Proxy variables, also known as information carriers, do not directly influence Y given full information. While this is a similar property as $\mathbf{s}$, they are not considered to be protected characteristics. A common example of a suspect covariate or an information carrier is location. Living in a particular location does not *make* someone of higher merit for many applications, but it may be indicative of things that do so. For example, suppose that information on education is missing in the data set. Location can be used as a proxy for education. This, however, may be concerning given that most neighborhoods are racially homogeneous. Impartial estimation in this setting should allow location to be a proxy for a legitimate variable but must *not* use location as a proxy for race or other sensitive covariates. Given the difficulty of precisely separating these two components, conservative estimates can be used to ensure that location is not used for redlining.

For simplicity, our current discussion excludes observed legitimate covariates which are added in Section 4.6. The data are assumed to have a joint distribution $\mathbb{P}(Y, \mathbf{s}, \mathbf{x}_u, \mathbf{w})$, where $\mathbf{w}$ is a vector of suspect covariates, from which n observations are drawn. As before, consider a linear model of credit risk given the sensitive and suspect covariates. The coefficients are given the subscript p for proxy, and ϵ_p has mean 0.

$$\text{Full SEO Model:} \quad Y = \beta_{0,p} + \beta_{s,p}^\top \mathbf{s} + \beta_{w,p}^\top \mathbf{w} + \epsilon_p$$

DAGs similar to those in the previous subsection visually represent the fairness assumption. In the observed data, $\mathbf{s}$ and $\mathbf{w}$ are often associated with Y. As before, this observed data structure can be generated from multiple full data models which include the unobserved, legitimate covariates. An unrestricted data model, Figure 4b, posits no conditional independence between covariates. The fair data model, Figure 4c, respects both constraints on $\mathbf{s}$ and $\mathbf{w}$ implied by the fairness assumption. Furthermore, this captures the intuition of a proxy variable; if location is a proxy for education and education is already in the model, then location will be uninformative.

As before, decompose the estimates from the full SEO model, where $\mathbf{W}$ is the matrix of proxy variables with $\mathbf{w}_i'$ as rows. All explanatory variables are separated into shared and unique components. In both decompositions, there are components identified as disparate treatment and disparate impact. Further



(a) Observed Data (b) Unrestricted Data (c) Fair Data

Fig 4: DAGs Using Suspect or Proxy Variables

information can be gained by considering the decompositions of $\mathbf{W}$ and $\mathbf{X}$ into their constituent parts. There is a unique component, orthogonal to $\mathbf{S}$, as well as components labeled $sd+$ and $sd-$.

$$\text{FEO:} \qquad \hat{Y}_f = \hat{\beta}_{0,f} + \underbrace{\mathbf{H}_{\mathbf{X}_o}\mathbf{S}}_{di}\hat{\beta}_{s,f} + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{X}_o})\mathbf{S}}_{dt}\hat{\beta}_{s,f} \tag{7}$$

$$\underbrace{\mathbf{H}_{\mathbf{S}}\mathbf{X}_o}_{sd+}\hat{\beta}_{x,f} + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{S}})\mathbf{X}_o}_{u}\hat{\beta}_{x,f} \tag{8}$$

$$\text{F-SEO:} \qquad \hat{Y}_p = \hat{\beta}_{0,p} + \underbrace{\mathbf{H}_{\mathbf{W}}\mathbf{S}}_{di}\hat{\beta}_{s,p} + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{W}})\mathbf{S}}_{dt}\hat{\beta}_{s,p} \tag{9}$$

$$\underbrace{\mathbf{H}_{\mathbf{S}}\mathbf{W}}_{sd-}\hat{\beta}_{w,p} + \underbrace{(\mathbf{I}-\mathbf{H}_{\mathbf{S}})\mathbf{W}}_{u}\hat{\beta}_{w,p} \tag{10}$$

In equations (7) and (8), previous discussions of redlining do not distinguish between the terms di and $sd+$ (Kamishima et al., 2012; Kamiran et al., 2013) because they are both due to the correlation between $\mathbf{X}_o$ and $\mathbf{S}$. It is clear that they are different, as $\mathbf{H}_{\mathbf{X}_o}\mathbf{S}$ is in the space spanned by $\mathbf{X}_o$ and $\mathbf{H}_{\mathbf{S}}\mathbf{X}_o$ is in the space spanned by $\mathbf{S}$. Furthermore, the coefficients attached to these terms are estimated from different sources. Intuition may suggest we remove all components in the space spanned by $\mathbf{S}$, but this is often incorrect. The term $sd+$ can be included in many models because it accounts for the group means of $\mathbf{X}$. Excluding $sd+$ implies that the *level* of $\mathbf{X}$ is not important but that an individual's *deviation* from their group mean is. This makes group membership a hindrance or advantage and is inappropriate for a legitimate covariate. Therefore, $sd+$ should be included if $\mathbf{x}$ is legitimate.

