Fairness of linear regression in decision making

Vincent Cohen-Addad, Surya Teja Gavva, C S Karthik, Claire Mathieu, Namrata Namrata

Fairness of Linear Regression in Decision Making

Vincent Cohen-Addad
Google Research
vcohenad@gmail.com

Surya Teja Gavva
Rutgers University
suryateja@math.rutgers.edu

Karthik C. S.*
Rutgers University
karthik.cs@rutgers.edu

Claire Mathieu†
CNRS, Université de Paris, IRIF
claire.mathieu@irif.fr

Namrata
University of Warwick
namrata@warwick.ac.uk

Abstract

Ranking systems conceived on historical data are central to our societies. Given a set of applicants and the information as to whether a past-applicant should have been selected or not, the task of fairly ranking the applicants (either by humans or by computers) is critical to the success of any institution. These tasks are typically carried out using regression methods and considering the impact of these selection processes on our lives, it is natural to expect various fairness guarantees. In this article, we assume that affirmative action is enforced and that the number of candidates to admit from each protected group is predetermined. We demonstrate that even with this safety-net, classical linear regression methods may increase discrimination in the selection process, reinforcing implicit biases against minorities, in particular by poorly ranking the top minority applicants. We show that this phenomenon is intrinsic to linear regression methods and may happen even if the sensitive attribute is explicitly part of the input, or if a linear regression is computed on each each minority group individually. We show that to better rank applicants it might be needed to adapt the choice of the regression methods (linear, polynomial, etc.) to each minority group individually.

*This work was supported by a grant from the Simons Foundation, Grant Number 825876, Awardee Thu D. Nguyen, the Israel Science Foundation (grant number 552/16), the Len Blavatnik and the Blavatnik Family foundation, and Subhash Khots Simons Investigator Award.

†This work was partially funded by the grant ANR-19-CE48-0016 from the French National Research Agency (ANR).

1 Introduction

Fair and efficient ranking of individuals based on their achievements is the lifeblood of an efficient company, equitable school system, and more generally of a functioning democracy. Thanks to their simplicity, explainability, and efficiency, regression methods are at the heart of both human- and computer-based selection and ranking processes. For example, selecting candidates for a college [1–8], for a job [9–13], or for a loan [14–20], is often done (at least as an initial filtering process) by computing a ranking of all the candidates based on a combination of the attributes of the candidates, so as to give a better score to the candidates who are the most likely to be successful. In the above contexts, the application of a candidate is summarized by a set of *features* representing, for example, the academic achievements, the work experience, or the financial record of the candidate. For both human- and computer-based decisions, the success prediction is performed according to historical data: given the performance of previously selected applicants, the predictor may give more weight to more informative features.

An iconic example is the use of weighted average of school grades to measure academic achievement, so as to select students at various stages of their curriculum. Thus, for selecting students to enroll in a biology program, the predictor should give more weight to scientific subjects [21].

One advantage of such a method over more advanced machine learning methods is from the viewpoint of transparency, that is, it can be easily explained, computed, and the rules can be published. Here we ask: Is such a method *fair*?

1.1 State-of-the-Art

The above question has motivated research since the 60s and the seminal work of Cleary on graduation tests [22], which states that a fair regression line should fit all minorities similarly (see also [23], and the work of Guion [24]). In the 70s, Thorndike [25] objected that the above criterion is neither sufficient nor necessary to achieve fairness, since different groups may have different success rates and require different regression lines. Instead, he proposes that a fair application of linear regression should imply a fair treatment of the students, in the sense that the ratio of the true positive plus false positive to the true positive plus false negative is the same for all student groups (see also the more recent work by Kleinberg et al. [26]). Later, Darlington [27] showed that the above two fairness criteria are essentially incompatible. These results inspired various new works from a variety of research areas ranging from psychology to philosophy, with the goal of designing quantifiable fairness criteria, see [28–36]. One important outcome of this early research was the distinction between the intrinsic fairness of the test (whether the test is fair to all the participants), and the usage of the test as a fair predictor for ranking and selecting students. In this paper, we focus on the latter.

The above research sparked a broad societal debate on whether school tests, employment criteria, or other large scale *one-size-fits-all* evaluation framework were perpetuating racial discrimination, and how to address it (see [37]). To that end, the United States Employment Services started a strategy called *race-norming*, to balance the scores of minority races [38], but it was finally banned because of accusations of *reverse discrimination* [39, 40].

The increase of automated decision processes in our lives has spurred various studies on

the impact of machine learning methods on latent social discrimination and required a new level of awareness on these problems (see the survey of [29]). Recent work has exhibited this phenomenon at large scale in various settings including justice decisions (through e.g.: the *ProPublica* scandal [41, 42]; see also [43–45]). A tempting way to address this issue is to *hide* the sensitive attributes (such as race, gender, sexual orientation, etc.) from the predictor (e.g., a regression method). However, various works (see e.g.: [46, Chapter 9] and references therein) have shown that the sensitive information is often implicitly encoded in the other features that can thus serve as a proxy to determine the sensitive attribute. In such cases, the predictor can recover the sensitive attribute (intentionally or not) and in their prediction increase the discrimination against disprivileged groups. Thus, the recommendation has become to explicitly take into account the sensitive attributes, but to design selection methods that use the sensitive attributes only to obtain more accurate predictions or to achieve affirmative action, the reader may refer to [47, 48].

1.2 Our Results

First, we consider a simple theoretical decision making model and mathematically prove that the error of a predictor based on linear regression is sensitive to the geometry of the applicant data. Therefore, if different groups have different geometries then, linear regression would be a better predictor for certain groups over others, and so the choice of using a linear predictor as decision-making tool may already favor certain groups over others.

Next, we analyze the performance of various regression methods on the 1988 NELS dataset [49]. The main advantage of regression methods for regression processes is their explainability which makes them possible to use for ranking applicants in practice. It is indeed possible to explain which features have led to rank a given applicant higher than another, an arguably basic requirement for public ranking systems. In fact, regression methods encompass several natural and common grading systems, such as for example systems that give a grade resulting from a weighted average of grades.

We show that the geometry varies between the data of the applicants of each type. Namely, the probabilities of success of applicants of different types are different functions of the attributes. Since regression methods are acutely affected by the geometry of the data, it immediately implies that the choice of the regression method has to be adapted to each group of applicants. This reinforces our aforementioned theoretical result.

In particular, using one regression method, say linear regression, for ranking all students will achieve poor prediction, even if applied on each group independently. Applicants from different groups require group-specific regression methods because of the group-specific correlation between features and success.

In a broader context, our work fits in the qualitative study of *implicit bias* in society. While social and cognitive psychologists have argued for the existence of implicit or automatic bias in various domains and individual behaviours [50–65], in many important cases questions have been raised as to how to measure and consequently reduce the implicit bias [66–68]. Moreover, in certain scenarios, the extent of impact of implicit bias is unclear. A case in point is the concept of *microagression*, which was coined by C. Pierce in 1970, later widely popularized in [69–71] and recently criticized in [72, 73].

Thus, while it is argued in literature that implicit biases combine with subjective organiza-

tional decision-making practices to perpetuate racial inequality [74], we argue further that even when if the decision making is a simple and explainable automated process, the choice of the tools used (in this case linear regression) may lead to biased decisions (against minority groups). Moreover, our findings of bias has the added benefit of being both qualitative and quantitative which makes it easier to identify and consequently address.

1.3 Outline of the Article

First, we discuss the basics of our study and argue mathematically how the heterogeneity of the geometries of the minority group impact regression methods (Section 2). Then, we provide details of the dataset that is considered for all the experiments in this article (Section 3). Next, we discuss how linear regression is a decision making tool that is less favorable to minority students by exploring the assumption of linear relationships between variables and quantitatively estimating the disparity in the number of influential features for minority applicants (Sections 4 and 5). Then, we discuss the importance of understanding the underlying geometry of the data for each type (Section 6) and propose polynomial regression as a methodology to capture this aspect of the data (Section 7).

