Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Shuxuan (Annie) Luo April 3, 2023

Detecting Training Data Biases: MLR And Graphical LASSO Based Methods

by

Shuxuan (Annie) Luo

Kevin McAlister, Ph.D.
Advisor

Quantitative Theory and Methods

Kevin McAlister, Ph.D.

Advisor

Lauren Klein, Ph.D.

Committee Member

Jessica Sun, Ph.D.

Committee Member

2023

Detecting Training Data Biases: MLR And Graphical LASSO Based Methods

By

Shuxuan (Annie) Luo

Kevin McAlister, Ph.D.
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2023

Abstract

Detecting Training Data Biases: MLR And Graphical LASSO Based Methods
By Shuxuan (Annie) Luo

As the use of algorithms for automated decision-making became increasingly prevalent, many have pointed out the discriminatory results produced. This paper aims to extract and evaluate one source of such discrimination—the unintentional biases captured in the training data through high correlations between the predictors and the protected characteristics. To see if a predictor is systematically excluding qualified members belonging to a protected group, we examine the "direct" correlation between this predictor and the protected characteristic, controlling for all other predictors in the training data. We first propose a Multivariable Linear Regression test, adapted from the "Input Accountability Test." We also propose using a Graphical LASSO based test. We applied all three tests on detecting biases in our simulated datasets, and we found GLASSO to work the best. Finally, we discuss limitations of GLASSO and where we can improve.

Detecting Training Data Biases: MLR And Graphical LASSO Based Methods

By

Shuxuan (Annie) Luo

Kevin McAlister, Ph.D.

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2023

Acknowledgments

I would like to extend a huge thank you to Dr. Kevin McAlister, my advisor, for his help with forming and sharpening my research question, for pointing me to relevant statistical methods, and for guiding me through the whole research process. I would also like to thank my committee members, Dr. Lauren Klein and Dr. Jessica Sun, for supporting me through my defense, and for their constructive feedback.

Contents

1 Introduction **1**

2 Problem Statement **4**

3 Methods **6**

3.1 Input Accountability Test 6

3.1.1 "Indirect" Relationships 8

3.1.2 Significance Testing 9

3.1.3 The "Direct" Relationship 10

3.2 Multiple Linear Regression 10

3.3 Graphical LASSO 12

3.3.1 Precision Matrix and Conditional Independence 12

3.3.2 LASSO 14

4 Data **16**

4.1 Variables 17

4.2 Biased Data Simulation 17

4.3 Unbiased Data Simulation 18

5 Results **20**

5.1 IAT 20

5.1.1 Biased Dataset 20

5.1.2 Unbiased Dataset 21
5.2 MLR 22
5.2.1 Biased Dataset 22
5.2.2 Unbiased Dataset 22
5.3 GLASSO 23
5.3.1 Biased Dataset 23
5.3.2 Unbiased Dataset 25

6 Conclusion **27**

7 Future Works **29**

8 Discussion **31**

Appendix A IAT Regression Results **33**

Appendix B MLR Results **36**

Bibliography **38**

List of Figures

2.1 Correlation of interest 5

3.1 Target pathway closed 7
3.2 Leftover correlation 8
3.3 More precise "direct" correlation 12

4.1 "Direct" relationship exists 18
4.2 Lack of "direct" relationship 19

5.1 Partial correlations in unbiased dataset 23
5.2 GLASSO for biased dataset 24
5.3 GLASSO for unbiased dataset 26

A.1 Summary for first step IAT, biased 33
A.2 Summary for second step IAT, biased 34
A.3 Summary for second step IAT, unbiased, large sample size 34
A.4 Summary for first step IAT, unbiased 34
A.5 Summary for second step IAT, unbiased 35

B.1 Summary for MLR, biased 37
B.2 Summary for MLR, biased 37

List of Tables

5.1 Summary statistics for first step of IAT in biased dataset 20
5.2 Summary statistics for second step of IAT in biased dataset 20
5.3 Summary statistics for second step of IAT with large sample size in unbiased dataset 21
5.4 Summary statistics for first step of IAT in unbiased dataset 21
5.5 Summary statistics for second step of IAT in unbiased dataset 22
5.6 Summary statistics for MLR in biased dataset 22
5.7 Summary statistics for MLR in unbiased dataset 22

6.1 Summary of how well the three methods worked 28

Chapter 1

Introduction

Automated machine learning systems, powered by extensive data, have been used prevalently to make decisions in all kinds of fields in our society, such as finance, medicine, and criminal systems.[2, 17, 3] The application of algorithms can lead to minor decisions, such as fraud detection, targeted advertisements, and optimized routes for delivery.[2] In the case of route optimization, one can easily see the positive effects of an increased efficiency that mathematics and algorithms can bring to our lives. However, machine learning is also applied to aid or directly make decisions that have important, often life-long implications to individuals, such as determining the limit of credit cards, deciding whether to loan mortgages, evaluating the level of sickness, and predicting probabilities of recidivism.[17] Problematically, some algorithm-based decisions have been found to involve racial biases. For example, the Optum algorithm developed by UnitedHealth Group, initially designed to identify patients who need extra care and to optimize medical resources, was found "less likely" to assign black patients "than white patients to get extra medical help, despite [the black patients] being sicker."[10] As a result, the algorithm is letting "healthier white patients cut in line ahead of sicker patients."[10] Such algorithms that clearly involve racial biases intuitively violate anti-discrimination initiatives. This paper aims to propose a statis-

tical method to detect potential algorithmic biases, with the intention of contributing to the overall evaluation and regulation of algorithms.

The classification models, which, by definition, are used to classify and thus to discriminate among different classes, seem to contradict our intention of eliminating discriminations. However, employing Bartlett et al.'s analysis regarding "antidiscrimination principles of Title VII of the Civil Rights Act of 1964," compounded with their discussion on the "burden-shifting framework",[3] we can see that the metrics against which the algorithms would be justified to discriminate are of "legitimate 'business necessity'," such as required job-related skills for employment, or actual level of sickness in the UnitedHealth Group example.[3] The discriminations with respect to other aspects, such as race and ethnicity, are of the kinds that we intend to diminish. In other words, our definition of a non-discriminatory algorithm is one that will not systematically misclassify qualified members (qualification as defined with respect to legitimate business necessity) due to their genders, races, and/or other protected characteristics. Applying the legal terms defined in Bartlett et al.'s essay, we allow "disparate treatment" based on qualifications, but our method aims to identify "disparate impact" across racial, gender, and/or other protected groups that certain algorithms can illegitimately lead to.[3]

Such unjustified disparate impact can come not only from the algorithms themselves through the manual programming that gives "certain factors inappropriate weights," but also from the data-mining process[2] where the input data already involves unintentional biases. Continuing the UnitedHealth Group's example, the reason why the Optum algorithm assigned less Black patients than it should have to receive extra medical help is because the level of sickness in that algorithm was approximated using an inappropriate proxy variable of "medical cost." In this case, adjusting the weight of cost, meaning to adjust how much medical cost factors into assessing the level of sickness, is one way to alleviate the issue from a programmer's perspective.