In the decomposition of the full SEO model, equation (10), $sd-$ addresses the concern that the group means are potentially unfair or could be used to discriminate. For example, if $\mathbf{w}$ is location, $sd-$ may measure racial differences between neighborhoods. Given that proxy variables $\mathbf{w}$ are not considered directly informative, it is unclear what these differences can legitimately contribute. If there is racial bias in neighborhood demographics, using this information would perpetuate this discrimination. Ensuring that this does not occur requires removing $sd-$ from the estimates of $\mathbf{Y}$. This identifies a new type of redlining effect that we call *uninformative* redlining; it is the sum of di and $sd-$. Uninformative redlining can be identified visually using the graphs in Figure 4. Fairness constrains the information contained in the arrow $\mathbf{s} \to Y$ as well as information conveyed in the path $\mathbf{s} \to \mathbf{w} \to Y$. This is because $\mathbf{s} \to \mathbf{w}$ is potentially discriminatory. Therefore, fair estimates with suspect or proxy variables only use the unique variability in $\mathbf{W}$. An important consequence of the F-SEO estimate is that average estimates are the same for different groups of $\mathbf{s}$. This is an alternate construction of the initial estimates used by Calders et al. (2013).

Definition 5 (Impartial Estimate: Full Substantive Equality of Opportunity). *Using a linear regression model, $\hat{Y}$ is impartial if:*

$$\hat{Y} = \hat{\beta}_{0,p} + \hat{\beta}_{w,p}^{\top}(\mathbf{w} - \hat{\Lambda}_{\mathbf{s}}^{\top}\mathbf{s}),$$

where $\hat{\beta}_{0,p}$ and $\hat{\beta}_{w,p}$ are estimated in the model

$$Y = \beta_{0,p} + \beta_{s,p}^{\top}\mathbf{s} + \beta_{w,p}^{\top}\mathbf{w},$$

and $\hat{\Lambda}_{\mathbf{s}}$ is the matrix of coefficients estimated in the model

$$\mathbf{w} = \lambda_{\mathbf{0}} + \Lambda_{\mathbf{s}}^{\top}\mathbf{s} + \boldsymbol{\epsilon}_w.$$

The estimates in Definition 5 can also be constructed by first projecting $\mathbf{w}$ off of $\mathbf{s}$ to create the unique component found in equation (10), and then treating the result as a legitimate covariate in the FEO framework.

4.5 Simple Example: FEO vs SEO

This section provides a simplified example to compare the estimates implied by FEO and SEO. Comparing estimates on an individual level has been overlooked in the literature, which favors providing a mathematical statement of discrimination for the set of predictions and demonstrating that the measure has been satisfied. It is important to understand what the estimates themselves look like. Claims about fairness are often made my individuals: the applicant in our loan example wants to be treated fairly. We demonstrate how ignoring this is such a large oversight of previous works. Without a proper generative story, "fair" estimates can appear decidedly *unfair*.

Consider a simple example with only two covariates: education level, x, and sensitive group, s. Covariate names are no longer in bold because they are not vectors. Suppose the data is collected on individuals who took out a loan of a given size. In this case, suppose higher education is indicative of better repayment. As an additional simplification, suppose that education is split into only two categories: high and low. Lastly, to see the relevant issues, s and x need to be associated. The two sensitive groups will be written as s_+ and s_-, merely to indicate which group, on average, has higher education. As such, the majority of s_- have low education and the majority of s_+ have high education. The response is the indicator of default, D.

The data and estimates are provided in Table 1, in which there exist direct effects for both s and x. This is consistent with the observed data DAGs in previous sections. While the framework presented in this paper is equally applicable to logistic regression and generalized linear models, this data example is simple enough that linear regression produces accurate conditional probability estimates. Therefore, different estimates can be directly analyzed for fairness.

Five possible estimates are compared in Table 1: the full OLS model, the restricted regression which excludes s, the FEO model in which education is considered a legitimate covariate, the SEO model in which education is considered

TABLE 1
Simplified Loan Repayment data.