2 Our Message: Learning Geometry is Important!

In order to present our message in a broad and general way, we first establish some mathematical notations.

For every $d \in \mathbb{N}$, $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$ and $x \in \mathbb{R}^d$, let $\|x\|_p$ denote the p^{th}-norm of x. Let $\mathcal{T} = \{T_1, \ldots, T_d\}$ be a set of types/attributes. Each type $T \in \mathcal{T}$ is a set which captures information about the type. For example, $T = [0, 100]$ if the type T represents an academic exam attribute, or $T = \{\text{male, female, non-binary}\}$ if the type T represents gender. Each candidate applicant is thought of as a d-dimensional vector in $T_1 \times T_2 \times \cdots \times T_d$, where the i^{th} coordinate denotes the attribute for the applicant on type T_i. Note that the attribute can represent academic grades, age, or even membership to certain groups.

We denote the sensitive attribute types by $\mathcal{T}^* \subseteq \mathcal{T}$. Let $\mathcal{T}' := \mathcal{T} \setminus \mathcal{T}^*$ be the rest of (non-sensitive) attributes. We further identify for every type $T \in \mathcal{T}^*$ a minority[1] subgroup $T_{\texttt{min}} \subset T$ with respect to the type T.

A predictor is trained on historical data, i.e., on a record of all the information (not just the sensitive attributes) of applicants from the past and whether each of them were deemed successful or not post the decision making process. For example, if the predictor was for determining a subset of applicants to admit to a college, then the predictor is trained on the application information of applicants from the past and whether each of the admitted applicants later did well in college.

Given a set of (current) applicants and a selection threshold fraction τ, a predictor – that has no information on the future success of the given applicants – produces a ranking that aims at maximizing the number of applicants in the top τ fraction applicants of the ranking that would

[1]We use the term *minority* group here to essentially include all subgroups that have been historically discriminated against on the basis of type T, and does not reflect the quantitative representation of the subgroup in the total population.

succeed later on. The top τ fraction applicants of the ranking are then referred to as the selected applicants.

A Type Aware mechanism is a collection of individual predictors, one for each possible collection of sensitive types (i.e., consisting of one predictor for every group formed based on the attributes in $\mathcal{T}^*$), and decisions for the subgroup of applicants whose sensitive attributes matches the profile of the group is made by the corresponding predictor. On the other hand, a Type Blind mechanism is a single predictor that does not have access to the entries of the sensitive types of the applicants. For example, if race (with say 6 attributes) and gender (with say 3 attributes) were the only sensitive set of attributes in consideration then, a type aware mechanism would have 18 individual predictors whereas a type blind mechanism would have just one single predictor.

Our Message. It has been argued in literature that Type Aware mechanisms are more fair towards minority subgroups than Type Blind mechanisms (for example see [26]). Our main (theoretical) message is that it is just not enough to use Type Aware mechanisms, but one should customize each individual predictor in the Type Aware mechanism based on the features of the data of that type. We introduce below a theoretical generating model for applicants in some decision making process and show that the accuracy of linear regression based predictors are dependent on the geometry of the attribute data on which they are trained. Thus, if two subgroups have different geometry then linear regression based predictions will be more erroneous towards one subgroup. If these subgroups have been historically discriminated against, then using linear regression only adds to the bias against them. Indeed, in the subsequent sections, we analyze empirical data and show that the geometry of the African-American applicants differs significantly from the rest of the applicant pool. Moreover, we observe that linear regression is more erroneous to the subgroup of African-American applicants (over other racial subgroups), thus reinforcing the theoretical message in this section.

Setting. We now introduce a generative model of applicants from which a decision making process would like to select the top 50% of the applicants (i.e., τ defined above is set to $1/2$). We assume that for all $T \in \mathcal{T}'$ we have $T := [0,1]$ and $\mathcal{T}'$ has m sets of attributes. For every fixing Γ of the set sensitive attributes (in $\mathcal{T}^*$), we generate a corresponding population of applicants as follows. Sample N vectors in $[0,1]^d$ uniformly and independently at random. Each of these vectors corresponds to the non-sensitive attributes of a unique applicant who sensitive attribute profile is fixed to Γ.

Our main theorem (proof deferred to Appendix) states that suppose the optimal predictor (i.e., the predictor whose recall is 1) for the aforementioned group of applicants generated corresponding to Γ is given by ranking the applicants based on the p^{th} norm of their entries in $\mathcal{T}'$ (for some p), then the error in prediction of linear regression increases with p. More formally, for any applicant x, we assume that their success probability after selection is given by:

$$\frac{\|x|_{\mathcal{T}'}\|_p}{d^{1/p}}.$$

Theorem 1. *If we use linear regression in any type aware mechanism to select* 50% *fraction of applicants from the group of applicants generated corresponding to the fixing of the sensitive attributes* Γ*, then the recall of this procedure for the group would be asymptotically (as the number of attributes, applicants grows) equal*

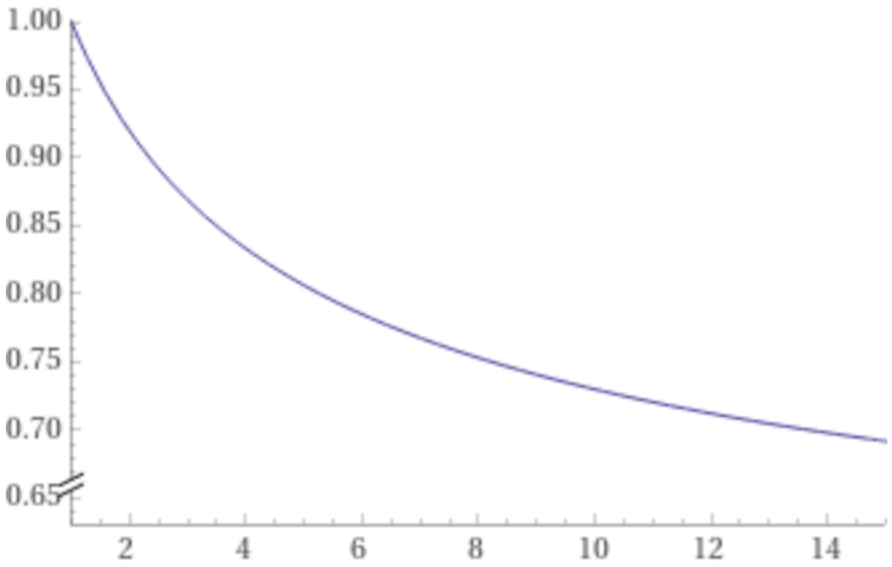

Figure 1: A plot of the error predicted by Theorem 1, with increasing values of p on the x-axis and the recall of the predictor on the y-axis.

to

$$1 - \frac{\tan^{-1}\left(\sqrt{\frac{(p+2)^2}{6p+3} - 1}\right)}{\pi}.$$

In Figure 1, we notice that for a subgroup of applicants, if the optimal predictor aligns with the ℓ_p-norm predictor for some large p then the error of the linear regression based predictor approaches 50% as p increases in our generative model (in fact for $p = \infty$, i.e., when the optimal predictor ranks based on the maximum entry of an applicant, the error of the linear regression predictor in our model is exactly 50%, i.e., linear regression simply makes a uniformly random decision!).

