Independent Study: Debiasing Learned Representations

Steven Beattie

STATS 489, Winter 2019

1 – Background on Previous Work

The Orthogonal to Groups (OG) method proposed in Aliverti et al. (2018) (**1**) is a data preprocessing method aimed at producing predictive model output which is not only as accurate as possible, but also comparatively unbiased with respect to some set of group features. Through adjusting each predictor to be orthogonal to a specific set of group features, Aliverti et al. seek to create a matrix of predictors that, when used in a model, will produce predictive results that are orthogonal to the set of group features as well. Orthogonality is desirable because, if two predictors are orthogonal to one another, their values (and indeed the predictors overall) are said to be uncorrelated. The idea is that, if predictors in the dataset are uncorrelated with the set of group predictors the predictive results generated with the dataset will be uncorrelated with the group predictors as well - in other words, any "influence" that group membership exerts on predictive results will be removed. The OG method constructs the adjusted matrix $\tilde{X}$ matrix as follows:

$$SVD(X) = V\Sigma U^T$$

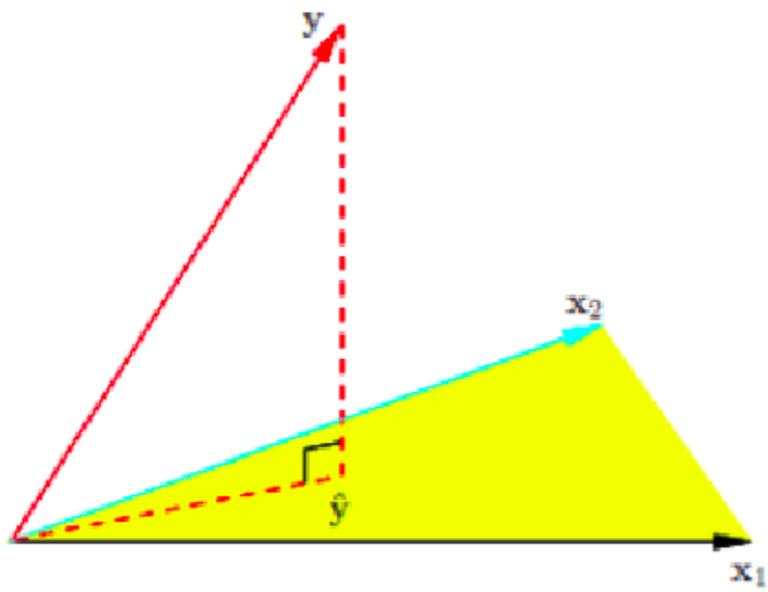
$$\tilde{X} = V_k^* \tilde{\Sigma}_k U^T = \tilde{S}U^T$$

Where k is the desired new rank of $\tilde{X}$, S is $n \times k$, U is $p \times k$, and $\tilde{S}U^T$ solves

$$\arg\min_{S,U} \|X - SU^T\|_F^2 \text{ subject to } \langle SU^T, Z\rangle = 0$$

($Z \subset X$ is the matrix of "protected" group memberships for debiasing)

The Eckhart-Young theorem[2] says that a rank-reduced matrix which has minimized distance (in terms of Frobenius norm) from the unadjusted matrix can be found through an adjusted singular value decomposition. According to the theorem, the only adjustment required is hard thresholding on the values of Σ, so that only the k largest values remain (where k is the desired reduced rank) in $\tilde{\Sigma}$. The authors make use of the guarantee given in the theorem, to reconstruct a rank-reduced $\tilde{X}$ matrix, from the singular value decomposition of X. They also demonstrate that the distance between X and $\tilde{X}$ is still minimized, even under the additional OG (orthogonality) constraint that they impose.



To implement the OG (orthogonality) constraint, each predictor in $S = V\tilde{\Sigma}$. is taken individually and regressed on Z.[1] As shown by the geometry of least squares this regression gives the projection of column S_i into the subspace spanned by the columns of Z; this is the value of the proportion of each value in S_i which is influenced by the group memberships encoded in Z. To remove this influence from the predictor S_i (i.e. "orthogonalize" S_i to Z), the projection is simply subtracted from X_i. This process is repeated for each $\forall S_i \in S$.

The (potentially rank-reduced) $\tilde{X}$ is constructed using S and U^T. The OG method utilizes the fact that, if we have a prediction $\hat{y}(\tilde{x})$ which is a linear function of $\tilde{x}$, then $cov(\hat{y}, Z$ follows if $cov(\tilde{X}, Z)$. In this way, they have created a technique to eliminate predictive bias (with respect to some group predictor columns Z) which works independently of any model. The guarantees that they provide for minimized distance between X and $\tilde{X}$ only hold for linear models, however the authors demonstrate that the method still can yield significant reductions in bias when used in non-linear models as well.