Income ($\mathbb{P}(x)$)	Low (.6)		High (.4)			
Group ($\mathbb{P}(s\|x)$)	s_- (.75)	s_+	s_- (.25)	s_+	Total	
Default Yes	225	60	20	30	335	
Default No	225	90	80	270	665	

	$\hat{\mathbb{P}}(D_i = 1\|x_i, s_i)$				DS	RMSE
Full Model	.5	.4	.2	.1	-.25	13.84
Exclude **s**	.475	.475	.125	.125	-.17	13.91
FEO	.455	.455	.155	.155	-.15	13.93
SEO	.39	.535	.09	.235	0.00	14.37
Marginal	.35	.35	.35	.35	0.00	14.93

a suspect covariate, and the marginal model which estimates the marginal probability of default without any covariates. Estimates are presented along with the RMSE from estimating the true default indicator and the discrimination score (DS) (Calders and Verwer, 2010). While we have argued that "discrimination score" is a misnomer since it does not separate explainable from discriminatory variation, it provides a useful perspective given its widespread use in the literature.

Since education is the only covariate that can measure similarity, the colloquial notion of fairness dictates that estimates should be constant for individuals with the same education. This is easily accomplished by the legal prescription of excluding s. If the information is not observed, it cannot lead to disparate treatment directly related to group membership. The FEO model satisfies this as well. As seen in the standard decomposition in equation (5) the only difference between the two estimates is the coefficient on x. Said differently, the term di in equation (6) lies in the space spanned by x. Therefore its removal only changes estimates for education groups. Excluding s permits redlining because it increases the estimated disparity between low- and high-education groups. This disproportionately effects those in s_- as they constitute the majority of the low education group. The FEO estimates result in some average differences between groups, but this is acceptable if the association between x and s is benign. This accurately measures the proportional differences desired by Banks (2001) for fair treatment.

The SEO estimates appear counter-intuitive: although s_+ performs better in our data set even after accounting for education, these "fair" estimates predict the opposite. Understanding this requires accepting the world view implicit in the SEO estimates: average education differences between groups are the result of structural differences in society. Members of s_- in the high education group have a much higher education than average for s_-. Similarly, members of s_+ who are in the high education group have a higher education than average for s_+, but not by as much. The magnitude of these differences is given importance, not the education level.

Broadly speaking, this type of correction for group differences occurs in university admissions processes when students are compared regionally. Consider the explanatory covariate to be "SAT score," which is a standardized test score commonly used in the United States to assess students' readiness for university. The sensitive covariate is the state from which the student is applying. Average

SAT scores differ drastically between states, with the spread between the best and worst performing states being over 17% (Zhang, 2016). This may be due, in-part, to factors such as differences in education budgets, emphasis on high school education, and focus on test preparation. Regional comparisons remove some of these structural differences.

In our example, the SEO model balances the differences in education distributions, resulting in both groups having the same average estimated default. This is seen in the discrimination score of 0. Without claiming that education is partially the result of structural differences, the SEO estimates appear to discriminate against s_+. All estimation methods previously considered in the literature produce estimates relevantly similar to SEO in this regard. This was not acknowledged because a direct comparison of the change in individual estimates was not provided. Furthermore, if not all s_+ individuals are given a benefit or not all s_- individuals are given a detriment, then these models are merely approximations of the fair correction. An ideal protected or sensitive covariate s is exactly that which accounts for differences in the opportunity of being high merit. This is more in line with Rawl's conception of equality of opportunity (Rawls, 2001), but this line of inquiry is beyond the scope of this paper.

The SEO estimates show another important property: their RMSE is lower than that of the marginal estimate of default while still minimizing the discrimination score. As such, if a bank is required to minimize differences between groups in the interest of fairness, it would rather use the SEO estimates than the marginal estimate. The difference between the two estimates is that SEO still acknowledges that education is an informative predictor and contains an education effect. Furthermore, equality of opportunity is not satisfied when marginal estimates are used because all merit information is ignored. See Arneson (2015) for a more detailed discussion.

4.6 Total Model

After analyzing the covariate groups in isolation, we now consider models with sensitive, legitimate, and suspect variables. The data are assumed to have a joint distribution $\mathbb{P}(Y, \mathbf{s}, \mathbf{x}_o, \mathbf{x}_u, \mathbf{w})$, from which n observations are drawn. The relevant DAGs are given in Figure 5. As before, the observed data and unrestricted full data contain no conditional independence relationships. The fair data model captures the fairness assumption. Again consider a simple linear model for clarity. Coefficients are given the subscript t and ϵ_t has mean 0.

$$\text{Total Model:} \quad Y = \beta_{0,t} + \beta_{s,t}^{\top}\mathbf{s} + \beta_{x,t}^{\top}\mathbf{x}_o + \beta_{w,t}^{\top}\mathbf{w} + \epsilon_t \tag{11}$$

The now familiar decomposition into unique and shared components is more complex because the shared components exist across multiple dimensions. Equation (12) separates each term into its unique component and the component which