A decision making would be particularly harmful to society if many or all of the minority subgroups $T_{\texttt{min}}$ (for some sensitive type T) have their optimal predictors ranking aligning with high ℓ_p-norms. One may interpret this supposition that in minority groups optimal predictors prioritize exceptionalism in certain types/features over well-roundedness (see Section 6 for additional exposition). Indeed one may argue that for applicants in minority groups, there is not adequate grooming to make them well-rounded and balanced. On the other hand, exception abilities in certain areas is mainly related to talent and can be argued is spread across groups roughly uniformly. However, our theorem above indicates that linear regression would simply disregard this information and would instead try to prioritize selecting balanced individuals across groups.

3 Data

The experiments in this article are based on the public-use version of the United States of America Department of Educations National Education Longitudinal Study of 1988 dataset [49]. This dataset contains information for a nationally representative sample of students who entered the

eighth grade in the fall of 1988, with a follow-up in the years 1990, 1992, 1994, and finally in 2000. More precisely, the dataset contains surveys of students reporting on school, work, home experiences, educational resources and support, the role of parents and peers in education, neighborhood characteristics, educational and occupational aspirations, and other student perceptions.

The decision that we focus on is college admission, specifically whether to admit a student to a four-year college or university. The admit decision is only based on the predicted student performance (college grade point average). Therefore, we limit our analysis of the dataset only to the samples (i.e., records of students) for whom we know their college GPA at the time of college graduation. We focus on three ethnic groups: Caucasian, Hispanic, and African American students. The goal is to admit a certain fraction of the student applicants of each race (explicit threshold fraction will be specified later) such that we maximize recall[2] for each race.

We define a student to be *successful* after graduating from college if their GPA at graduation is greater than or equal to 3.25. After this thresholding, we note that the percentage of successful students in our sample set is higher for Caucasian students than African American students and Hispanic students (48.62% versus 30.88% and 39.14% respectively).

4 On Assuming Linear Relationships

We would like to now understand if linear regression is biased to be a better predictor for a particular race. To address this question in the simplest manner possible, we restrict the predictor to only access the grades extracted from the ninth grade transcripts for Math. The set of grades that can be obtained is $\{3, \ldots, 10\}$, and Grade 10 is the highest possible grade, whereas Grade 3 is the lowest possible grade[3]. As part of the cleaning process, we have removed all samples which have a missing grade in Math, after which we are left with 4,173 many Caucasian, 506 many Hispanic and 442 many African American students in our sample set.

In Figure 2, we have a data plot capturing the fraction of successful students of each race obtaining a particular Math grade. For example, the blue point $(6, .395)$ with label 0.087 means that among the Caucasian students whose Math grade in ninth grade was equal to 6 (which is 8.7% of the set of all Caucasian students), 39.5% ended up being successful in college. (The same-color labels do not sum to exactly 1 due to rounding error.) Additionally, we included the best line that fits the data points for each race.

The natural intuition before looking at the data is that a candidate with a higher Math grade is more likely to succeed in college. The first observation that one might make from Figure 2 is that this is flat out wrong for Hispanic students, and that the line fit to Hispanic students is very poor. Any predictor using a linear relationship will be highly erroneous for Hispanic students. Moreover, although the intuition is correct for the bulk of Caucasian students, particularly those with grades in $[7, 10]$, that relationship is less strong for African American students in the dataset. Thus, linear regression carried out on this single feature will have smaller error for Caucasian students than for Hispanic or African American students.

As a consequence, if linear regression was used as a predictor to admit students based simply

[2]Recall for a race is defined as the ratio of successful students admitted of that race to the total number of successful students of that race in the pool of applicants.

[3]In [49], the order of the grades are reversed, i.e., Grade 3 is the highest possible grade and Grade 10 is the lowest possible grade. We have however reversed this order for clarity of presentation.

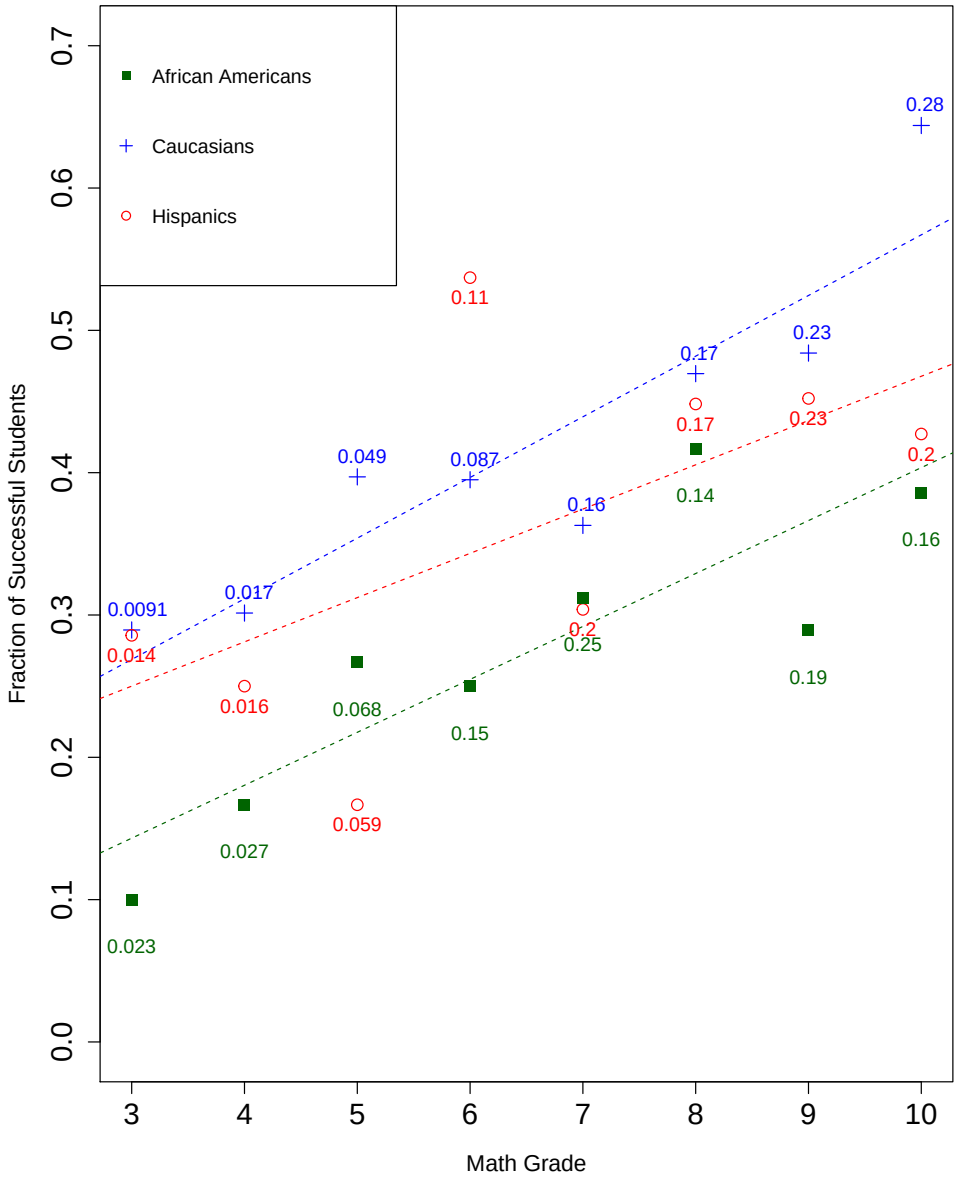

Figure 2: Data plot of Math grade for each race versus the fraction of successful students obtaining that grade. Each data point is labeled with the fraction of students of that race obtaining that grade.

on the Math grade, it would reinforce the misconception that the racial minority students cannot compete with Caucasian students. For example, consider admitting the top 10% of students of each race. The predictor would simply then choose for each race from the pool of students who obtained grade 10. While this selection is optimal for Caucasian students, the best strategy for the Hispanic (resp. African American) race would have been to choose from the pool of students who obtained grade 6 (resp. grade 8). We tentatively explain this phenomenon in the next session.