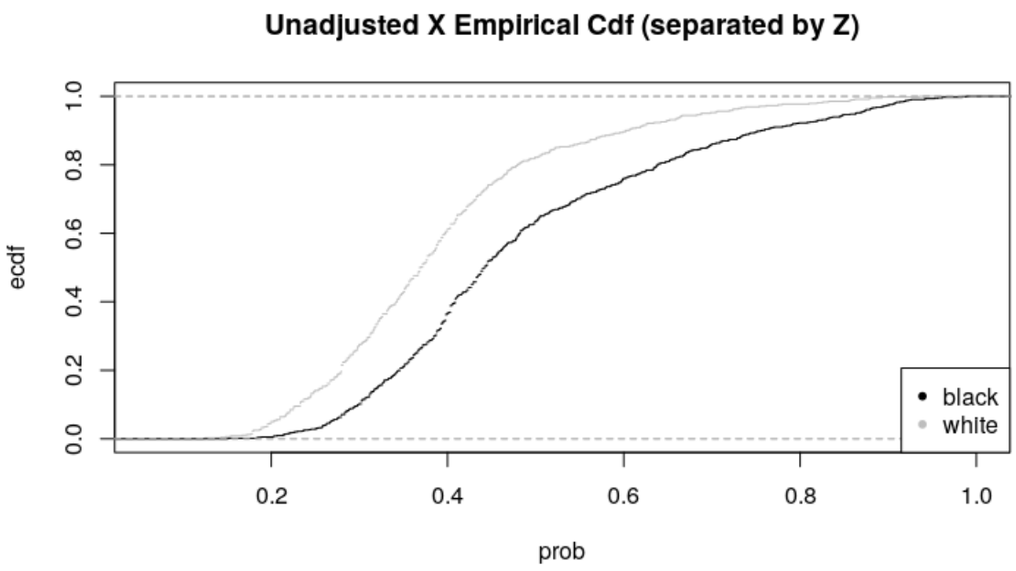
2 – Reconstructing the Previous Work

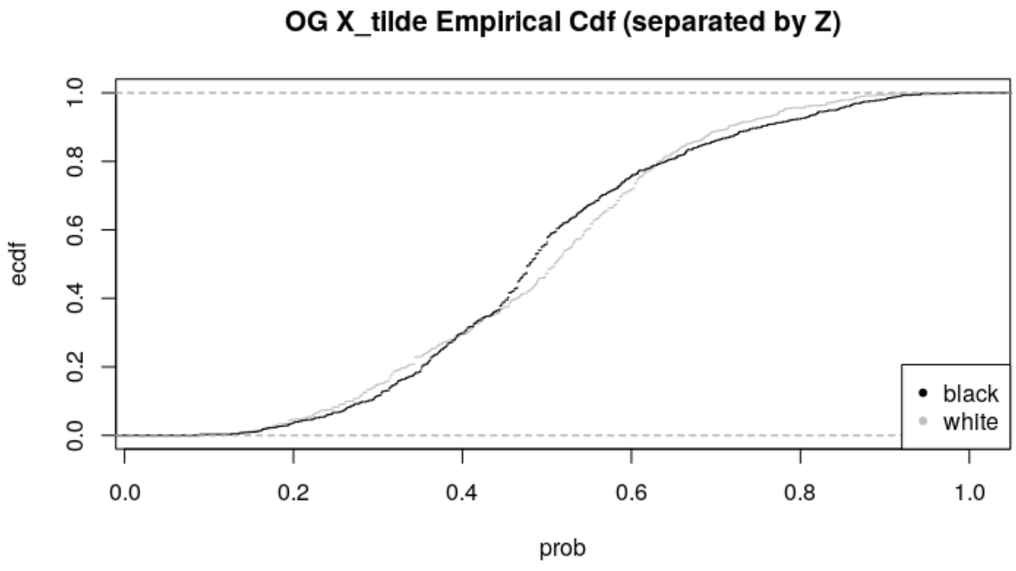
The first steps of this project involved implementing the OG method in R, and then generating $\tilde{X}$ that allow us to reproduce the OG method results reported by Aliverti et al. To generate these results, a subset of the "COMPAS dataset" is used[3]. ProPublica found that a popular predictive model being used to predict recidivism risk among defendants accused (or convicted) of criminal offenses is actually quite biased with respect to race[4]. The COMPAS dataset (colloquially named for the biased algorithm its data is sampled from) was released along with these findings, and has been cited many times in work on algorithmic fairness. In this work and the previous work, Z is an Nx1 matrix (where N = rows in COMPAS data) that corresponds to the race of the defendant, and the predictive target is two-year recidivism; whether or not a defendant will be arrested again within two years for another crime, if released from jail.

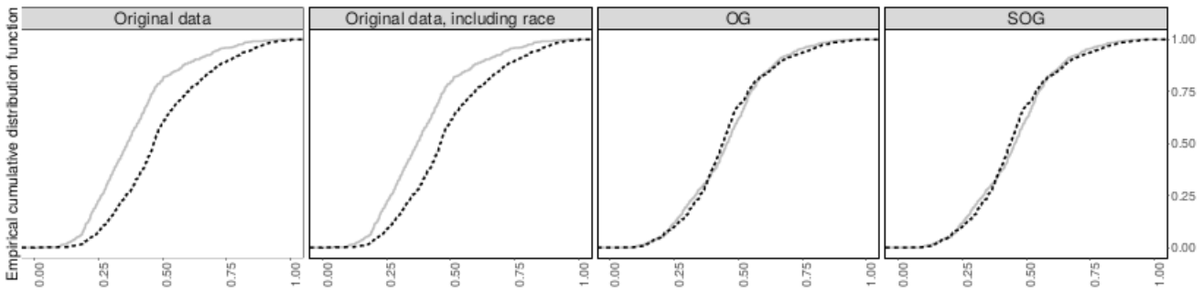
To begin, we loaded and cleaned the COMPAS data for use in an R data frame, and initially used a subset of the classifications for race. After defining the relevant parameters and creating a simplified interaction matrix to mirror the one used in the Aliverti et al. paper, we generated an unadjusted X matrix (though svd, rank reduction, and reconstruction; no orthogonalization), and an $\tilde{X}$ matrix using the OG method outlined above.