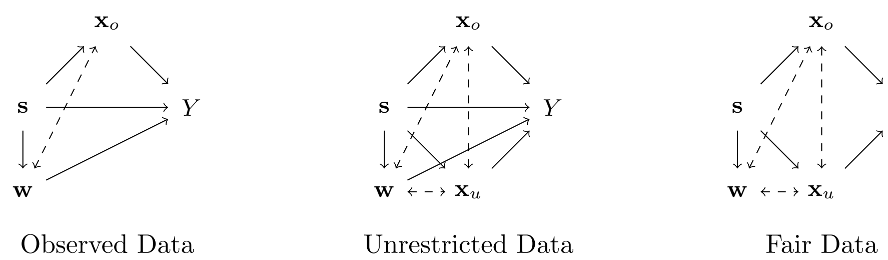


Fig 5: DAGs of Total Model

is correlated with the other variables.

$$\begin{aligned}\hat{Y} = \hat{\beta}_{0,t} &+ \underbrace{\mathbf{H}_{[\mathbf{X}_o,\mathbf{W}]}\mathbf{S}}_{di}\hat{\beta}_{s,t} + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{X}_o,\mathbf{W}]})\mathbf{S}}_{dt}\hat{\beta}_{s,t} \\ &+ \underbrace{\mathbf{H}_{[\mathbf{S},\mathbf{W}]}\mathbf{X}_o}_{sd+}\hat{\beta}_{x,t} + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{S},\mathbf{W}]})\mathbf{X}_o}_{u}\hat{\beta}_{x,t} \\ &+ \underbrace{\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]}\mathbf{W}}_{sd^-,sd^+}\hat{\beta}_{w,t} + \underbrace{(\mathbf{I}-\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]})\mathbf{W}}_{u}\hat{\beta}_{w,t} \end{aligned} \tag{12}$$

The sensitive covariates are again separated into disparate impact and disparate treatment components. Similarly, the legitimate covariates are separated into permissible statistical discrimination and a unique component. The suspect covariates, however, display different behavior in the total model. The unique component has the same interpretation, but the component correlated with other covariates is labeled both $sd+$ and $sd-$ to indicate that this combines both legal and illegal forms of statistical discrimination. The notation, $\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]}\mathbf{W}$, indicates that the shared component is the best linear estimate of $\mathbf{W}$ given both $\mathbf{S}$ and $\mathbf{X}$. As $\mathbf{w}$ is a suspect variable, we need an impartial estimate of it. This is the FEO model for impartial estimates as there are only sensitive and legitimate covariates. In this case, $\mathbf{w}$ has taken the place of Y as the response.

Definition 6 (Impartial Estimate: Total Model). *Impartial estimates are created with the following multi-step procedure:*

1. *Estimate the total model (11) to produce* $\hat{\beta}_{0,t}$, $\hat{\beta}_{s,t}$, $\hat{\beta}_{x,t}$, *and* $\hat{\beta}_{w,t}$.
2. *Create an impartial estimate of each element of* $\mathbf{w}$ *per Definition 4, and collect the impartial estimates as* $\hat{\mathbf{W}}$.
3. *Set* $\hat{Y} = \hat{\beta}_{0,t} + \mathbf{X}_o\hat{\beta}_{x,t} + \hat{\mathbf{W}}\hat{\beta}_{w,t} + (\mathbf{I}-\mathbf{H}_{[\mathbf{X}_o,\mathbf{S}]})\mathbf{W}\hat{\beta}_{w,t}$.

The conditions given for impartial estimates in Definition 2 merely specify the properties of the residuals after creating estimates using Definition 6 under the assumption that the full regression model holds. A more direct construction of these estimates first pre-processes the suspect covariates $\mathbf{w}$ by projecting them off of the sensitive covariates $\mathbf{s}$. These adjusted covariates can then be treated as legitimate variables because they are orthogonal to $\mathbf{s}$. As noted previously, this does not change the value of $\hat{\beta}_{w,t}$.

4.7 Applications of Impartial Estimates

Now that impartial estimates have been provided, a simple comparison can be made to identify the cost of disparate impact and governmental policies. The cost of fairness in these cases can be quantified by considering the decisions that would be made by different actors. Suppose that a privately owned, profit-maximizing bank is providing loans. As such, the best estimates are those which most accurately predict the riskiness of a loan while operating under Title VII of the Civil Rights Act. Contrast this with a government-owned company that is constrained to be impartial. This bank can constrain profit to achieve fairness goals. Conceptually, the cost of the government policy is the difference in the expected profit between the estimation methods these two banks would use.

The FEO models provide the minimally constrained impartial estimates. A private bank may argue in favor of this method even if some covariates are prejudicial, because many covariates are generated outside the scope of the bank's operation. For example, discrimination in education may not be within the power of the bank to change. The government-owned bank, however, may want to use a partially or fully SEO model due to discrimination in the generation of some covariates. The state-owned bank's estimates are from the corresponding SEO model. Using these estimates of risk, a bank would provide different loans at different rates.