5 Quantitative Analysis of Influential Features

Here we show that the African American and Hispanic datasets have quantitatively more *influential* features than the Caucasian dataset. To that end, we now look at more features from the dataset of [49]. We include grades from ninth and eleventh grade transcripts, standardized test scores, proficiency test results, and hours spent (and accomplishment levels) of various extracurricular activities. In all, we now consider 68 different features, and since the measure of performance in each of these features were different from one another, we normalized all the scores to be in the interval $(0, 10]$, where 10 corresponded to the highest possible score. Furthermore, as part of the clean-up of the dataset we set any missing information to 0.

We performed linear regression separately for each of the three groups, Caucasians, Hispanics, and African Americas and learnt the coefficients of each feature for each race. The magnitude of the coefficient of a feature represents the importance of that feature for predicting success in college, or in other words, if the absolute value of the coefficient of a particular feature (say f) is high, then it significantly influences the success prediction between two students who have identical scores in all features except f.

In Figure 3, we have a plot of the number of features for each race whose coefficient (in absolute value) given by the regressor exceeds a threshold value. For example, the coefficient of 18 features exceed .015 for Caucasians, but that number of coefficients is 38 for Hispanics and 37 for African Americans.

We observe that for any threshold value, the number of features that influence the success of an African American or a Hispanic student is more than that for a Caucasian student, i.e., Figure 3 emphasizes that the predictors of minority students rely on a more diverse set of features. Additionally, as we demonstrated in Figure 2, linear regression is more erroneous for minority students, and the observation that minority students also have a larger number of influential features only adds to the total error in their prediction.

6 Geometric Interpretation of Data

Next, one might wonder about the performance of simple criterion for ranking (and admitting) students. Consider the following two policies.

(i) For each student consider their average grade over all subjects. Rank and admit the top students based on this average grade. This admission policy favors students with a balanced academic record.

(ii) For each student consider their maximum grade over all subjects. Rank and admit the top

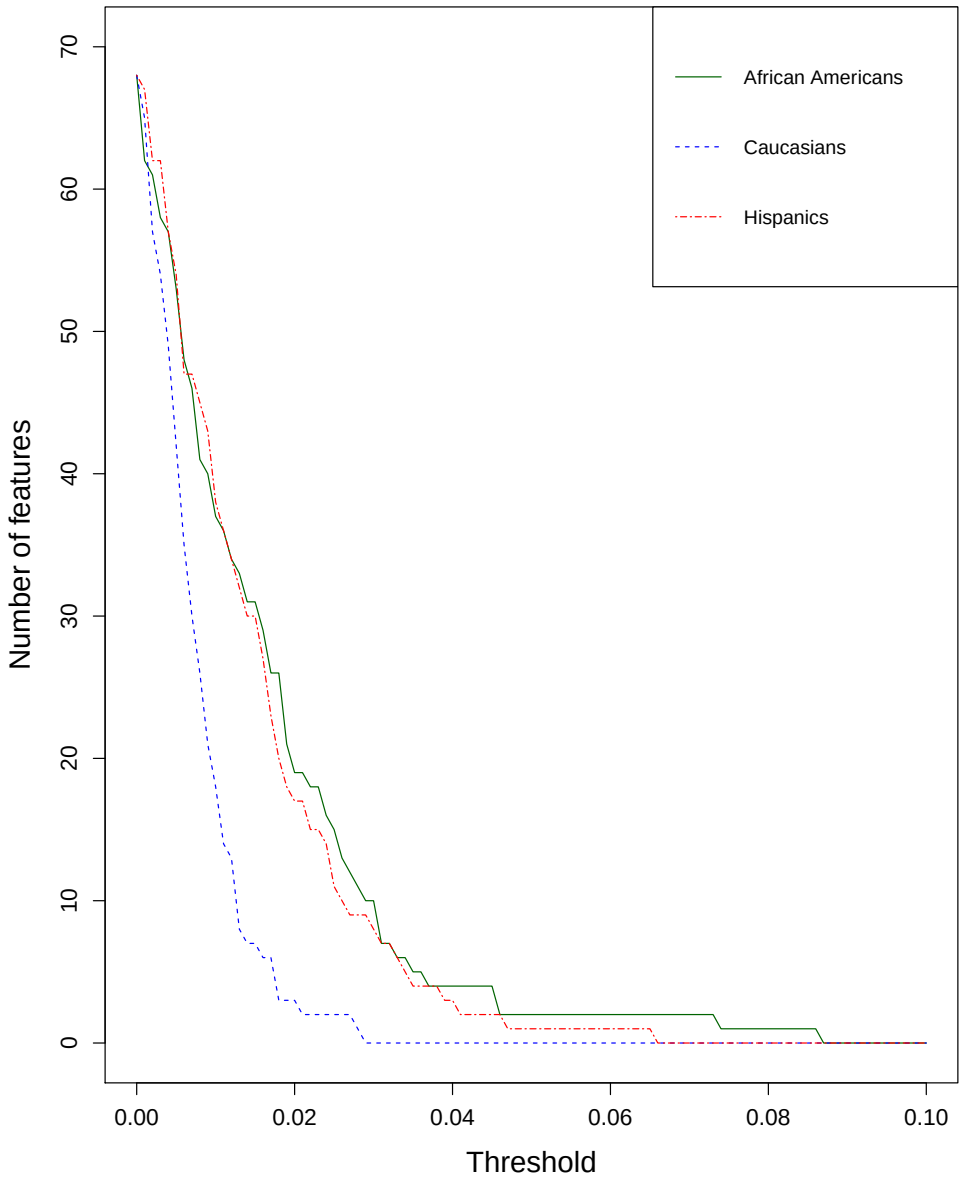

Figure 3: Number of Coefficients of the Regressor above Threshold.

students based on this maximum grade. This admission policy favors students who are excellent in at least one subject.

In order to understand which of these two policies perform well, we designed the following experiment. Let the overall set of grades of each student be represented as a d-dimensional vector, where d is the total number of features (i.e., subjects), and each coordinate entry of the vector represents the grade awarded to the student for that feature. For every $p \geq 1$, we define the p^{th} score of a student as the p^{th} norm of the vector representing the overall set of grades. Under this representation ranking students based on the p^{th} score where $p = 1$ corresponds to policy (i) above and ranking students based on the p^{th} score where p is infinity corresponds to policy (ii). The transition in the p^{th} norm of a vector from $p = 1$ to ∞ (or any large constant) is smooth and therefore through this experiment we can determine which of the two policies perform better.

We perform the above experiment by focusing on four specific features of the dataset – the Math, Science, English, and History grades extracted from the ninth grade transcripts (i.e., $d = 4$ above). We have cleaned the data and removed the records of all students with a missing grade in these specific four features for this part of the experiment. As a result, we are left with 3,559 many Caucasian, 419 many Hispanic and 345 many African American students in our sample set of which 49.76%, 40.81%, and 32.17% are successful respectively. We shall use this cleaned sample set for the rest of the experiments in this article. Finally, we note that we did not consider the extensive set of 68 features as in Figure 3 mainly because there is a lot of missing information in the 68 features which we had set to 0 by default and while this fix was fine to estimate the number of influential features, it would introduce a lot of noise for performing predictions.

We computed for every p ranging from 1 to 15 (increments of 0.05) the p^{th} score of each student and used it to rank them[4]. Then for each p, we computed the fraction of successful students ranked in the top 50% of students for each race. We plot below the relative percentage change in this fraction when compared to the fractional value computed using the p^{th} score when $p = 1$. We report our findings in Figure 4.