However, we can also observe the social, systemic problem of "health-care spending for black patients" being generally "less than for white patients with similar medical conditions."[10] This example illustrates that the wider social mechanisms at work can also lead to biased "training data"—"the data that train the model to behave in a certain way," due to the high correlation between the protected groups and the predictors used in the algorithms.[2] This correlation makes predictors to contain the protected demographic information, and in turn leads to discriminatory models.

As mentioned in the guidelines protecting individuals against algorithmic discriminations,[13] the White House specifically proposed avoiding the use of the predictors that significantly correlate with protected characteristics. We will use the help from statistics to extract and examine such correlations and thus to detect those biased predictors.

Chapter 2

Problem Statement

To aid the White House's initiative of excluding biased predictors, this paper aims to statistically determine if a contested predictor in the training data, hereon referred to as a proxy, significantly correlates with a protected group. We will explore methods to extract the relationships between the proxy of interest and the protected group, and to evaluate the significance of such relationships. It's important to note that, because we allow disparate treatment while disallowing disparate impact, we need to categorize the correlations between the proxy and the protected feature into direct ones and indirect ones. An "indirect" relationship, corresponding to disparate treatment, means that genders/ races correlate with the proxy of interest through a pathway with target variable or some other proxies in between. The target refers to the variable of business necessity. A "direct" relationship, corresponding to disparate impact, refers to a change in genders/ races that directly leads to a change in the proxy, given the target variable and other proxies controlled. See Figure 2.1. for "direct" relationship. This categorization demonstrates the intricacies of the spirit of antidiscrimination, where some specific correlations are allowed, while others, though between the same pair of variables, are not.

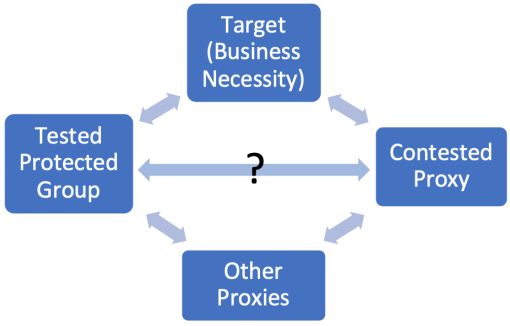

Figure 2.1: Correlation of interest

Chapter 3

Methods

One source of algorithmic bias comes from using an inappropriate proxy variable to train a predictive model. The "inappropriateness" is defined by a significant correlation between the proxy and a protected group, controlling for the target variable and other proxies. In short, a proxy is inappropriate when the protected group has a significant "direct" relationship with that proxy. We therefore need to first address how to tease out the "direct" relationship. Then, with the aim of defining a "significant" correlation, we turn to the methods for significance testing.

3.1 Input Accountability Test

The "Input Accountability Test (IAT)," proposed by Bartlett et al., is a two-step process that each tackles a corresponding area we laid out above, namely, finding the "direct" relationship and determining its significance. In the first step, the IAT calculates the "residual", the part in the proxy's variation that does not perfectly correlate with the variation of the target variable. By controlling for the target in the first step, the IAT defines a "direct" correlation as what correlations through the target are not. It then determines if this residual has any statistically significant relationship with the protected group.[3] In practice, the first step runs a simple linear

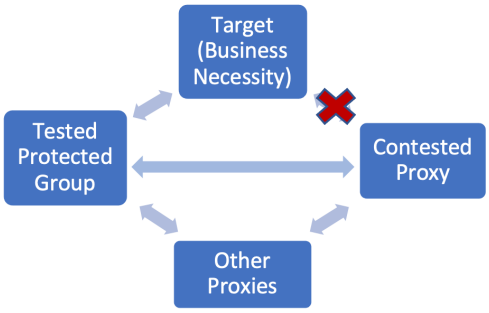

Figure 3.1: Target pathway closed

regression, $proxy = \alpha_0 * target + \varepsilon_1$, where the residual vector ε_1 would capture any variations of the proxy unrelated to that of the target. Notice here that the proxy is the dependent variable, and the target is the independent variable. While counter-intuitive since the unobserved target usually sits on the left-hand side waiting to be determined by the predictive models, it's important to remember that the objective is not to see how well the proxy can predict the target. Rather, this first step is to close off the pathway of gender/ race correlating with the proxy through the target. See Figure 3.1.

In the second step, the IAT runs another simple linear regression, $\varepsilon_1 = \alpha_1 * ProtectedGroup + \varepsilon_2$, to test if the residual has any statistically significant correlations with the protected group. As the residual serves to control the target, the second step quantifies and evaluates the left-over correlations denoted with a question mark in Figure 3.2.

The significance testing employed by Bartlett et al. is a two-sided p-value test, where the null hypothesis states that $\alpha_1 = 0$. The decision rule for statistical significance is if $p < 0.05$.[3]

While having the merit of running the significance testing against only the conception of a "direct" relationship between the protected characteristic and the proxy of interest, thus trying to investigate the correlations disallowed under antidiscrimination's definitions,[3] a close read of the IAT leaves us yet two major kinds of issues. We will

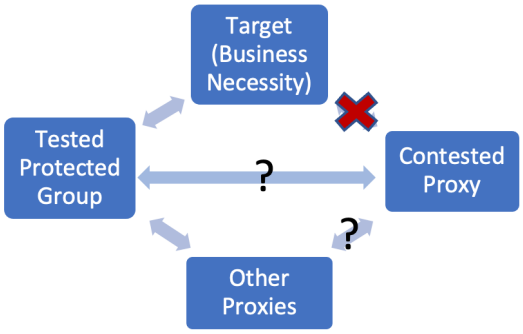

Figure 3.2: Leftover correlation

first question how IAT's loosely defined method of eliminating "indirect" relationships between the proxy and the target can be employed in bad faith. We then criticize its use of p-tests for significance testing. We will then address a wider difficulty of carving out all possible "indirect" relationships with the intention of obtaining the exact "direct" relationship. As the IAT disregards situations where a protected group can correlate with the contested proxy through the pathways of other proxies in between, this method bears the risk of falsely claiming the proxy to be inappropriate even when it in fact does not directly discriminate against the protected characteristic. We can easily imagine this risk manifesting in real-world cases, due to the complexities introduced by correlations among different proxies.