One could consider that loans are provided at rates determined by the state-owned bank's estimates but whose cost to the bank is computed using the privately-owned bank's estimates. The bank expects this to result in lower profit given its estimates of risk. The difference in expected profit between this loan scheme and the desired, profit-maximizing scheme is an estimate of the cost of the government program. Said differently, it is the price that the government could pay the bank for accomplishing the government's goals. Examples of the different estimates can be seen in our simplified data example in Table 1.

The cost of disparate impact can be estimated similarly. Instead of comparing the ideal private bank and the state-owned bank, the comparison is between the rates the bank actually provided and what the ideal private bank *would* have provided given the same data. Therefore, instead of comparing the SEO and FEO models, one would compare the bank's actual lending history to what would be implied by the required impartial estimates. For example, suppose the bank followed the current minimal legal prescription of excluding the sensitive covariates as opposed to using impartial estimates. The difference between the actual and impartial estimates are again seen in our simplified data example in Table 1. The additional revenue the bank received due to their discriminatory lending practices can be estimated as the difference in the two estimation methods. Elaboration of these applications with real data is a subject of current research.

5. CORRECTING ESTIMATES

Suppose that we have estimates of credit risk, $Y^\dagger$, given by an unknown model with unknown inputs. The model may use sensitive information to be intentionally or unintentionally discriminatory. This is a challenging but necessary case to consider as most models used by private companies are proprietary. Therefore, we must be completely agnostic as to the construction of these black-box estimates. It is perhaps surprising that these estimates can easily be made impartial.

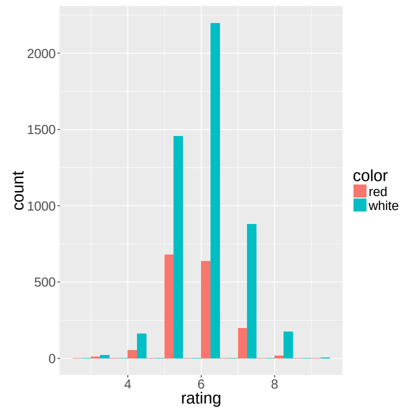


Fig 6: Histogram of wine data rankings.

Consider $Y^\dagger$ to be an additional explanatory covariate or covariates and include them in the modeling similar to a stacked regression (Breiman, 1996):

$$Y = \beta_{0,c} + \beta_{s,c}^\top \mathbf{s} + \beta_{x,c}^\top \mathbf{x}_o + \beta_{w,c}^\top \mathbf{w} + \beta_{Y^\dagger}^\top Y^\dagger + \epsilon_c.$$

These black-box estimates are potentially predictive, but there is no guarantee that they are impartial. This identically matches the description of suspect covariates. If we treat $Y^\dagger$ as a suspect covariate, its information can be used but not in a way that makes distinctions between protected groups. This allows us to easily correct black-box estimates.

A numeric example will solidify this idea and demonstrate the efficacy of our methods. Our data contains rankings of wine quality and is taken from the UCI Machine Learning Repository (Lichman, 2013). There are ratings for both red and white wines and a reasonable request is that ranking be impartial between the two groups. This data set captures all of the relevant issues that arise in FADM and was also considered in Calders et al. (2013). It also demonstrates the similarity between FADM and controlling for batch effects if data is aggregated from multiple sources.

The response is the rating of wine quality measured between 0 and 10, the sensitive feature is wine type (red or white), and there are 10 explanatory variables such as acidity, sugar, and pH. Of the approximately 6,500 ratings, about 25% are for red wines. The distributions of wine ratings are very similar, as seen in Figure 6.

Calders and Verwer (2010) and Calders et al. (2013) claim that the appropriate measure of unfairness in this case is given by the average rating difference between the two groups. This paper has argued that explainable differences need to be taken into account. This is evidence for treating the multiple regression coefficient as an estimate of prejudice. We do not advocate doing so for reasons made clear shortly, but the difference in measures is worth noting. The two measures can disagree substantially, not only in magnitude but even in sign. The average difference in wine rating is .24, with white wine being preferred; however, a multiple regression analysis indicates that white wine is, on average, rated .14 *lower* than red wines, ceteris paribus. The difference in sign is an example of Simpson's Paradox and is more generally related to the problem of suppression (Johnson et al., 2015).