In Figure 4, we notice that policy (i) is better than policy (ii) for African American students and the reverse is true for Hispanic students. On the other hand for Caucasian students policy (i) is marginally better than policy (ii). Notice however that $p \approx 9$ gives the optimal ranking for African American students in this experiment, with a gain of about 3.25% in the recall. In other words, quantitatively speaking, about 4 additional successful African American students would benefit by using the p^{th} score at $p = 9$ instead of $p = 1$. Therefore informally, one may conclude that interpreting the data points of African American students in the ℓ_9-normed space is better than in the ℓ_∞-normed space, which is in turn better than in the ℓ_1-normed space. On the contrary, it is better to interpret the data points of Caucasian and Hispanic students in the ℓ_1-normed space. Another conclusion that one may draw is that Caucasian students are not as sensitive to geometric interpretation (i.e., their relative increase or decrease in recall is smaller than the students of other races). Therefore we need to be extra careful while handling the applications of minority races. These empirical inferences support the claims made in Theorem 1, but how may we use this observation for decision making?

[4]Given the grade set is $\{3, \ldots, 10\}$ and that we have four subjects, it is easy to very that the ranking given at $p_0 = 15$ is the same as as the ranking given at any $p \geq p_0$. In other words, the ranking at p_0 is the same as the ranking at $p = \infty$.

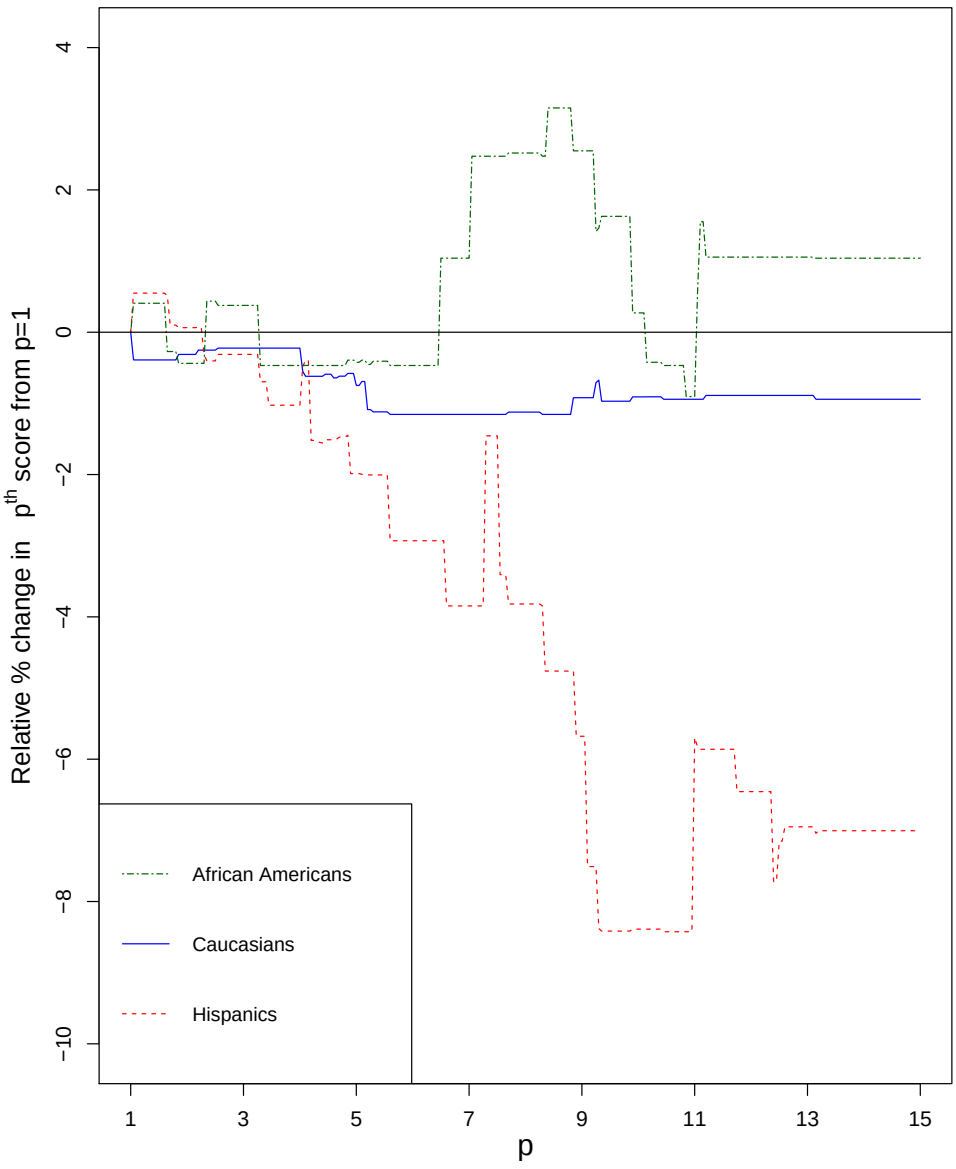

Figure 4: A plot of the relative change in recall for students of the three races as we vary p. For example, among the top 50% of Hispanic candidates in the ranking computed using the p-norm for $p = 10$, the number of successful students is about 8% less than if the ranking had been done using the p-norm for $p = 1$.

7 Polynomial Regression

Below we make a connection between algebraic tools for regression and geometric insights in ranking to provide a better way for decision making.

Notice that given a vector $\mathbf{x} \in \mathbb{R}^d$, the p^{th} norm of $\mathbf{x}$ can be interpreted as (p^{th} root of) the evaluation of the power sum symmetric polynomial of degree p in d variables on the point $\mathbf{x}$. Therefore, we can generalize the ranking experiment based on the p^{th} scores to design a ranking based on polynomial regression (with no interaction terms).

We perform polynomial regression (where the degree p of the polynomial varies) on the same cleaned dataset as described in the experiment on p^{th} score. We plot our findings in Figure 5, where we randomly split the (cleaned) dataset into training set (50% of the dataset) and test set (remaining 50% of the dataset). The polynomial regressor is then trained on the training set and then we use it to predict the success probability of students in the test set. Based on this predicted success probability of students in the test set, we admit the top 50% of the students of each race in the test set.

Qualitatively, the findings in Figure 5 broadly agree with Figure 4. The shift in the value of p between the two figures (from $p = 9$ to $p = 6$) which maximizes the recall for African American students can be attributed to the following two differences between polynomial regression and p^{th} score based ranking. First, note that the coefficients of the monomials in polynomial regression are not 1 but learnt from the training set (to potentially improve the ranking predictions). Second, note that the polynomial regression we performed not only included all monomials of degree p with no interaction terms but also lower order monomials (excluding interaction terms).

We further note that if we extrapolated back to $p = 0$, then we would obtain the mechanism of just randomly ranking students (as all students would be tied with same predicted success probability in college). A final side remark is that Figure 5 is not as smooth as Figure 4 because we ran the regressor only on integral values of p.

8 Conclusion

Prior to our work, it was known in the community that it is *fairer* to use separate predictors for each sensitive type (for example, see [26]) and we deepen the current understanding by showing that not only should each sensitive type have its own predictor, but even the methodology needs to be different for each type to guarantee good prediction across types. Case in point, we have demonstrated above that polynomial regression (where the degree of each polynomial regressor is chosen individually for each race) performs better than simply using a linear regressor for all races.

References

[1] Brent Bridgeman, Judith Pollack, and Nancy Burton. Predicting grades in college courses: A comparison of multiple regression and percent succeeding approaches. *Journal of College Admission*, 199:19–25, 2008.