For testing, we created a simple random split in the data - first 80% for training, last 20% for testing - and fit a logistic regression model using X_{unadj} and $\tilde{X}$. In the model, we used the variables age, priors_count, juv_other_count, juv_fel_count, juv_misd_count, sex, race, is_recid, and is_violent_recid.

Next we separated the predicted class probabilities (for "will reoffend: yes/no") on the basis of race (our one-dimensional Z matrix), and plotted the empirical CDF of each set of probabilities. The plot for X_{unadj} showed the same distinct patterns of separation between the set of probabilities for each race as was seen in the previous work. Our plot of the probabilities generated using $\tilde{X}$ showed that the separation between the two sets of probabilities had been largely eliminated, again mirroring the previous work.







To confirm that our method was in fact reproducing the previous results, we generated error statistics for the predictions from the models trained on X_{unadj} and $\tilde{X}$, again separated by race in order to show the disparity in prediction (or lack thereof). Our results again mirrored those seen in the previous work. We looked at True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Prediction Accuracy, Classification Error Rate, and Area Under Curve. These results are reported in the results section, alongside the performance of the proposed "treatment effect" adjustment method that was described and implemented (in a very basic "alpha or beta" version form) for this project towards the end of the semester *(described in detail later in this report)*.

3 – Review of Previous Work; Describing an Improved Method

At this point, we began constructing an improved method of generating $\tilde{X}$ which addresses two important issues that we found with the OG procedure. First, orthogonalizing X with respect to the subspace spanned by Z doesn't actually make sense - at least in the context of racial bias, though possibly more generally as well. The projection onto the subspace spanned by Z should be thought of as the "collective influence" that race (in this case) has on the columns of X. Thus, when we remove this projection from X, we are trying to somehow remove "the influence of race" from any predictions generated by X, so that the predictions are unbiased with respect to race.

```
Aliverti OG
_____________

Decompose X by SVD(X)

Hard-threshold Sigma, keeping k largest values

Reconstruct rank-reduced X

Get coefficients Beta from X ~ Z

Debias X:

    return X_tilde = X - [ Z Beta^(T) ]
```

What actually happens though, is that the projection measures the correlation between $\hat{y}$ and *all* of the subspace spanned by Z; in other words, the correlation between $\hat{y}$ and all of the group designations in Z, simultaneously. In general, when there is more than one z_i which exerts influence on $\hat{y}$ through the columns of X, these quantities projected out of the columns of X are aggregate reflections of the effect that *all* levels of Z, as a whole. In the case where there are only two levels of Z, this correlation can be thought of as corresponding to the differential effect that the distinct levels of Z have on $\hat{y}$ - and the removal of (some amount of) this correlation is indeed what we want. However, when $\|Z\| > 2$, this correlation is best thought of as an "aggregate measure of the effect of race" on $\hat{y}$, which does a poor job of reflecting the bias corresponding to each individual z_i.

For example, in the context of our racial bias problem, we want to remove the predictive bias which can be attributed to having $z =$ "$AfricanAmerican$" - not some aggregate value of the influence of race in general. Further on this note, when $\|Z\| > 2$, it seems absurd that this same aggregate value should be removed from all observations without regard to their z_i. If one can be developed, a method should be preferred which is able to remove the "influence of being African American" from observations where $z =$ "$AfricanAmerican$", and so on; an aggregate level of racial bias is somewhat nebulous, and far less appropriate in this context than using a reflection of the actual influence that the specific race of the observation has on prediction.

Other than the lack of ability to handle more than two categories in Z effectively, the other significant issue with the OG procedure is that it essentially interprets correlation as causation. It always removes 100% of the measure of bias calculated, from the predictors X_i. We need to consider the fact that Z may justifiably share some measure of correlation in the data with the other columns of X, or with the predictive output overall [1]. If this happens to be the

[1] In the case that X and an accurate $\hat{y}$ share correlation with Z, though, we need to be even more careful to remember that correlation

case, removing the entirety of the correlation that they share, through removing $proj(X_i, Z)$ from X_i, unnecessarily discards important information. We claim that the actual point of interest in debiasing representations is quantifying and removing the causal influence of Z on X and/or $\hat{y}$.