3.1.1 "Indirect" Relationships

Defined as the correlations between the target and the proxy, the variations in the proxy due to those "indirect" relationships are allowed in Bartlett et al.s' conception of the antidiscrimination laws. This is the reason why they picked out the residual using the first step. However, as already acknowledged, their spelled-out way of excluding "indirect" correlations "focused on linear settings, but the IAT could in principle be amended to handle nonlinear models."[3] Without specifying how to find such an "indirect" relationship other than applying simple linear regressions,

we worry that a test, while fulfilling the ill-defined "spirit" of the IAT as it tries to quantify the relationship between the proxy and the target, may extract a residual using an overfitted model that could involve $target^2$, $target^3$, etc.. The overfit can reduce the residual to a bare minimum. Such practices can lead to a statistically insignificant correlation between the residual and protected group, resulting in a strong justification for using virtually all proxies. Therefore, the loosely defined first-step of the IAT can defeat its own purposes.

3.1.2 Significance Testing

The opposite result of deeming all proxies biased and thus inappropriate, which would also defeat the IAT's purposes, can happen in the second step of the IAT when the sample size of the training data is large enough. As already pointed out by Bartlett et al., their reliance on the p-value test renders the IAT's claims of any statistically significant correlations questionable.[3] There are two main critiques on the p-test's ability of determining the significance of the relationship between the residual and the protected group. When the sample size is large enough, the p-value would quickly diminish to a value smaller than the 0.05 threshold.[8] The formula calculating the p-value for a two-sided test is $p = 2 * \Phi(\frac{\hat{\alpha_1}-0}{sd/\sqrt{n}})$, where Φ is the normal cumulative distribution function (normal CDF) and $\frac{\hat{\alpha_1}-0}{sd/\sqrt{n}}$ calculates the t-statistic. The bigger the sample size n is, the bigger the t-statistic will be, and the less probable that we would observe our samples under the null hypothesis. A direct implication of this sample size critique is that "a company that brings a large dataset to bear on an IAT test might be disadvantaged relative to firms with less data."[3]

Another critique against p-test is that, even when the null hypothesis failed to be rejected, we cannot directly translate this situation as claiming that the "direct" relationship we tested for is insignificant. This is because there's a gap between failure to reject the null vs acceptance of the null.[14]

3.1.3 The "Direct" Relationship

Defined using a negative approach, where the "direct" relationship is what the "indirect" ones are not, the "direct" correlation depends on an exact exclusion of the "indirect" relationships. Under Bartlett's construction, we should only exclude correlations through the pathway of the target, i.e., the business necessity variable. This is because, with the spirit of antidiscrimination laws, the only legitimate base to offer disparate treatment is the difference in values of the target variable.[3] However, a literal read of specific antidiscrimination laws, such as employment antidiscrimination statements,[9] defines "direct" relationships using a positive approach. Under this statement, we can see that disparate treatment against, say, family income seems to be implicitly allowed as family income was not positively listed. As the protected characteristics have been positively defined in the statement, the "direct" correlations that statistical methods are looking for should control for all variables not listed. This expands our previous conception of allowed "indirect" relationships from business necessity only to include pathways of all proxies. See Figure 3.3. By having the need to control for different pathways, we propose to use a multivariable linear regression (MLR) method to control for all background variables. Here, background variables refer to all other proxies. Another benefit of running a MLR is that, by reducing omitted-variable bias, we can obtain more accurate estimates of the coefficients for the protected group.

3.2 Multiple Linear Regression

The two-step IAT process for one proxy with the aim of excluding "indirect" relationships through controlling the target variable can be simplified in a one-step multivariable linear regression model. In MLR, we also have additional proxies controlled to investigate the direct pathway noted in Figure 3.3. The function of the

model is

$$Proxy = \beta_0 * Target + \beta_1 * ProtectedGroup + \beta_2 * OtherProxy + \beta_3 * OtherProxy + \varepsilon_0$$

, where β_1 has a clear interpretation that the change in value in *ProtectedGroup* would lead to a β_1 unit of change in Proxy, when Target and other proxies are controlled. In other words, β_1 characterizes the "direct" relationship we are looking for, as the "indirect" relationships through Target and other proxies are always controlled. Note that this is not a mathematical simplification of the IAT, meaning that the value of β_0 is expected to be different from that of α_0, and β_1 different from α_1. By the Frisch-Waugh-Lovell Theorem, obtaining the same coefficients would require regressing a second error term on the first one,[15] where the second error term comes from regressing between *ProtectedGroup* and Target. IAT's regression of the residual on a protected group is different from this process, thus the MLR would return different coefficients, as noted by the betas. The advantage of using the betas retrieved from the MLR is that the betas account for the correlations both between *ProtectedGroup* and Target and between other pairs of explanatory variables. Therefore, the betas are better estimates of the coefficients by alleviating the omitted-variable bias. These better estimates can then influence the significance testing, as each beta is a component of the formula calculating the p-value.

The MLR, however, like the IAT relies on a good method of significance testing. As we discussed above, the most prevalent p-test is not an ideal method. We therefore proceed to our second proposal of using the Graphical LASSO method, which not only calculates partial correlations, and thus the "direct" relationships, between each pair of variables, but also provides a new way of testing the significance through its relationship-selection feature provided by the L_1 regularization term. GLASSO is a graphical model designed for network analysis, which we think is better equipped to

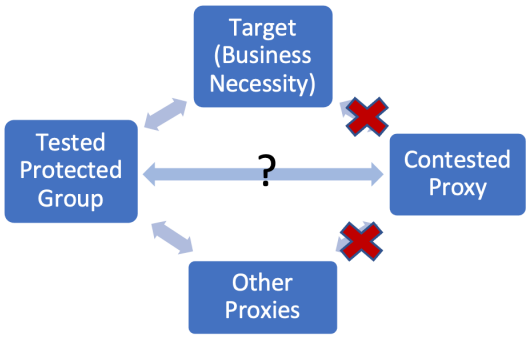

Figure 3.3: More precise "direct" correlation

tackle the network of proxies, a protected group, and the target.

3.3 Graphical LASSO

The Graphical LASSO is a method of selecting the strongest relationships between pairs of variables, conditional on all rest of the variables. By applying "a lasso penalty... to the inverse covariance matrix,"[11] GLASSO mainly does two things. First, we calculate a covariance matrix for the training dataset, invert it, and obtain a precision matrix. As the entries in the precision matrix have a specific meaning of conditional correlation, the entries are measuring "direct" relationships. Thus, GLASSO is also measuring the correlation shown in Figure 3.3. Second, we make our precision matrix sparser, meaning to reduce more entries to zero, based on a lasso penalty. This second step is to test the significance of non-zero entries in the precision matrix.

3.3.1 Precision Matrix and Conditional Independence

After transforming the training data to a matrix, we first scale all columns so that the variables won't show high correlations due to the difference in magnitude of each

variable's values. Denote X to be the vector of all variables. The covariance matrix calculates the covariance between each pair of scaled variables, one coming from X^T and the other coming from X. As a result, the covariance matrix has a diagonal of variances of each variable, and the non-diagonal entries are calculated by

$$cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])^T]$$

, thus always returning a symmetric matrix. As the formula does not involve the rest of variables in X for specific pairs, we cannot differentiate "direct" relationships from "indirect" ones. Using Schur complement of a block matrix,[6] it has been proved that inverting the covariance matrix calculates partial correlations. Under a multivariate normal distribution (MND), such partial correlations are equal to conditional correlations,[1] which are the "direct" relationships we're looking for.