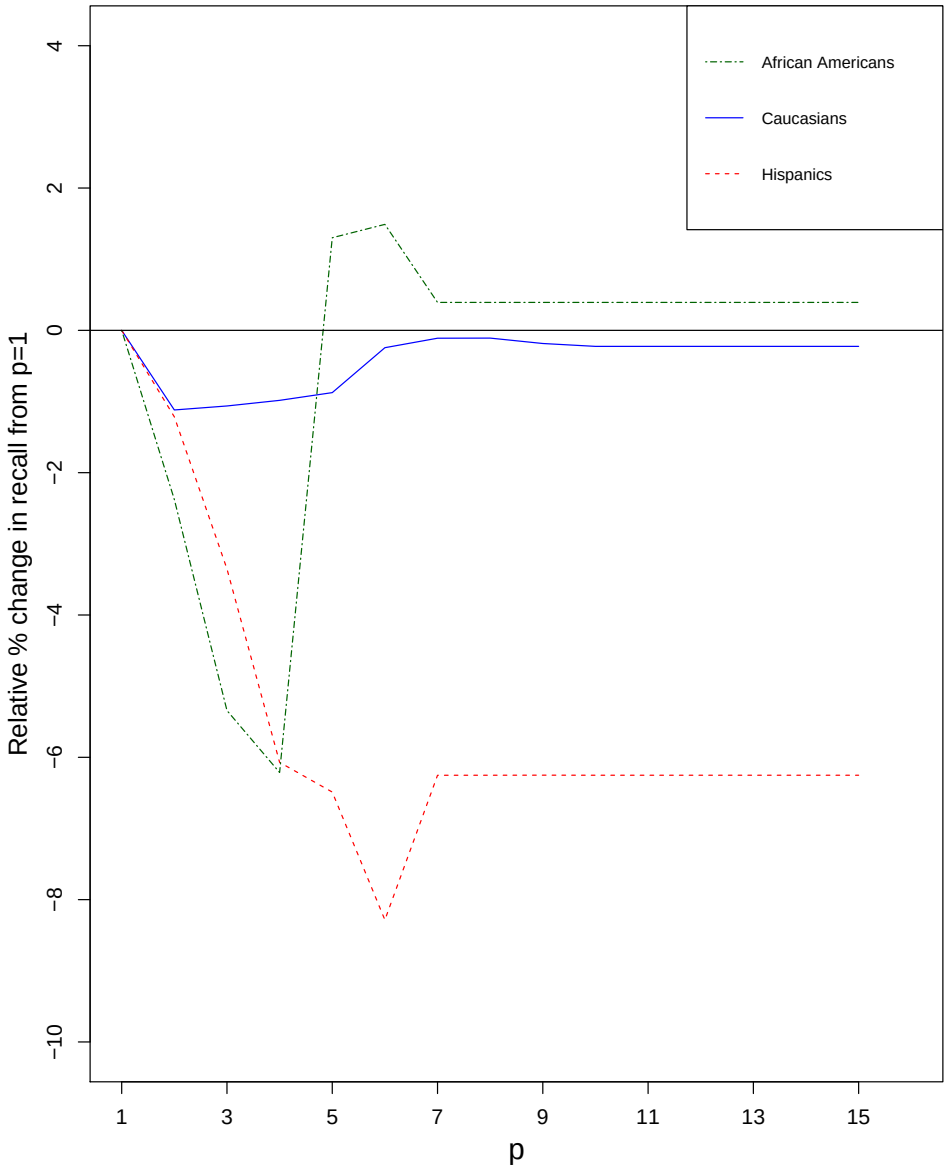

Figure 5: A plot of the relative change in recall for students of the three races as we vary the degree p while performing polynomial regression.

[2] Julie Noble and Richard Sawyer. Predicting different levels of academic success in college using high school gpa and act composite score. act research report series. 2002.

[3] Alyssa Nguyen, Brianna Hays, and Matthew Wetstein. Showing incoming students the campus ropes: Predicting student persistence using a logistic regression model. *Journal of Applied Research in the Community College*, 18(1):11–16, 2010.

[4] Eric L Dey and Alexander W Astin. Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in higher education*, 34(5):569–581, 1993.

[5] Roy D Goldman and Barbara Newlin Hewitt. Predicting the success of black, chicano, oriental and white college students. *Journal of Educational Measurement*, pages 107–117, 1976.

[6] Joshua D Angrist and Miikka Rokkanen. Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344, 2015.

[7] Salvatore Corrente, Salvatore Greco, and Roman SłOwińSki. Multiple criteria hierarchy process in robust ordinal regression. *Decision Support Systems*, 53(3):660–674, 2012.

[8] Alexander G Wesman and George K Bennett. Multiple regression vs. simple addition of scores in prediction of college grades. *Educational and Psychological Measurement*, 19(2):243–246, 1959.

[9] Brian Jacob, Jonah E Rockoff, Eric S Taylor, Benjamin Lindy, and Rachel Rosen. Teacher applicant hiring and teacher performance: Evidence from dc public schools. Technical report, National Bureau of Economic Research, 2016.

[10] Walter C Borman, Leonard A White, Elaine D Pulakos, and Scott H Oppler. Models of supervisory job performance ratings. *Journal of Applied Psychology*, 76(6):863, 1991.

[11] Jeffrey J McHenry, Leaetta M Hough, Jody L Toquam, Mary Ann Hanson, and Steven Ashworth. Project a validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, 43(2):335–354, 1990.

[12] Malcolm James Ree and James A Earles. Predicting training success: Not much more than g. *Personnel psychology*, 44(2):321–332, 1991.

[13] Nambury S. Raju, Stephen D. Steinhaus, Jack E. Edwards, and Juyne DeLessio. A logistic regression model for personnel selection. *Applied Psychological Measurement*, 15(2):139–152, 1991.

[14] Edinam Agbemava, Israel Kofi Nyarko, Thomas Clarkson Adade, and Albert K Bediako. Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in ghana. *European Scientific Journal*, 12(1), 2016.

[15] John C Wiginton. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, pages 757–770, 1980.

[16] Kevin J Leonard. Empirical bayes analysis of the commercial loan evaluation process. *Statistics & probability letters*, 18(4):289–296, 1993.

[17] Lisa R Gilbert, Krishnagopal Menon, and Kenneth B Schwartz. Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting*, 17(1):161–171, 1990.

[18] Taha Zaghdoudi. Bank failure prediction with logistic regression. *International Journal of Economics and Financial Issues*, 3(2):537, 2013.

[19] Balaji Vasan Srinivasan, Nathan Gnanasambandam, Shi Zhao, and Raj Minhas. Domain-specific adaptation of a partial least squares regression model for loan defaults prediction. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 474–479. IEEE, 2011.

[20] Zaghdoudi Khemais, Djebali Nesrine, Mezni Mohamed, et al. Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *International Journal of Economics and Finance*, 8(4):39, 2016.

[21] E David Thompson, Bethany V Bowling, and Ross E Markle. Predicting student success in a majors introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Research in Science Education*, 48(1):151–163, 2018.

[22] T Anne Cleary. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968.

[23] T. Anne Cleary and Thomas L. Hilton. An investigation of item bias. *Educational and Psychological Measurement*, 28(1):61–75, 1968.

[24] Robert M Guion. Employment tests and discriminatory hiring. *Industrial Relations: A Journal of Economy and Society*, 5(2):20–37, 1966.

[25] Robert L Thorndike. Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70, 1971.

[26] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, May 2018.

[27] Richard B Darlington. Another look at cultural fairness 1. *Journal of Educational Measurement*, 8(2):71–82, 1971.

[28] Nancy S Cole. Bias in selection. *Journal of educational measurement*, 10(4):237–255, 1973.