4 – Developing an Improved Method

4.1 First Approach

Our first approach attempted to distinguish between correlational and causal effects on prediction. Upon analysis, we noticed that our second issue with the OG procedure can be described mathematically. From (**1**), we can see

$$\begin{aligned}\tilde{X} &= (I_n - P_z)X \\ &= X - Z(Z^TZ)^{-1}Z^TX.\end{aligned}$$

Consider a model $\tilde{X}$ which has a collection of "protected" features Z, and a subset of features Y which we assume to be crucial to accurate prediction (plus some noise ϵ). For example $\tilde{X} = \beta_1 Z + \beta_2 Y + \epsilon$. Now, if we use the current procedure for extracting the influence of Z, we will lose information from Y as shown by the expansion of the equation which the authors provide for $\tilde{X}$:

$$\begin{aligned}\tilde{X} &= (I_n - P_z)X \\ &= (I - P_z)Z + (I - P_z)Y \\ &= Y - \Upsilon \text{ where } \Upsilon \text{ is any column of } Y \text{ which is not orthogonal to } Z.\end{aligned}$$

While this will indeed find the components of X that are correlated with Z and eliminate them from X, extracting the influence of Z in this way leads to unnecessary loss of decision-making information. Put another way, reconsider the equation $\tilde{X} = (I_n - P_z)X$. Consider this formulation from the authors in (**5**):

$$P_zX = P_zZB^T + P_zW\Gamma^T + P_zE$$

where X and Z are as before, B and Γ are coefficient matrices, W is a representation of non-protected attributes in X, and E is a matrix of error terms. Since $P_z \perp Z$, $P_zZ = 0$. The influence of Z is eliminated, but information from W is eliminated as well, despite that W contains attributes which are perfectly acceptable to base predictive analysis on.

In this first approach, we defined $X = ZB^T + W\Gamma^T + E$, so that we could estimate B^T and simply subtract ZB^T from X. We decided to estimate B^T via linear regression, by regressing Z X. After dropping the intercept and extracting the coefficients B, we subtracted the part of X directly attributable to Z by subtracting

$$X_{adj} = X_{unadj} - \frac{ZB^T}{\sum_{i=1}^{k} B_i^2}$$

Unfortunately, despite theoretical effectiveness, we were unable to produce results distinct from X_{unadj} with this method. This approach indeed could prove to be a simple and effective way to remove only the "causal influence" attributable to Z on predictions generated using X. For now we will leave this to future research for exploration and implementation.

4.2 Background on Second Approach

For an individual observation, the "causal influence" attributable to value a for predictor X_j can be thought of as the difference between the responses (or predicted responses) where the observation has $X_{i,j} = a$ (called the "treatment value") and, alternately, $X_{i,j} = b$ for some baseline value b of predictor X_j (called the "control value"). This difference is often called an observation of the treatment effect attributable to that specific non-baseline value a of the predictor X_j. (see **5**) The idea of a treatment effect is derived from counterfactual causality.

For example, in a clinical trial, we want to quantify the effect that a drug treatment has on a patient with a disease. To quantify this effect with complete accuracy, we would need to observe the patient simultaneously receiving treatment and not receiving treatment, in order to ensure that all other variables are equal and the difference in outcome can wholly and justifiably be attributed to the treatment. Since this is impossible, clinical trials rely heavily on "matching methods." Matching methods help clinicians to assess the difference in treatment outcomes between patients who are identical as possible and (ideally) only differ in their treatment status (e.g. "did they receive the drug or a placebo").

4.3 Matching Methods

The first step to finding sets of identical or near-identical observations which differ only in our protected "treatment" attribute(s) is to define a measure of similarity for observations. The first obvious choice is "exact matching", in which a match is registered only when two observations are identical in the designated "matching variables"; this gives reliable observations of treatment effect, but may result in a small and non-comprehensive set of matches. Alternatives include the Mahalanobis distance, and the popular Propensity Score.

---

doesn't imply causation; aside from being inaccurate, such an assumption leads directly to many of the misguided arguments which still reinforce informal and institutionalized racism around the world today.

After closeness is defined between two observations, various algorithms can be used to evaluate potential matches in the dataset. There are two main types of matching method: Nearest Neighbor matching (may not use all observations), and Subclassification matching (uses all observations).

Within Nearest Neighbor, there is Greedy Matching (each treated observation is matched with its closest control observation), and Optimal Matching (treated and control are matched so that the average distance between matched observations is minimized). Standard nearest neighbor matching is done using a 1:1 treated:control ratio, but a 1:k ratio can also be used (where k is selected by the researcher). Nearest Neighbor methods can be implemented with replacement as well; in either of these cases, weighting techniques need to be used in order to account for the number of times a control observation is matched with a treated observation.

Within Subclassification, there is N-tile Grouping (observations are split into N equally-sized groups; treatment effect assessed within groups and then aggregated), and Full Matching (same as N-tile Grouping, but with an optimized number of sub-groups; also known as variable-ratio matching). In the case of Full Matching, some form of weighting technique will also need to be used.

There are other important points to consider when using matching methods, as well, to ensure that quality matches are obtained. For example, whether or not the treated and control observations have a large enough common support. For more information on these topics (and related ones), please refer to the matching methods review paper cited in (**6**) at the end of this report.

5 – Second Approach

Through the use of matching methods and the quantification of treatment effects, we can get an idea of the causal influence that a difference in Z has on predictive output. For a well-matched pair of observations which (ideally) vary only in their value of z, the "treatment effect" - difference in predictive output on the two observations - can be considered an observation of the overall "treatment effect" that exist for group membership in a non-control group. In other words, this is an observation of the unacceptable predictive variation between groups in Z. The challenge, then, is how to adjust X so that this unacceptable variation is removed from any $\hat{y}$ generated using X.