For a symmetric matrix to be invertible, it needs to be positive definite, which can be checked by seeing if both its trace and its determinant are positive.[16] Since we know the variances must be positive, the covariance matrix would be invertible if its determinant is positive.

under the MND assumption, we can infer conditional independence between two variables, given a precision matrix whose entries are zero,[1] for reasons explained above. If the two variables happen to be the contested proxy and the protected characteristic, then we decide that there is no "direct" relationship between them, therefore not constituting any discriminations. However, there is an important caveat that, due to the categorical nature of a protected characteristic variable, the MND assumption can rarely be satisfied.

3.3.2 LASSO

Applying a lasso penalty to our precision matrix Θ serves as a form of significance testing for the case where the entry in Θ between the proxy and the protected group is not zero. The objective function that GLASSO maximizes is "the penalized log-likelihood" $\log\det\hat{\Theta} - tr\left(S\hat{\Theta}\right) - \rho||\hat{\Theta}||_1$, where $\hat{\Theta}$ is the estimated sparse precision matrix, tr denoting the trace, and $||\hat{\Theta}||_1$ being the $L1$ norm that calculates the sum of the absolute values of the entries in matrix $\hat{\Theta}$.[11] The log-likelihood function is calculated based on the assumed Wishart distribution of $\hat{\Theta}$, which in turn depends on assuming the variables in the original training dataset to have MND.[5] ρ, the coefficient of the penalty term, is a hyperparameter that determines to what degree the entries in the precision matrix will be shrunk to maximize the above objective function. ρ ranges from 0 to infinity. When $\rho = 0$, the log-likelihood function would not be penalized and thus will return the same precision matrix. When ρ is infinite, the objective function would be infinitely penalized, forcing all entries in $\hat{\Theta}$ to be reduced to zero.

The stronger the conditional correlation a pair of variables have, the bigger the value of ρ required for this entry to be eliminated. Considering the effect that a change of ρ would have on the elimination process, our method creates plots for increasing ρ values to demonstrate how relationships are shrunk to zero. In our plots, the edges represent nonzero entries in the estimated precision matrix, and the nodes represent the variables. As ρ grows, we expect our estimated precision matrix to be sparser, thus having less edges in the graphs.

Employing GLASSO as a significance test of "direct" relationships between the proxy of interest and a protected characteristic, we observe the relative position of the graph in which the edge between the node of a contested proxy and the node of the protected group first got eliminated. The later this elimination first happens, the stronger the conditional correlation between two variables is. For future works, we need to create

an index to standardize this "relative position" and to determine a sensible threshold as a decision rule for significant conditional correlation.

Chapter 4

Data

To demonstrate and assess the methods listed above, we simulated two main datasets with a sample size of 800. One dataset is biased and the other is unbiased. Both datasets have one contested proxy and one protected group of interest. To illustrate the p-test issue, we also created a separate unbiased dataset with a sample size of $450,000$. To put into context, we replicated a scenario of employing bodyguards, where the target is *strength*, the contested proxy is *height*, and the protected group of interest is *gender*.[3] All biasedness and unbiasedness we refer to are in terms of the "direct" relationship between *gender* and *height*. We also created some background variables: *householdIncome*, *age*, *edu*, and *race*, all of which contributed to variations in *height*, either by appearing in the formula for constructing *height*, or through the correlations with the terms that appeared in the formula.

Recognizing that the advanced predictive models used in the real world rarely renders a clean decision-making criterion, the height cutoff, we will not incorporate any height cutoffs determined by specific algorithms trained on our simulated data. Instead, we will treat *height* as a continuous variable.

The replication material for simulating the datasets and for applying the three methods are available.

4.1 Variables

We start with simulating two *strength* distributions, one for males and the other for females. The mean strength for males is higher, and males are more represented in the dataset by having more observations labeled as males. While gender, named as the g_binary variable, was not explicitly created, we coded all other variables separately for the male group and the female group, thereby obtaining g_binary through combining the rows of the two datasets. With a different *strength* between two genders, we create correlations between the target variable and the protected characteristic.

For background variables, we created $householdIncome$ and *age* to be two random variables with a normal distribution and a Poisson distribution, respectively. To acknowledge the strong relationships between $householdIncome$ and *edu* and between $householdIncome$ and *race* due to social mechanisms at work as well as for historical reasons, we created *edu* and *race* using formulas containing the random variable $householdIncome$, thereby having our background variables correlated to each other. To create some correlations between g_binary and all background variables, the functions of those background variables for males and females are slightly different. As our definition of unbiasedness has nothing to do with "indirect" relationships, g_binary is constructed to be correlating with all non-height variables to better mimic the real world and to add difficulty in extracting "direct" relationships. The functions also always include random noise to avoid perfect correlations and to make our simulations more realistic. We will explain below how we constructed *height*, our contested proxy variable, differently under our definitions of biased and unbiased datasets.

4.2 Biased Data Simulation

The conception of being biased or not against gender is purely based on the construction of *height*, which is our proxy to *strength*. Again, the disparate treatment allows

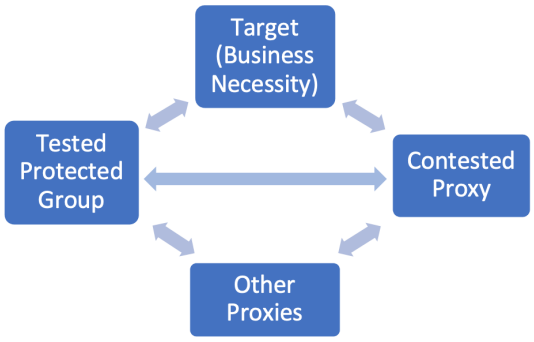

Figure 4.1: "Direct" relationship exists

us to end up employing more males than females if more males indeed have qualified level of strength. However, the disparate impact, which we will call bias, refers to the case where *height* systematically disqualifies the otherwise qualified females more than qualified males, because a female can be shorter than a male even when they have the same strengths. Since our definition of bias is to have significant "direct" relationships between $g_binary$and *height*, we instilled such "direct" correlations by encoding heights differently for different gender groups. More specifically, *height* is created with formulas containing *strength*, *age*, and *race*, and female heights and male heights have different formulas due to different coefficients. The difference in *height* formulas is the "direct" correlation between gender and height, thus bias, that we are trying to detect. See Figure 4.1.