[29] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

[30] Hillel J Einhorn and Alan R Bass. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 75(4):261, 1971.

[31] Ronald L Flaugher. Bias in testing: A review and discussion, TM report no 36. *Educational Testing Services*, 1974.

[32] Ronald L Flaugher. The many definitions of test bias. *American Psychologist*, 33(7):671, 1978.

[33] Marshall B Jones. Moderated regression and equal opportunity. *Educational and Psychological Measurement*, 33(3):591–602, 1973.

[34] Robert L Linn. Fair test use in selection. *Review of Educational Research*, 43(2):139–161, 1973.

[35] Robert L Linn. In search of fair selection procedures. *Journal of Educational Measurement*, 13(1):53–58, 1976.

[36] Nancy S Petersen and Melvin R Novick. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, pages 3–29, 1976.

[37] Rebecca Zwick and Neil J Dorans. Philosophical perspectives on assessment fairness. *Fairness in educational assessment and measurement*, pages 267–282, 2016.

[38] Mitchell F Rice and Brad Baptiste. Race norming, validity generalization, and employment testing. *Handbook of Public Personnel Administration*, 58:451, 1994.

[39] John A Hartigan and Alexandra K Wigdor. *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* National Academy Press, 1989.

[40] Kimberly West-Faulcon. Fairness feuds: Competing conceptions of title vii discriminatory testing. *Wake Forest L. Rev.*, 46:1035, 2011.

[41] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

[42] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

[43] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

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

[45] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[46] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[47] Roland G Fryer Jr and Glenn C Loury. Valuing diversity. *Journal of political Economy*, 121(4):747–774, 2013.

[48] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[49] *National Education Longitudinal Study of 1988*, 1988. http://nces.ed.gov/surveys/nels88.

[50] M. R. Banaji and A. G. Greenwald. Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology*, 68(2):181–198, 1995.

[51] M. R. Banaji, C. Hardin, and A. J. Rothman. Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, 65:272–281, 1993.

[52] M. R. Banaji and C. Hardin. Automatic stereotyping. *Psychological Science*, 7:136–141, 1996.

[53] J. A. Bargh and F. Pratto. Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology*, 22:293–311, 1986.

[54] G. V. Bodenhausen. Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science*, 1:319–322, 1990.

[55] J. M. Darley and P. H. Gross. A hypothesis-confirming bias in labeling effects. *Journal of Personality & Social Psychology*, 44:20–33, 1983.

[56] P. G. Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56:5–18, 1989.

[57] J. F. Dovidio, N. Evans, and R. B. Tyler. Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, 22:22–37, 1986.

[58] J. F. Dovidio, K. Kawakami, C. Johnson, B. Johnson, and A. Howard. On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology*, 33:510–540, 1997.

[59] R. H. Fazio, J. R. Jackson, B. C. Dunton, and C. J. Williams. Variability in automatic activation as an unobtrusive measure of racial attitudes. a bona fide pipeline? *Journal of Personality and Social Psychology*, 69:1013–1027, 1995.

[60] R. H. Fazio, D. M. Sanbonmatsu, M. C. Powell, and F. R. Kardes. On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50:229–323, 1986.

[61] S. L. Gaertner and J. P. McLaughlin. Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 46:23–30, 1983.

[62] C. N. Macrae, G. V. Bodenhausen, A. B. Milne, and J. Jetten. Out of mind but back in sight: Stereotypes on the rebound. *Journal of Personality and Social Psychology*, 67:808–817, 1994.

[63] C. W. Perdue and M. B. Gurtman. Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology*, 26:199–216, 1990.

[64] L. A. Rudman and E. Borgida. The afterglow of construct accessibility: The behavioral consequences of priming men to view women as sexual objects. *Journal of Experimental Social Psychology*, 31:493–517, 1995.

[65] C. Stangor, L. A. Sullivan, and T. E. Ford. Affective and cognitive determinants of prejudice. *Social Cognition*, 9:359–380, 1991.

[66] Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology*, 117(3):522, 2019.

[67] Franziska Meissner, Laura Anne Grigutsch, Nicolas Koranyi, Florian Müller, and Klaus Rothermund. Predicting behavior with implicit measures: Disillusioning findings, reasonable explanations, and sophisticated solutions. *Frontiers in Psychology*, 10:2483, 2019.

[68] Corneille O. and Hütter M. Implicit? what do you mean? a comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, 24(3):212–232, 2020.

[69] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271, 2007.

[70] Derald Wing Sue. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons, 2010.

[71] Michele A Paludi. *Managing Diversity in Today's Workplace: Strategies for Employees and Employers [4 volumes]*. ABC-CLIO, 2012.

[72] Greg Lukianoff and Jonathan Haidt. *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin Books, 2019.

[73] Edward Cantu and Lee Jussim. Microaggressions, questionable science, and free speech. *Texas Review of Law & Politics, Forthcoming*, 2021.

[74] Daniel Hirschman and Emily Adlin Bosk. Standardizing biases: Selection devices and the quantification of race. *Sociology of Race and Ethnicity*, 6(3):348–364, 2020.

A Proof of Theorem 1

Suppose the linear regression predictor tries to fit the population to the equation $A\vec{\beta} = \vec{y}$, where each row of A corresponds to the non-sensitive attributes of an applicant (i.e., each row of A is a uniformly random vector in A) and the i^{th} coordinate of y is given by the p^{th} norm of the i^{th} row of A. Note that $\vec{\beta}$ is the vector that minimizes the least square error, $\|A\vec{\beta} - \vec{y}\|_2$.

First, we show below that $\vec{\beta}$ must have (almost) the same entry in all coordinates.

Theorem 2. *As $n \to \infty$, with high probability $1 - o(1)$, the optimal vector $\vec{\beta} = (b, \ldots, b) + o(1)$ for some $b \in \mathbb{R}$.*

Proof. The error $\|A\vec{\beta} - \vec{y}\|_2^2$ is equal to $\sum_{i=1}^{n} |A_i \cdot \vec{\beta} - y_i|^2$, where y_i is equal to the (scaled) p^{th} norm of the row vector A_i. So we have a sum of N independent identically distributed copies which by large of large numbers tends to $\sim N \cdot \mathbb{E}(|A_i \cdot \vec{\beta} - y_i|^2)$. The expected value $\mathbb{E}(|A_i \cdot \vec{\beta} - y_i|^2)$ is a quadratic form in $\vec{\beta}$ given by:

$$\mathbb{E}(|A_i \cdot \vec{\beta} - y_i|^2) = \mathbb{E}((\left(\vec{\beta}^T \cdot A_i^T - y_i\right)\left(A_i \cdot \vec{\beta} - y_i\right)) = \mathbb{E}(\vec{\beta}^T A_i^T A_i \vec{\beta}) - 2\mathbb{E}(y_i A_i \cdot \vec{\beta}) + \mathbb{E}(y_i^2).$$

Because the form $Q(\vec{\beta}) = \mathbb{E}(|A_i \cdot \vec{\beta} - y_i|^2) = \mathbb{E}(\vec{\beta}^T A_i^T A_i \vec{\beta}) - 2\mathbb{E}(y_i A_i \cdot \vec{\beta}) + \mathbb{E}(y_i^2)$ is non-negative, it has a unique minimum $\vec{\beta}$ which satisfies gradient condition $\frac{dQ(\vec{\beta}+t\mathbf{v})}{dt} = 0$ for every vector $\mathbf{v}$.