Bias refers to a form of deviation from a baseline or target value; in order to effectively remove bias, we need to be clear on where our baseline for unbiasedness is. To define this baseline, we needed to choose a race that all other races should be treated like; we chose Caucasian. The justification for this will be covered further in the "Discussion" section. This "baseline race" constitutes the control group, and each other race represents a "level" of treatment group (to be compared to the control group individually). For our example application, we've only debiased with respect to one treatment group (African American), but this same approach can be repeated for any number of other treatment levels.

This method proceeds on the assumption that, if the model used in debiasing is the same as in post-processing analysis/prediction, and we have both a treatment observation and its closest matched control observation, then debiasing of post-processing predictive output can be accomplished with an adjustment to the value of Z which counter-balances the differential in prediction between the two observations. This debiasing approach requires that Z be converted from a categorical variable to a continuous one, with $z_1 - z_0 = 1$ (more on the choice of Z later). In its current form, it requires a linear classification model - in any form, the matching and debiasing adjustment should be carried out on the same type of model (with the same parameters) as will be used after debiasing. A basic outline of the algorithm used to develop this debiasing adjustment method is shown below; more detail can be found in the attached R script for this project (inStud-wn2019.R).

```
Adjustment Method
_____________

(Within X, convert Z to a double)
(Decide on baseline/control value of Z)
(Create a binary {0.1} predictor of Z for matching)
(Set control observations' Z value to 2, set all others to 1)

Fit desired predictive model on X

For each pair of control/treated Z values:

    Obtain matched rows of X

    For each matched pair/group:

        Predict response for treated observation

        Predict response for matched control observation

        Scale treated observation's Z-value by the percent difference in predicted
        responses (treated - control)
```

In our example implementation, we used Propensity Score matching, and 1:1 Nearest Neighbor matching. For predictive modeling, we used a logistic regression model with the same predictors as used in our analysis of the OG method (age, priors_count, juv_other_count, juv_fel_count, juv_misd_count, sex, race, is_recid, and is_violent_recid).

We took the same non-random training/testing split as before (80/20), and we performed analysis on the same unadjusted predictors in X, again including only observations from the COMPAS data with African American or Caucasian race. After matching and debiasing the "treated" (African American) observations, we compared the logistic model output for X_{adj} and X_{unadj} through empirical cdf plots and computed error statistics.

6 – Comparison of Results From Methods Implemented

After using the OG method on the unadjusted data, there was a reduction in classification error and the difference in classification error between black and white observations. There was a slight increase in true positive rate as well. Using the adjustment method on unadjustment data, there was an ∼0.1% increase in classification error on average. The true positive rate was consistent.

While the OG method was slightly better than the proposed adjustment method in terms of model performance, the proposed adjustment method had a noticeable edge on the OG method in terms of debiasing. For example, between unadjusted and OG-processed black and white observations, we saw a 0.46% decrease in classification error difference, a 0.7% reduction in true positive rate (dropping the difference to zero), and almost no change in the true negative rate.

On the other hand, the proposed adjustment method saw a 3.25% decrease in true negative rate with zero difference in true positive rate. Also, the difference in classification error between groups (black and white observations) decreased from 1.2% to 0.06%. By comparison, the proposed method ended up with far less separation in predictive output compared to the OG method; difference in OG classification error between groups was 0.3%. Also, the differences in true negative rate between OG and proposed method, respectively, are 2.2% and 0.34%. Notwithstanding that the proposed adjustment method may provide greater benefit after tuning (discussed later), or the mass of further development still needed to address potentially serious issues with the proposed method (also discussed later), this represents a significant improvement in debiasing ability.

Interestingly, only about ∼28% of treatment observations were adjusted in a (using the debiasing adjustment method); this accounts for only 16.8% of the total observations in matrix b (of which treated observations themselves account for ∼60%).

Note: in the attached results, **xun** is the unadjusted matrix of predictors used in conjunction with the OG method; **xog** is the matrix of predictors adjusted using the OG method; b is the matrix of predictors used in conjunction with the proposed adjustment method; a is the matrix of adjusted predictors generated using the proposed adjustment method. Both versions of the unadjusted matrix include all of the same predictors and observations - they are stored as separate matrices simply because of the difference in data structures required for the two methods. Therefore, the unadjusted and adjusted output between these two methods are indeed comparable despite being generated from different matrices.

Also note that the disparities between our cited OG method results and the OG method results cited in Aliverti et al. are due to our decision to forgo the implementation of dimension reduction for this project. The output for our implementation of the OG method using full-dimensional data still accurately reflects the characteristics of the relevant adjusted and unadjusted method results reported in Aliverti et al, as well as the trends in the differences between them.

In the interest of space, figures outlining project results are included in the Appendix at the end of this report.

7 – Discussion

We should also discuss two points of consideration which are relevant to this adjustment method. The first is the use (and choice) of a baseline value for Z; the second is a clarification of the problem that debiasing is trying to solve.

It seems that the question of baseline needs to be answered using application-specific knowledge. While taking the majority class of Z as the baseline might make sense, in many situations, this should not be considered the default choice. Two perfect examples of why can be found in our application: recidivism rates in our subset of the COMPAS data. In our subset, African American is the most frequent racial classification. Yet, as we've seen, predictive accuracy is significantly lower on the unadjusted data for African American observations compared to Caucasians; if we adjust other races' treatment (or just Caucasians, in our case) to reflect that of African American observations, the predictive accuracy overall could decrease as well. If, on the other hand, we debias with respect to Caucasians, we might avoid this potential decrease in accuracy.