4.3 Unbiased Data Simulation

Unbiasedness does not mean hiring the same number of males and females, or that those two groups have the same level of strength. Rather, it simply refers to a lack of "direct" relationship between g_binary and *height*. g_binary and *height* can still be correlated, even significantly correlated, through the pathway of *strength* or the background variables. The pathways have been set up in both biased and unbiased

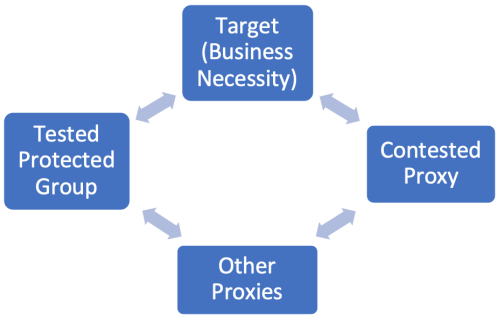

Figure 4.2: Lack of "direct" relationship

datasets when we have g_binary to be correlating with all non-height variables. A lack of "direct" relationship means that, two individuals, when having the same strengths and same background variables, wouldn't significantly differ in height when their genders are different. With this conception in mind, we constructed *height* using the same function across genders. Even though *height* still has *strength*, *age*, and *race* in its formulas, thus allowing g_binary to correlate with *height* indirectly through the pathways, we know that *height* is an unbiased proxy against gender because they are not "directly" correlated as male heights and female heights were constructed in the exact same way. See Figure 4.2.

Chapter 5

Results

5.1 IAT

5.1.1 Biased Dataset

Since we constructed our *height* in different ways for different genders, we expect a significant α_1, the coefficient of gender in the second step of the IAT.

First step: See Table 5.1.

Second step: See Table 5.2.

Table 5.1: Summary statistics for first step of IAT in biased dataset

Variable	Estimate	P-value	Significance
$strength$	0.91522	$< 2e-16$	yes

We see that, in our replication of the IAT, we get our expected result of a significant

Table 5.2: Summary statistics for second step of IAT in biased dataset

Variable	Estimate	P-value	Significance
g_binary	-5.75418	$< 2e-16$	yes

α_1. We record the result of $\alpha_0 = 0.91522$ and $\alpha_1 = -5.75418$ for future comparison

with the betas calculated using the MLR, where the background variables will be controlled.

5.1.2 Unbiased Dataset

To see the issue of a large sample size, we created an unbiased dataset with a total of 450000 observations. With such a large sample, we expect the coefficient for g_binary will be significant, even though we exactly replicated the IAT.

As shown in Table 5.3, we indeed obtained a significant coefficient, whose p values is extremely small, even though there aren't any "direct" relationships between gender and height by construction.

We now turn to another unbiased dataset with a sample size of 800. While we know

Table 5.3: Summary statistics for second step of IAT with large sample size in unbiased dataset

Variable	P-value	Significance
g_binary	$< 2e-16$	yes

that the true α_1 equals zero, we expect the coefficient of g_binary to be significant, due to a failure to control for other proxies through which gender correlates with height.

First step: see Table 5.4

Second step: see Table 5.5

Table 5.4: Summary statistics for first step of IAT in unbiased dataset

Variable	Estimate	P-value	Significance
$strength$	0.662335	$< 2e-16$	yes

Indeed, the coefficient between gender and height is deemed as significant by the IAT, even though there aren't any "direct" correlations. Having witnessed the problems with the IAT, we now turn to our first proposed method of using the MLR.

Table 5.5: Summary statistics for second step of IAT in unbiased dataset

Variable	Estimate	P-value	Significance
g_binary	2.04320	$< 2e-16$	yes

5.2 MLR

5.2.1 Biased Dataset

Using the same biased dataset as the IAT but controlling for more variables at the same time thus obtaining more accurate estimates of the coefficients, we expect the non-zero β_1 to be statistically significant. In addition, as explained before with the Frisch-Waugh-Lovell Theorem, we also expect $\alpha_0 \neq \beta_0$, and $\alpha_1 \neq \beta_1$. The summary statistics of MLR is in Table 5.6.

We indeed obtained a significant β_1. In addition, $\alpha_0 = 0.91522 \neq 0.8124$, and

Table 5.6: Summary statistics for MLR in biased dataset

Variable	Estimate	P-value	Significance
$strength$	0.8124	$< 2e-16$	yes
g_binary	-6.188	$< 2e-16$	yes

$\alpha_1 = -5.75418 \neq -6.188$.

5.2.2 Unbiased Dataset

We use the unbiased dataset with a sample size of 800. We expect β_1 to be insignificant. See Table 5.7.

Unexpectedly, g_binary is still deemed with a significant correlation with $height$.

Table 5.7: Summary statistics for MLR in unbiased dataset

Variable	P-value	Significance
$strength$	$< 2e-16$	yes
g_binary	$< 2e-16$	yes

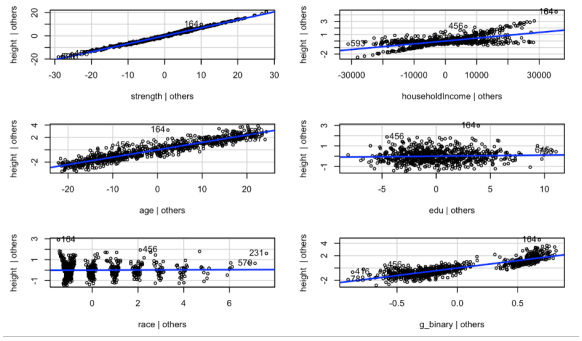

Figure 5.1: Partial correlations in unbiased dataset

Since the construction of each variable is either random or based on other variables within the dataset, there cannot be any omitted-variable bias. To better understand the "direct" relationship between g_binary and $height$, we plotted the partial correlation scatter plots for the unbiased dataset.

Unlike what we expected, the fitted line does seem to fit the g_binary and $height$ graph well. We expect the source of error to be coming from the use of p values for significance testing.

5.3 GLASSO

5.3.1 Biased Dataset

Using the same biased dataset in the MLR, we know that there is a significant conditional correlation between gender and height, since there is a true "direct" relationship. Therefore, we expect the line between g_binary and $height$ to be eliminated in a relatively late stage.