Estimating the extent of discrimination is a far more challenging problem than providing fair estimates. It is akin to asking for an estimate of the average treatment effect, where the treatment is wine type. It requires a set of causal assumptions to be made about the comparability of the two groups. In the language of causal inference, there may be insufficient covariate overlap between the two groups given the required differences between the two types of wine. This prevents, or increases the difficulty in, accurately estimating a measure of discrimination. Interested readers are directed toward Pearl (2009); Rosenbaum (2010) for introductions to various perspectives on causal inference.

Our validation framework follows that of Calders et al. (2013). Models will be trained on intentionally biased data and tested on the real data. Note that this is not exactly the desired measure; ideally one should test on impartial data, but this presupposes an answer to the problem at hand. By biasing the training data, the test data is at least more impartial. Data is biased by randomly selecting 70% of the white wines and increasing their rating by 1. This results in group mean differences (white-red) of .94 on the biased data and .24 on the raw data.

This bias is picked up by the multiple regression coefficient on wine type (WHITE). Averaged over 1,000 biased samples, the multiple regression coefficient for wine type increases by .7 on the biased data while all other coefficients change by a negligible amount. Therefore, our impartial estimates are exactly the same if produced using the biased data or raw data. This is precisely the type of result necessary for impartial algorithms: the multiple regression model is unaffected by the bias added to the data. Admittedly, this is trivial since the bias is randomly attributed to 70% of white wines. That being said, this is heretofore unrecognized in the literature and not acknowledged in Calders et al. (2013).

The hope is that the impartial estimates will improve performance on the test sample even when they are trained on the biased data. We compare our models with those of Calders et al. (2013), which are indicated by "Calders" in Tables 2 and 3. They attempt to capture explainable variability through the use of propensity score stratification. This method estimates the probability that an observation is a red or white wine. Observations are then binned into 5 groups based on this propensity score. In each group, a full SEO model is fit where all explanatory variables are considered discriminatory. Therefore, within each strata, there is no average difference between groups. There is a large problem with this perspective since there is often enough information to predict the sensitive attribute almost perfectly. This results in largely homogeneous strata and the method fails. Furthermore merely using 5 bins does not provide sufficient covariate overlap for the guarantees surrounding propensity score matching to hold, even approximately. Other methods such as those of Kamiran et al. (2013) only handle a single sensitive and single legitimate covariate. Our estimates satisfy all of their fairness requirements in a more general and flexible framework.

The results we present are averaged over 20 simulations where out-of-sample error is estimated using 5-fold cross-validation. Variability in the performance estimates is negligible. Table 2 shows the linear regression results on this data. As a baseline measure, we use ordinary least squares (OLS), which is merely the estimates from the full regression model. This contains the disparate treatment and disparate impact effects. The FEO model treats all non-sensitive explanatory variables as legitimate, while the SEO model treats all non-sensitive covariates as proxy variables. There is a spectrum of SEO models that consider different sets as fair or discriminatory. We chose to consider the full SEO model so as to permit easy comparison to the Calders estimates.

Four performance measures are given using the out-of-sample estimates: RMSE measured on the biased data (out-of-sample), RMSE on the raw data, the mean difference in estimates between groups (DS), and our impartiality score (IS). Our models have an impartiality score of 0 in-sample by construction, and Tables 2 and 3 demonstrate that this is not noticeably worsened out-of-sample. Note that the impartiality score is measured with respect to the desired type of impartiality (FEO or SEO). For the Calders estimates, SEO impartiality was used as it is conceptually closest to their procedure. If FEO impartiality is used instead, the IS is roughly half the value presented in the tables.

TABLE 2
Regression Performance

	OLS	Formal EO	Sub. EO	Calders
RMSE-biased	0.84	0.85	0.93	0.86
RMSE-raw	0.95	0.92	0.91	0.92
DS	0.94	0.60	0.00	0.79
IS	0.04	0.00	0.00	0.12

The general direction of performance results are as expected: requiring estimates to be impartial worsens performance on the biased data but improves performance on the test data. Even using linear regression models, we provide more accurate estimates of the test data while minimizing the mean difference between groups. This significantly improves upon the results of the Calders estimates.

From a different perspective, the results in Table 2 may not be particularly impressive in that wine ratings are not well estimated; merely using the average rating in the biased data to estimate the raw data yields an RMSE of 1.02. This is to be expected, in part because wine quality is more complex than these 10 explanatory covariates, but also because rating is most likely not a linear function of the explanatory covariates. Therefore, consider a more complex, black-box estimate of rating that can account for these nonlinearities. We will use a random forest (Breiman, 2001) as a canonical black-box, as it is an off-the-shelf method which performs well in a variety of scenarios but has largely unknown complexity. It is challenging to consider how $\mathbf{s}$ should be used constructively in a random forest algorithm while ensuring impartiality. Furthermore, we allow the random forest model to use $\mathbf{s}$ in order to demonstrate that estimates can be easily corrected even when algorithms are trained on sensitive attributes.