Whether or not accuracy increases or decreases when using this new debiasing method will be a matter of whether or not the distribution of observations' responses between groups is similar. This adjustment method implicitly assumes that the true probability of the response for one group in Z resembles that of another group. We can safely assert that this assumption is reasonable for the true distribution of responses in our application, and that sufficient data will reflect this fact; however, in other applications, this assumption might not hold. If a similar distribution of responses between groups cannot be assumed, another debiasing method might be preferred to this one. While accuracy and unbiasedness are different goals *(discussed below)*, accuracy in predictive modeling is always desired; if this assumption on the distribution of the data is strongly violated, we can reasonably expect model accuracy to meaningfully decrease for non-control observations.

The second example of why the majority class in Z shouldn't automatically be chosen is contextual. In America, the law has long been acknowledged to be more "just" for Caucasians than for other races; indeed, at many times and in many places, it has even prioritized justice for Caucasians over that of other races. By choosing Caucasian as the baseline race, not only are we increasing model accuracy (to the extent that our distributional assumption is met), but we are also supporting the reduction of disparities in justice which have historically been tied to race.

7.1 On Fairness, Justice, and Algorithmic Decision-Making

A main motivation for interest in the area of algorithmic fairness and fairness in data analysis is injustice. Rightly so, there is significant concern about the outcomes of algorithmic and model-based assessments and decisions - and the perception of these decisions in society. To be fair, there is certainly some exception to this, but in general, scientific and computed findings are already considered to be unbiased in our society. The computational nature of these findings, presumably due to the fact that they are done by machines and not humans, are considered to be for example more objective and less political. Perceived traits like this easily give rise to a general perception that computed findings or recommendations are not only more objective and unbiased than those made by a human, but also as having some emergent trait of being more "insightful" than a human is capable of being.

While in some ways this perception is accurate, there is significant gap between perception and reality in terms of fairness. Computers don't exist independently, and didn't evolve along some evolutionary pathway parallel to humanity; we are the ones that create the machines, and we are the ones who program them. In general, we tell them how to compute their findings for us. In doing so, we make their computations a progression of the analyses that we have been doing ourselves for centuries - complete with any bias or oppressive perspective that may lie in our current understanding. In this way, the analysis performed by machines, despite the perception of impartiality, can in fact produce the same injustice or inaccuracy that we currently produce, or even expand upon it. I would even argue that computer-generated injustice is worse because of scaling (computers can simply make unfair decisions at a faster rate than we can) and because of the current public perspective of "computers" as possessing inherent impartiality and the ability to somehow provide superior insight.

In general, fairness is valued because we value justice; we value perspectives, actions, processes, etc that are not unjust. However, fairness is not justice. While it can certainly be said that unfairness constitutes injustice, to eliminate unfairness is not to eliminate injustice. Even once fairness is established, justice requires an added moral evaluation to be applied to the fair treatment that everyone within the institution is now receiving. Without fairness, we not only contend with whether or not the treatment of groups is just, but also with whether or not differential treatment of those groups is just. When fairness between groups is established, the problem of injustice collapses into one of evaluating the morality of how people in general are treated; this is a powerful simplification to be able to make.

7.2 What "Is" This Adjustment Method?

Empirically, we have shown so far that this adjustment is effective on a specific dataset when $z_1 - z_0 = 1$, under some conditions. So, what is this adjustment? Is it interpretable? Is it generalizable? Intuitively, the "treatment effect" $\hat{y_1} - \hat{y_0}$ can be thought of as an observation of the bias between a control group and a specific non-control group. In other words, an observation of the unacceptable variation between predictions from the two groups. The value of z_1 is then scaled to eliminate this variation. It is important to understand that this method has a specific kind of scenario in which its application (or application of a similar, more developed method in the future) makes sense. Treatment effects are estimated, but the application to clinical trials stops there. It should be stressed that the treatment effect observations are only used to get an idea of the unacceptable variation between observations; this debiasing adjustment method is simply an attempt at finding a ways to effectively debias based on observations of treatment effect.

In the attached PDF scans, more information is shown about what the adjusted "z_{new}" value is for a given treatment observation - and what it is not. Pathological and edge-case situations are also arrived at analytically, which show circumstances in which this debiasing adjustment may be ineffective or non-sensical. Note that, while $z_{new} = z_0$ is possible, it is actually a relatively an uncommon outcome; in general, it is untrue that $z_{new} = z_0$ or some other constant. If treatment effects are similar for treatment groups z_i and z_j where $j \neq i$, a situation where $z_{new,i} = z_{new,j}$ could possibly occur, too. For this reason, it is recommended to save a copy of the original categorical Z so that interpretability isn't lost (as reflected in the linked R code for this project).

Note also that the interpretation of the model changes from pre-adjustment to post-adjustment. A change in $\hat{y}$ is no longer associated with a corresponding one-unit change in z (holding all other variables constant), after adjustment. Instead, afterwards we can think of z_{new} as being a score which represents an observation of the treatment effect associated with group membership $z_i \neq z_0$.