To demonstrate the weakest and the strongest penalties, our ρ values range from 0.001 to 1000. We expect that all edges would remain when $\rho = 0.001$, and all edges would be eliminated when $\rho = 1000$. In between the two extremes, we have ρ values from 0.05 to 0.325, increasing for 0.025 at a time. See Figure 5.2 after applying GLASSO onto our biased dataset. As we see from the graphs, the second to last graph still

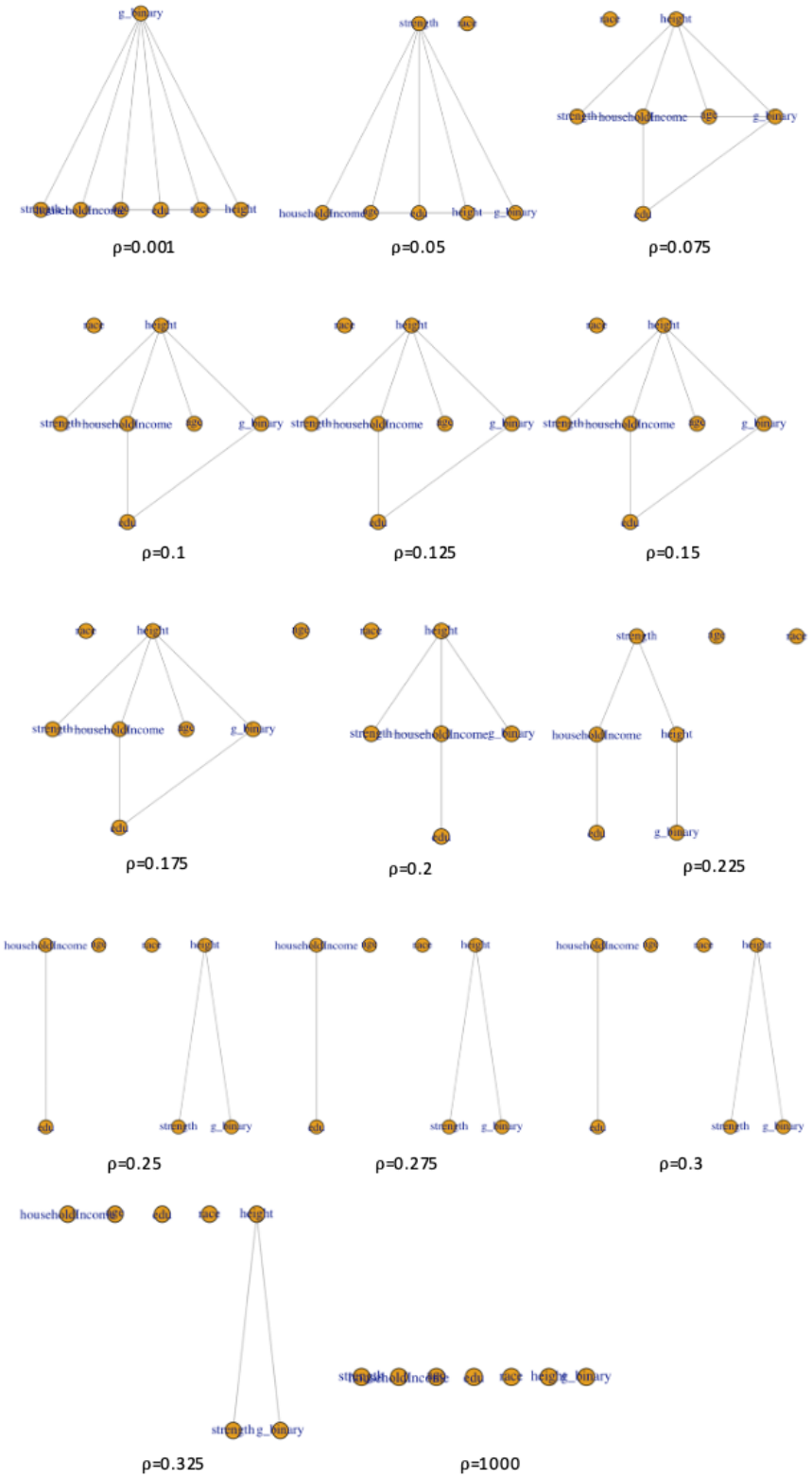

Figure 5.2: GLASSO for biased dataset

has an edge between *height* and $g_b inary$, showing us that the conditional correlation between those two variables indeed seems to be significant.

5.3.2 Unbiased Dataset

We used the unbiased dataset with 800 observations. We also used the same ρ values. Since there is no true "direct" relationship between gender and height, we expect the edge in between to be eliminated in a relatively early stage. This is because the conditional correlation between gender and height should be less significant, compared to correlations between other pairs. See Figure 5.3 for our application of GLASSO onto our unbiased dataset. Indeed, we see that the edge between *height* and $g_b inary$ got eliminated starting from the third graph, pointing us to the conclusion that the "direct" relationships between height and gender, if any, is not significant.