Table 3 contains the results of the random forest models. The FEO model treats the random forest estimates as potentially discriminatory but the others as legitimate, while the SEO model considers all variables as potentially discriminatory. These are again compared to the base-line random forest estimates and the Calders estimates. The general trend is the same as before: constraining estimates to be impartial improves performance on the test sample but worsens performance on the biased sample. Using a random forest significantly reduced RMSE while not worsening the fairness measures. The corrected random forest estimates significantly outperform the Calders estimates along all measures. It does so even while exactly satisfying their desired constraint of zero mean difference between groups.

TABLE 3
Corrected Random Forest Performance

	RF	Formal EO RF	Sub. EO RF	Calders
RMSE-biased	0.73	0.74	0.83	0.86
RMSE-raw	0.87	0.82	0.81	0.92
DS	0.93	0.62	0.00	0.79
IS	0.03	0.00	0.00	0.12

6. DISCUSSION

Our work has several important implications and contributions: we can construct impartial estimates, ensure black-box estimates are impartial, and quantify regulation and disparate impact. All of this is done through providing a clear statistical theory of impartial estimation and explainable variability. While we discuss the classical method of multiple regression, its power and interpretability were heretofore not well-understood in the FADM literature. Future work moves away from assumptions of linearity. Such methods would use a different similarity metric to measure differences between people as well as a different functional form of the response.

There are many open questions for FADM even in the linear setting, however. For example, how should impartial inference be conducted? If estimates produced by a proprietary model are close but not identical to our methods, can we test their deviation from impartiality? This is crucially important in legal cases and is closely related to inference under model-misspecification (Buja et al., 2014). A second major challenge is to relax the assumption of known covariate groups. A method to interpolate between the categories may remove some of the inherent difficulties in the classification.

More generally, this paper provides a framework through which fairness and impartial estimation can be understood broadly. There are two key components that are required to generalize this approach to other methods. First, we identified the direct effect of $\mathbf{x}_o$ by estimating the regression model using all available information. This is required for the coefficient estimates in both the formal and substantive equality of opportunity models. Second, variables were residualized to remove the direct effect of $\mathbf{s}$ from $\mathbf{w}$. This was necessary for the substantive equality of opportunity models and to correct black-box estimates. Both generalized linear models and generalized additive models can perform these tasks. Therefore, the concepts outlined here are applicable in those domains. A more thorough investigation of these settings is currently being conducted.

7. APPENDIX

Our definition of impartial estimates follows easily from their construction in Section 4. The derivation of suitable conditions is only presented for the total model case as the others follow as special cases. As such, model subscripts are removed from the coefficient notation. Furthermore, we drop the notation for observed and unobserved legitimate covariates as only observed covariates are present in the estimated functions. Consider the population regression of Y on $\mathbf{s}$, $\mathbf{x}$, and $\mathbf{w}$:

$$Y = \beta_0 + \beta_{\mathbf{s}}^\top \mathbf{s} + \beta_x^\top \mathbf{x} + \beta_w^\top \mathbf{w} + \epsilon.$$

Regardless of the true functional form of Y, the residual ϵ always satisfies $\mathbb{E}[u] = 0$, $\mathrm{Cov}(\mathbf{s}, \epsilon) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}, \epsilon) = \mathbf{0}$, and $\mathrm{Cov}(\mathbf{w}, \epsilon) = \mathbf{0}$ (where in each case $\mathbf{0}$ is a vector of zeros of the appropriate dimension). The conditions in Definition 2 specify the properties of the residual when $\hat{Y}$ is an impartial estimate constructed via Definition 6.

Our estimate is given as

$$\hat{Y} = \hat{\beta}_0^\top + \hat{\beta}_{\mathbf{x}}^\top \mathbf{x} + \hat{\beta}_{\mathbf{w}}^\top (\mathbf{w} - \hat{\Lambda}_{\mathbf{s}}^\top \mathbf{s}),$$

where $\hat{\Lambda}_{\mathbf{s}}$ is the matrix of coefficients from estimating

$$\mathbf{w} = \lambda_{\mathbf{0}} + \Lambda_{\mathbf{s}}^\top \mathbf{s} + \Lambda_{\mathbf{x}}^\top \mathbf{x} + \boldsymbol{\epsilon}_w.$$

Therefore, the impartial residual is

$$u = (\beta_{\mathbf{s}} + \Lambda_{\mathbf{s}} \beta_{\mathbf{w}})^\top \mathbf{s} + \epsilon.$$

The first three properties of Definition 2 follow from simple computations after noting that

$$\mathrm{Cov}(u, \mathbf{s}) = (\beta_{\mathbf{s}} + \Lambda_{\mathbf{s}} \beta_{\mathbf{s}})^\top \mathrm{Var}(\mathbf{s}) \Rightarrow (\beta_{\mathbf{s}} + \Lambda_{\mathbf{s}} \beta_{\mathbf{s}})^\top = \mathrm{Cov}(u, \mathbf{s}) \mathrm{Var}(\mathbf{s})^{-1},$$

assuming the inverse exists.