7.3 Advantages for Adjustment Method

There seem to be at least a few advantages to a debiasing approach based on estimation of treatment effects. First, this adjustment method is still interpretable after adjustment. By contrast, the X matrix is uninterpretable after the OG method is applied due to the built-in dimension reduction capacity. Also, since the true group memberships Z can be retained and will still correspond to the same observations in X_{adj} as in X_{unadj}, the bias-adjustment score attribute may (with further development) provide interpretable information about group-relative bias which was latent in the original covariates, above and beyond retaining interpretability of the original covariates.

Most importantly, this debiasing adjustment method is easily extendable to situations in which X holds more than two values in Z. In this case, not only is interpretation retained, but the adjustments - which are done based

on the observed treatment effect between a single and specific z_1 and the control group z_0 - are only applied to observations which have the relevant group membership value. This benefit is enhanced by the possibility that the adjustments may more accurately reflect and eliminate only the unacceptable variation in predictions between groups.

7.4 Limitations, Conditions, and Assumptions for Adjustment Method

Due to the fact that it is a post-processing technique, debiasing may be required many times depending on the task at hand. Also, the effectiveness of the debiasing lies in the quality of matches used; accordingly, for each unique set of predictors generated from a data source (as well as for unique data sources), an optimal matching technique needs to be found in order to ensure optimal debiasing. As observations are added to a dataset, however, the same matching technique can be used and only the debiasing procedure itself needs to be repeated.

This adjustment method, in its current form, requires that *mean(Z)* $\neq 0$ and that $Z = \{z_0, z_1\} \neq \{0, 1\}$. In testing, both were shown empirically to produce effectively unchanged predictive results compared to the unadjusted data. For more information on why $Z = \{0, 1\}$ doesn't work, see attached handwritten sheet (WHICH PAGE?) Further analysis should be performed to discern the implications of this, which may lead towards a simpler and more optimized adjustment method based on treatment effect observations.

While the choice of distance metric and matching method are left up to the user, the theoretical foundation of the adjustment method assumes Exact Distance to be used on attributes other than the treatment attribute. Violations of this assumption weaken our ability to state that the difference in predictive output purely represents an observation of the unacceptable variation between two groups. At the same time, Exact Distance matching may produce fewer matches than another distance metric; since the strength of the adjustment currently relies on the number of observations available (as each observation of treatment effect is used to adjust exactly one treatment group observation), fewer matches means that fewer treatment group observations will be adjusted, and consequently the data will remained biased to a greater extent.

Part of the appeal to the Aliverti OG, and the Aliverti Sparse OG which we didn't discuss here, are that they are debiasing methods which provide dimensionality reduction as well. In this way, the adjustment method proposed in this paper is lacking; while one could easily use the same matrix decomposition and dimension reduction technique in our method as used in Aliverti's, the Aliverti approach still is able to provide accuracy guarantees even under dimensionality reduction. This is due to the way that the Frobenius norm is used in their debiasing technique (utilizes the Eckhart-Young theorem); unfortunately, we are not able to make any comparable guarantees about the accuracy of our method.

Personally, I would argue against this approach, and maintain that if debiasing is desired, it should be achieved independently of other goals; in my opinion, things like dimension reduction should be performed on data which has been already thoroughly debiased. A method which simultaneously tries to do both may not fully succeed at either one. Still, dimensionality reduction can be quite an important benefit in data analysis - as illustrated by the fact that matching observations, and predicting on $nrow(X)$ observations, are rather computationally taxing (significant lag can be noticed when executing these sections of R code in the provided script).

7.5 Immediate Problems for Adjustment Method

In order for this method to be accepted as a promising method which may benefit from further development, it should first be more clearly established that the adjustments this method utilizes do indeed reflect and coherently eliminate the unacceptable variation in predictions between groups. This seems to be the case empirically and intuitively, but a more concrete analytic case should be established to confirm this.

On this note, there is an issue of variable adjustment effectiveness due to the encoding of Z. So far, only combinations which vary by one (i.e. $z_1 - z_0 = 1$) have been tried; empirically, $Z = \{1, 2\}$ (which is the default numerical equivalent of the first two factor levels in R) and $Z = \{2, 3\}$ both give similar debiasing results. However, theoretically, since the adjustment is a percent scaling of the z_1 predictor value, the encoding could have a significant impact on adjustment effectiveness. Consider the situation when $\hat{y_1} - \hat{y_0} = 0.5$; in this case, $z_{new} = z_1 * 0.5$. With $Z = \{1, 2\}$, this results in $z_{new} = z_0 = 1$; with $Z = \{2, 3\}$, this results in $z_{new} = 1.5$. Analysis of this sensitivity could also help progress future work, too, if this method (or a similar approach) is worked on further.

Analysis of the variation in results seen with different configurations of the values of Z should also be performed. It may also be informative to see whether or not the misclassified observations using various unsuccessful configurations are the same between before-adjustment test results and after-adjustment test results; if the overall predictive ability is unchanged despite observations being adjusted differently, this could be helpful information to have.

These are the most pressing issues, in my opinion, though there may well be others; I would suggest that these and other issues with the proposed adjustment method be resolved before researchers consider investing significant resources into its development. If they cannot be resolved, or if a more efficient approach presents itself, then this method certainly should be abandoned. However, in principle, it does seem that adjusting to an extent which reflects only the *causal* influence that a specific group membership has on prediction (which "treatment effect" seems a promising way to measure) will be a necessity for any successful prediction debiasing method developed.