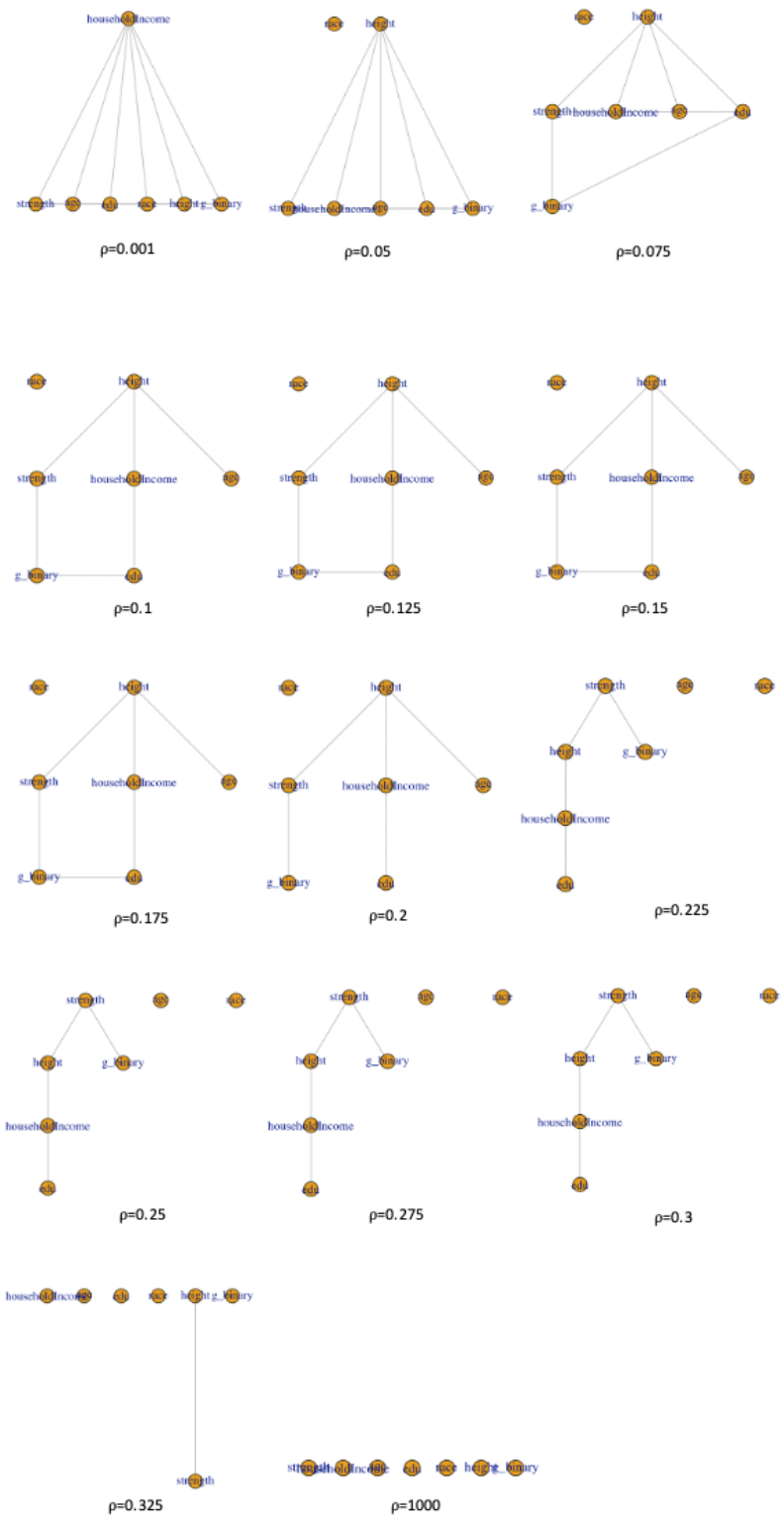

Figure 5.3: GLASSO for unbiased dataset

Chapter 6

Conclusion

While accepting Bartlett et al.'s framework of approaching the test of discrimination as differentiating the correlations between a contested proxy and a protected characteristic into allowed ones and disallowed ones, our paper had a different conception of what part of the correlation seems to be allowed by a literal read of existing policies. More specifically, we propose that all "indirect" relationships, including but not limited to the pathway through the business necessity variable (target), should all be excluded for us to find the exact "direct" correlation. This is not only because a real-world system of non-target variables usually have strong interrelationships, leading the IAT to deem all proxies inappropriate, but also because it seems wrong to agree with the propositions specifically accusing the use of height to be discriminatory against females when, say, height in fact only discriminates against the poor, given this example's imagined high correlation between gender and household income. Based on our conception of "direct" correlations, we require all rest of variables to be controlled. We proposed to use MLR and GLASSO to allow us to control for more than one variable.

We simulated datasets, one biased and the other unbiased, where bias is defined as having "direct" correlations between height and gender. Based on our results an-

alyzing those data, we see that both IAT and MLR tend to categorize height as discriminatory against gender, whether or not height being actually biased. The result obtained using the GLASSO method is completely congruent to our expectations, and thus seems to be a more desired method compared to the IAT and MLR. However, GLASSO also has a drawback of being dependent on ρ values.

See Table 6.1 for a summary of how the three methods worked.

Table 6.1: Summary of how well the three methods worked

IAT	MLR	GLASSO
expectedly accused an unbiased proxy to be biased	also not correctly classifying unbiased proxy	seems to work well, though heavily dependent on rho

Chapter 7

Future Works

While the GLASSO based method works better compared to IAT and MLR, we still need to create a standardized index for this method and determine a threshold of the index number as our decision-making rule of the significance test. One way is to extract the absolute value of the entry of interest in the penalized precision matrix and plot them against incrementally increasing values of ρ. This will give us a monotonically decreasing function because the absolute value of the entry of interest in the estimated precision matrix is always decreasing. The index can come from the integration of this function, meaning to calculate the area under the function, whose $x - axis$ is the ρ values and $y - axis$ is the absolute values of the entry of interest in the estimated precision matrix. Yet, to standardize this index, we need to investigate more closely how this index differs across different datasets. For example, it's possible to have an entry of interest to start relatively small in magnitude compared to other precision matrices calculated based on other training datasets, and yet stays non-zero even with a big ρ value. Such a case would result into a relatively small area after integration, compared to other datasets, making it difficult to set a threshold of the area under the curve.

Another concern is that, due to the nature of protected groups usually being cate-

gorical variables, and due to the skewed distribution with a long tail commonly seen in background variables, the multivariate-normal-distribution assumption can almost never be fulfilled. However, this MND assumption is vital for GLASSO to work since it is required both for the conditional independence inference and for the calculation of the loss function (negative of log likelihood function) under the Wishart distribution assumption. We need to investigate statistical methods that don't require an MND distribution in inferring conditional independence, as well as to find generalized log likelihood functions that lifts the MND assumption.

The third technical concern is that there may be omitted variables in the training data when we apply our methods to real-world datasets. Omitting important variables, thus not controlling for the significant "indirect" correlations, is harmful to the subsequent analysis on "direct" relationships, as they would include correlations through the omitted pathways. Unbiased predictors may be excluded in subsequent training process as a result.

Finally, we also need to acknowledge the conceptual difficulties with protected characteristics. Our paper only limits its analysis on one specific group, gender. However, not only are there many other protected groups, such as race and ethnicity, nationality, sexual orientation, etc., but there are also combinations of those protected features. By controlling for all other variables to extract conditional correlations, both MLR and GLASSO are unable to account for the social/ legal problem of "intersectionality," meaning the potential discriminations or other difficulties faced by, say, an African American woman. In a scenario where enough African American males and Caucasian females are hired while none of the qualified African American females were, MLR and GLASSO are bound to fail to detect this more complicated discrimination.

Chapter 8

Discussion

Our method is based on allowing "disparate treatment," which can be ethically and socially controversial due to the practical issue of "different legally protected groups" having "different base rates" regarding qualifications.[4] In reality, this difference leads to the incompatibility between unbiasedness/ accuracy and fairness of the outcome.[4] To achieve statistical fairness, the algorithm must be made biased to compensate for differences across groups.[4] The tradeoff between accuracy and fairness makes an unbiased algorithmic decision that this essay intends to promote to disproportionately filter out minorities, even though the selection process was based solely on qualifications. The unfair outcome can be deemed unethical especially after taking systemic racism/ sexism, etc. into consideration, where the protected groups are de facto lack of opportunities to develop such qualifications.[12] The result of not being selected, though by employing the algorithmic tools appropriately, will in turn strip more opportunities away from the minorities and thus exacerbate the existing structural racism/ sexism, etc..

Potential solutions include the notion of "algorithmic reparation" that intends for a fair algorithm, whether through adjusting weights in favor of minorities or "omitting and eradicating machine learning systems" for certain areas of application.[7] How-

ever, by aiming for a fair outcome (broadly defined), the purpose of using machine learning to eliminate human biases seems to be defeated since the bias for minorities would be manually instilled.