The last property, $\mathrm{Cor}(\eta, \mathbf{s}) = \mathbf{0}$ where $\eta = \hat{Y} - L(\hat{Y}|\mathbf{x})$, is definitional, and specifies the difference between legitimate and suspect covariates. Since suspect covariates are projected off of $\mathbf{s}$ initially, the only relationship between $\hat{Y}$ and $\mathbf{s}$ must be due solely to linear functions of $\mathbf{x}$. This condition by itself is insufficient to define impartial estimates because of the possibility of redlining.

REFERENCES

Arneson, R. (2015). Equality of Opportunity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Arrow, K. J. (1973). *The Theory of Discrimination*, pages 3–33. Princeton University Press.

Balkin, J. M. and Siegel, R. B. (2003). The American civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1).

Banks, R. R. (2001). Race-based suspect selection and colorblind equal protection doctrine and discourse. *UCLA Law Review*, 48.

Blank, R. M., Dabady, M., and Citro, C. F., editors (2004). *Measuring Racial Discrimination*. National Research Council.

Breiman, L. (1996). Stacked Regressions. *Machine Learning*, 24:49–64.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brighouse, H. and Swift, A. (2009). Legitimate parental partiality. *Philosophy & Public Affairs*, 37(1):43–80.

Brockner, J. (2006). It's so hard to be fair. *Harvard business review*, 84(3):122.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and Zhao, L. (2014). Models as Approximations: How Random Predictors and Model Violations Invalidate Classical Inference in Regression. *ArXiv e-prints*.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling Attribute Effect in Linear Regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80.

Calders, T. and Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Coate, S. and Loury, G. C. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, 83(5):1220–1240.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2011). Fairness Through Awareness. *CoRR*, abs/1104.3913.

Fan, H. and Moro, A. (2011). Theories of Statistical Discrimination and Affirmative Action: A Survey. volume 1A, pages 133–200. Elsevier B.V.

FHLBB (1935). Redline Map for Birmingham, Alabama, ca. 1935. In *Residential Secruity Maps 1933-1939*. NARA II RG 195.

Fiss, O. M. (1976). Groups and the Equal Protection Clause. *Philosophy & Public Affairs*, 5(2):107–177.

Foster, D. P. and Vohra, R. V. (1992). An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188.

Goldberger, A. S. (1984). Reverse Regression and Salary Discrimination. *The Journal of Human Resources*, 19(3):pp. 293–318.

Gosepath, S. (2011). Equality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition.

Hajian, S. and Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Trans. Knowl. Data Eng.*, 25(7):1445–1459.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Holland, P. W. (2003). Causation and Race. *ETS Research Report Series*, 2003(1):i–21.

House, W. (2014). Big Data - Seizing Opportunities, Preserving Values. Technical report.

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015). Submodularity in Statistics: Comparing the Success of Model Selection Methods. *ArXiv e-prints*. http://arxiv.org/abs/1510.06301.

Jollimore, T. (2014). Impartiality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. (2016). Fairness in Learning: Classic and Contextual Bandits.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination Aware Decision Tree Learning. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 869–874, Washington, DC, USA. IEEE Computer Society.

Kamiran, F., Zliobaite, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In Flach, P. A., Bie, T. D., and Cristianini, N., editors, *ECML/PKDD (2)*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer.

Lichman, M. (2013). UCI Machine Learning Repository.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In Li, Y. B. L. and Sarawagi, S., editors, *KDD*, pages 560–568. ACM.

Phelps, E. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review*, 62(4):659–611.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Risse, M. and Zeckhauser, R. (2004). Racial Profiling. *Philosophy & Public Affairs*, 32(2):131–170.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer-Verlag New York.

Sahlberg, P. and Hargreaves, A. (2011). *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* Teachers College Press.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Stine, R. and Foster, D. (2013). *Statistics for Business: Decision Making and Analysis*. Pearson Education Limited.

Wenar, L. (2013). John Rawls. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2013 edition.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning Fair Representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org.

Zhang, F. (2016). Average SAT & ACT Scores by State (Participation Adjusted). http://blog.prepscholar.com/average-sat-and-act-scores-by-stated-adjusted-for-participation-rate. Accessed: 2016-09-29.