$$Q(\vec{\beta}+t\mathbf{v}) = \mathbb{E}((\vec{\beta}+t\mathbf{v})^T A_i^T A_i(\vec{\beta}+t\mathbf{v}) - 2\mathbb{E}(y_i A_i \cdot (\vec{\beta}+t\mathbf{v})) + \mathbb{E}(y_i^2)$$
$$= Q(\vec{\beta}) + t\left(\mathbf{v}^T\mathbb{E}(A_i^T A_i)\vec{\beta} + \vec{\beta}^T\mathbb{E}(A_i^T A_i)\cdot\mathbf{v} - 2\mathbb{E}(y_i A_i)\cdot\mathbf{v}\right) + O(t^2).$$

Therefore for every $\mathbf{v}$, we have

$$\mathbf{v}^T\mathbb{E}(A_i^T A_i)\vec{\beta} + \vec{\beta}^T\mathbb{E}(A_i^T A_i)\mathbf{v} - 2\mathbb{E}(y_i A_i)\cdot\mathbf{v} = \mathbf{0}$$
$$\Rightarrow 2\mathbb{E}(A_i^T A_i)\vec{\beta} - 2\mathbb{E}(y_i A_i) = \mathbf{0}.$$

Hence the minimizer $\vec{\beta}$ satisfies

$$\mathbb{E}(A_i^T A_i)\vec{\beta} = \mathbb{E}(y_i A_i).$$

If y_i is any symmetric function of the coordinates of A_i (like pth norm in this case), we have that $y_i A_i$ is vector with identically distributed coordinates, so the vector $\mathbb{E}(y_i A_i)$ has all equal coordinates.

If we use that A_i is vector with iid coordinates from $[0,1]$, then $\mathbb{E}(A_i^T A_i)$ is given by a matrix M with the (r,s) entry given by $M_{r,s} = \mathbb{E}(X_r X_s)$, where X_i are iid uniform variables on $[0,1]$.

So we have $M = \frac{1}{3}I + (\frac{1}{4}J - \frac{1}{4}I) = \frac{1}{12}I + \frac{1}{4}J$, where J is the all ones matrix, and thus we see that $\vec{\beta}$ satisfies $\vec{\beta} = (b, b, \cdots, b)$, where $\frac{b}{12} + \frac{db}{4} = \mathbb{E}(y_i X_r)$.

Thus, we have that the minimizer of this quadratic form is of the form $\vec{\beta} = (b, b, \cdots, b)$ for some $b \in \mathbb{R}$. Therefore with probability $1 - o(1)$, the minimizer $\vec{\beta} = (b, b, \cdots, b) + o(1)$ (Because the quadratic form is very close to this expectation quadratic form with high probability, and the minimizer doesnt change under slight perturbations to the form) □

Therefore, informally, we may conclude that linear regression simply selects the top 50% of the applicants based on their ℓ_1 norm.

Thus ranking applicants using regression is equivalent to ranking according to $A_i \cdot \vec{\beta}$ which is proportional to the $\sum_{j=1}^{d} \vec{\beta} A_i(j)$ which essentially the ℓ_1 norm of A_i. (We are going to rank according to $\vec{\beta} = (b, b, \cdots, b) + o(1)$, so the ranking which is determined by the volumes of the region $\vec{\beta} \cdot X > \tau$ is essentially equal to the volumes $(b, b, \cdots, b) \cdot X > \tau$- which corresponds to the rank by ℓ_1 because in this case (non-negative entries) $\|X\|_1 = (1, 1, \cdots, 1) \cdot X$.)

Let S_p be a ranking of vectors in $[0,1]^d$ based on their ℓ_p-norm. Let S_p^τ be the restriction of the ranking to the top τ fraction of applicants. The value of $|S_1^\tau \setminus S_p^\tau|$ gives us the recall of the theorem statement and this is calculated below.

Theorem 3. *Let S be a uniformly random sample of N points in $[0,1]^d$. Let $p \in \mathbb{R}_{\geq 1} \cup \{\infty\}$. After ranking the points in S by their ℓ_1-norm (resp. ℓ_p-norm), let $S_1 \subset S$ (resp. $S_p \subset S$) be all points in S ranked in the top half (breaking ties randomly). Then for large enough d, n we have that*

$$\frac{|S_p \cap S_1|}{|S_1|} \sim 1 - \frac{\tan^{-1}\left(\sqrt{\frac{(p+2)^2}{6p+3} - 1}\right)}{\pi}.$$

Proof. By taking n large enough, it is enough to consider the case where you pick a point $\mathbf{x}$ randomly from $[0,1]^d$ and compute the following probability:

$$\Pr_{\mathbf{x}\sim[0,1]^d}\left[\|\mathbf{x}\|_1 \geq m_1 \text{ and } \|\mathbf{x}\|_p \geq m_p\right],$$

where m_1 (resp. m_p) is the median of the distribution of $\|\mathbf{x}\|_1$ (resp. $\|x\|_p$). Asymptotically, for large d, we have $m_p \sim \sqrt[p]{\frac{d}{p+1}}$. If $\mathbf{x} := (x_1, \ldots, x_d)$ then the above probability is essentially the following:

$$\Pr_{\mathbf{x}\sim[0,1]^d}\left[x_1 + \cdots + x_d \geq m_1 \text{ and } x_1^p + \cdots x_d^p \geq m_p^p\right].$$

For every $i \in [d]$, let $\mathbf{y_i} := (x_i, x_i^p)$. Then the probability can be seen as:

$$\Pr_{\mathbf{x}\sim[0,1]^d}\left[\mathbf{y_1} + \cdots + \mathbf{y_d} \in \mathcal{R}\right], \tag{1}$$

where $\mathcal{R} := [m_1, \infty) \times [m_p^p, \infty)$. Also note that for every $i \in [d]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim[0,1]^d}[\mathbf{y_i}] = \left(\frac{d}{2}, \frac{d}{p+1}\right) \sim (m_1, m_p^p).$$

Thus, applying central limit theorem to all the $\mathbf{y_i}$s, as $d \to \infty$, we have:

$$\frac{1}{\sqrt{d}} \cdot \sum_{i\in[d]} (\mathbf{y_i} - \mathbb{E}[\mathbf{y_i}]) \to \mathcal{N}(0, \Sigma),$$

where Σ is the covariance matrix of $\mathbf{y_i}$s given by $\mathbb{E}[\mathbf{y_i}^T\mathbf{y_i}]$. We can thus compute Σ to be:

$$\Sigma = \mathbb{E}\begin{bmatrix} x_i^2 & x_i^{p+1} \\ x_i^{p+1} & x_i^{2p} \end{bmatrix} = \begin{bmatrix} 1/3 & 1/p+2 \\ 1/p+2 & 1/2p+1 \end{bmatrix}.$$

So the probability in (1) converges to

$$\Pr_{\mathbf{Y}\sim\mathcal{N}(0,\Sigma)}[\mathbf{Y} \in [0,\infty) \times [0,\infty)].$$

Note that the distribution of $\mathcal{N}(0, \mathbf{\Sigma})$ is given by

$$\frac{\exp\left(-\frac{1}{2}(\mathbf{Y}^T\mathbf{\Sigma}^{-1}\mathbf{Y})\right)}{2\pi \cdot |\det \mathbf{\Sigma}|}$$

Moreover, we have the inverse of Σ is:

$$\mathbf{\Sigma}^{-1} = \frac{1}{\frac{1}{3(2p+1)} - \frac{1}{(p+2)^2}} \begin{bmatrix} \frac{1}{2p+1} & -\frac{1}{p+2} \\ -\frac{1}{p+2} & \frac{1}{3} \end{bmatrix}.$$

Thus, using the integral

$$\int_0^\infty \int_0^\infty \exp(ax^2 + bxy + cy^2)\, dx\, dy = \frac{1}{2\sqrt{4ac - b^2}} \left(\pi + 2 \arctan\left(\frac{b}{\sqrt{4ac - b^2}} \right) \right),$$

we can compute probability to be the expression given in the theorem statement. □