We propose to apply T.M. Scanlon's "justification for inequality"[18] that defends the use of unbiased algorithms in the selection stage that discriminate by qualifications. Scanlon's arguments involve both an accurate selection test, referred to as "procedural fairness," where "the process through which... others [being selected]... was procedurally fair," and the development opportunities, called "substantive opportunity," where everyone have the "means to do better" at developing their qualifications before the selection stage.[18] This framework defines the notion of equal opportunities into two separate parts, namely, the selection and the development of merits, and thus transforms the unbiasedness-and-fairness-tradeoff problem to be no longer a zero-sum game. An unbiased algorithm would ensure "procedural fairness," while social/ political interventions would cover the wider, background-related concern of "substantive opportunity."

Appendix A

IAT Regression Results

```
Call:
lm(formula = height ~ strength, data = cut_employee)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1042 -2.5862  0.4193  2.5013  8.2949

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.21384    0.81208   5.189 2.68e-07 ***
strength     0.91522    0.01215  75.339  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.199 on 798 degrees of freedom
Multiple R-squared:  0.8767,    Adjusted R-squared:  0.8766
F-statistic:  5676 on 1 and 798 DF,  p-value: < 2.2e-16
```

Figure A.1: Summary for first step IAT, biased

```
Call:
lm(formula = resid ~ g_binary, data = cut_employee)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6131 -1.3449 -0.1628  1.2295  6.4535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84134    0.07436   24.76   <2e-16 ***
g_binary    -5.75418    0.13145  -43.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.734 on 798 degrees of freedom
Multiple R-squared:  0.706,     Adjusted R-squared:  0.7056
F-statistic:  1916 on 1 and 798 DF,  p-value: < 2.2e-16
```

Figure A.2: Summary for second step IAT, biased

```
Call:
lm(formula = resid ~ g_binary, data = cut_employee)

Residuals:
   Min     1Q Median     3Q    Max
-1.807 -1.375 -0.549  0.379 39.994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.389125   0.003942  -98.71   <2e-16 ***
g_binary     1.167374   0.006828  170.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.159 on 449998 degrees of freedom
Multiple R-squared:  0.061, Adjusted R-squared:  0.06099
F-statistic: 2.923e+04 on 1 and 449998 DF,  p-value: < 2.2e-16
```

Figure A.3: Summary for second step IAT, unbiased, large sample size

```
Call:
lm(formula = height ~ strength, data = cut_employee)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4581 -1.5845 -0.0357  1.3564  7.1469

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.579618   0.495951   29.40   <2e-16 ***
strength     0.662335   0.007419   89.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.953 on 798 degrees of freedom
Multiple R-squared:  0.909,     Adjusted R-squared:  0.9089
F-statistic:  7970 on 1 and 798 DF,  p-value: < 2.2e-16
```

Figure A.4: Summary for first step IAT, unbiased

```
Call:
lm(formula = resid ~ g_binary, data = cut_employee)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6969 -1.4448  0.0148  1.3200  5.7575

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.65383    0.07308  -8.947   <2e-16 ***
g_binary     2.04320    0.12918  15.816   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.704 on 798 degrees of freedom
Multiple R-squared:  0.2387,    Adjusted R-squared:  0.2377
F-statistic: 250.2 on 1 and 798 DF,  p-value: < 2.2e-16
```

Figure A.5: Summary for second step IAT, unbiased

Appendix B

MLR Results

```
Call:
lm(formula = height ~ ., data = mlr_employee)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89707 -0.27005 -0.07324  0.14183  2.16125

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      7.094e+00  1.484e-01   47.819   <2e-16 ***
strength         8.124e-01  1.738e-03  467.289   <2e-16 ***
householdIncome  2.068e-05  1.388e-06   14.897   <2e-16 ***
age              1.137e-01  1.199e-03   94.880   <2e-16 ***
edu              6.245e-03  4.891e-03    1.277    0.202
race            -1.317e-03  9.309e-03   -0.141    0.888
g_binary        -6.188e+00  3.447e-02 -179.513   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.421 on 793 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9979
F-statistic: 6.214e+04 on 6 and 793 DF,  p-value: < 2.2e-16
```

Figure B.1: Summary for MLR, biased

```
Call:
lm(formula = height ~ ., data = mlr_employee)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44492 -0.39415 -0.04959  0.34889  2.91220

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.123e+00  2.095e-01  19.677   <2e-16 ***
strength         6.875e-01  2.455e-03 279.980   <2e-16 ***
householdIncome 4.367e-05  1.960e-06  22.278   <2e-16 ***
age              1.200e-01  1.693e-03  70.853   <2e-16 ***
edu              1.067e-02  6.908e-03   1.545    0.123
race             6.688e-03  1.315e-02   0.509    0.611
g_binary         2.418e+00  4.869e-02  49.661   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5946 on 793 degrees of freedom
Multiple R-squared:  0.9916,    Adjusted R-squared:  0.9916
F-statistic: 1.564e+04 on 6 and 793 DF,  p-value: < 2.2e-16
```

Figure B.2: Summary for MLR, biased

Bibliography

[1] Ritei Shibata Baba, Kunihiro and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian New Zealand Journal of Statistics*, 2004.

[2] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *SSRN Electronic Journal*, 2016.

[3] Adair Morse Nancy Wallace Bartlett, Robert P and Richard Stanton. Algorithmic accountability: A legal and economic framework. *n.d.*, n.d.

[4] Hoda Heidari Shahin Jabbari Michael Kearns Berk, Richard and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv*, 2017.

[5] Taras Bodnar and Yarema Okhrin. Properties of the singular, inverse and generalized inverse partitioned wishart distributions. *Journal of Multivariate Analysis*, 2008.

[6] Po C. Why does inversion of a covariance matrix yield partial correlations between random variables? *Cross Validated*, 2017.

[7] Apryl Williams Davis, Jenny L. and Michael W. Yang. Algorithmic reparation. *Big Data Society*, 2021.

[8] Eugene Demidenko. The p -value you can't buy. *The American Statistician*, 2016.

[9] US EEOC. 3. who is protected from employment discrimination? *n.d.*, n.d.

[10] Melanie Evans and Anna Wilde Mathews. Researchers find racial bias in hospital algorithm. *Wall Street Journal*, 2019.

[11] Trevor Hastie Friedman, Jerome and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

[12] Simone L. Hardeman-Jones Hardeman, Rachel R. and Eduardo M. Medina. Fighting for america's paradise: The struggle against structural racism. *Journal of Health Politics*, 2021.

[13] The White House. Algorithmic discrimination protections — ostp. *n.d.*, n.d.

[14] n.d. Odds are, it's wrong. *n.d.*, 2010.

[15] n.d. Fwltheorem. *n.d.*, n.d..

[16] n.d. M2-unconstrained. *n.d.*, n.d..

[17] Ricky Camplain Pro, George and Charles H. Lea III. The competing effects of racial discrimination and racial identity on the predicted number of days incarcerated in the us: A national profile of black, latino/latina, and american indian/alaska native populations. *PLoS ONE*, 2022.

[18] Thomas Scanlon. Why does inequality matter? *Oxford University Press*, 2018.