8 – Suggestions for Further Research

Again, if the immediate problems with this method are investigated and addressed, this method may potentially be worth developing farther. If this point is reached (i.e. if this method remains coherent and effective upon further investigation), these lines of inquiry could serve helpful in the development of this debiasing adjustment method.

The version of our debiasing algorithm which was used to generate the results cited in this paper was not optimized at all. We ensured that the treated and control groups in our data had sufficient common support, however we did not cross-validate any tuning parameters. For this specific application, better debiasing could be achieved using our proposed method with alternate choices of distance metric and matching method. If the method shows promise under further investigation, it would be worthwhile to investigate the conditions in which the method performs optimally. A first place to look in this direction is balancing the trade-off between match quality and number of matches, which was discussed earlier.

In order for the variable z_{new} to accurately reflect the stated interpretation, it is possible that an adjustment which starts out at the baseline z_0 value would be more ideal. Specifically, this suggestion is to investigate the effect of assigning z_{new} based on a modulation of the baseline z_0 value instead of the current z_1 value. Intuitively, this would allow us to more credibly claim that z_{new} is a score which represents an observation of the treatment effect associated with group membership $z_i \neq z_0$, as compared to the baseline of group membership in $z_i = z_0$.

In a linear model, just as the coefficients can scale the predictor values, the predictor values can scale the overall contribution to the model from each predictor. In the ideal (exact distance) matching scenario (where observations only differ in the value of their protected attribute z_i), the only possible source of variation in predicted response would be the value of z_i, as all other values will either be set in model fitting or consistent across all observations. However, the effectiveness of scaling with respect to this variation will depend on the size of the coefficient attached to Z in the model. In future work, it would be important to consider ways to take the value of this coefficient into account when defining values of z_{new}.

This approach also relies on a linear relationship between Z and $\hat{y}$. Further research should be done to investigate the use of estimated treatment effects to debias predictive output generated from non-linear or even non-parametric models. Also, as stated before, the approach currently assumes that $\hat{y_1}$ and $\hat{y_0}$ are both probabilities in $[0, 1]$. Since one can imagine scenarios in which debiasing in cases of continuous prediction would be desirable (such as when the output is a score of some sort), it would be worthwhile to consider if and how this method could be adjusted to be useful in such situations.

9 – Conclusion

This is a research report, but it is also an open proposal for future research. We ran out of time to go further in this semester; the next steps (outlined more thoroughly in the Discussion section) include implementing this method using exact distance and matching on all non-protected attributes. If the results of the adjustment using such an approach are consistent with the results reported here, then the steps suggested for further investigation and development might be considered. Otherwise, a lack of improvement (or worse results) will indicate that the assumptions and interpretations used thus far may be violated or inappropriate, or that this approach to debiasing may not actually be a promising one for future work (as the case may be).

In this paper, we have discussed the Orthogonal to Groups (OG) method - a current approach to debiasing predictive output which is proposed by Aliverti et al. (2018) We have also considered some drawbacks to the OG method, and some objections to its use. In response, we have explored multiple debiasing techniques which might address the drawbacks we outlined. Both methods which we have discussed require further work in order to be sure how much promise they hold, but intiutively they are both approaches that might perform well.

The debiasing adjustment method proposed is a method that seeks to remove unacceptable variations in predictive output between different protected groups through estimating and utilizing a "treatment effect." Intuitively, this approach could prove ideal for debiasing, but currently this implementation still needs significant work. It does seem to show some measure of promise, however, in both the empirical evidence gathered and the preliminary analytic work performed thus far (excerpts from which are included in Appendix).

10 – Credits and References

Credit:

Steven Beattie - literature review and analysis; method development (third method), implementation (all methods), and testing (all methods); project report authorship

Professor Yuekai Sun - feedback & review (entire project), method development (second method); practical and conceptual advising related to the project; advising for professional and academic development

Professor Emmanuel Aliverti - for graciously helping me get started in my understanding of the OG method and his group's testing method, so that I could begin to code and recreate their results for use in this project

University of Michigan Statistics Department - for allowing me this invaluable opportunity to gain exposure to the research process, and complete my senior seminar in an area of interest to me

References:

1 - Removing the influence of a group variable in high-dimensional predictive modelling. Aliverti et al., 2018. https://www.researchgate.net/publication/328429034$_{R}emoving_{t}he_{i}nfluence_{o}f_{a}g roup_{v}ariable_{i}n_{h}igh-dimensional_{p}redictive_{m}o$

2 - Singular Value Decomposition: Low Rank Approximation. https://en.wikipedia.org/wiki/Singular$_{v}alue_{d}ecompositionLow rank_{m}atrix_{a}pproximation$

3 - ProPublica COMPAS Dataset (and accompanying materials). https://github.com/propublica/compas-analysis

4 - ProPublica article giving background on COMPAS and describing their research which utilized the "COMPAS Dataset". https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

5 - Debiasing representations by removing unwanted variation due to protected attributes. Bower, Amanda; Niss, Laura; Sun, Yuekai; Vargo, Alexander. https://arxiv.org/pdf/1807.00461.pdf

6 - Matching Methods for Causal Inference: A Review and a Look Forward. Stuart, Elizabeth A. http://biostat.jhsph.edu/ estuart/Stuart10.StatSci.pdf

11 – Appendix

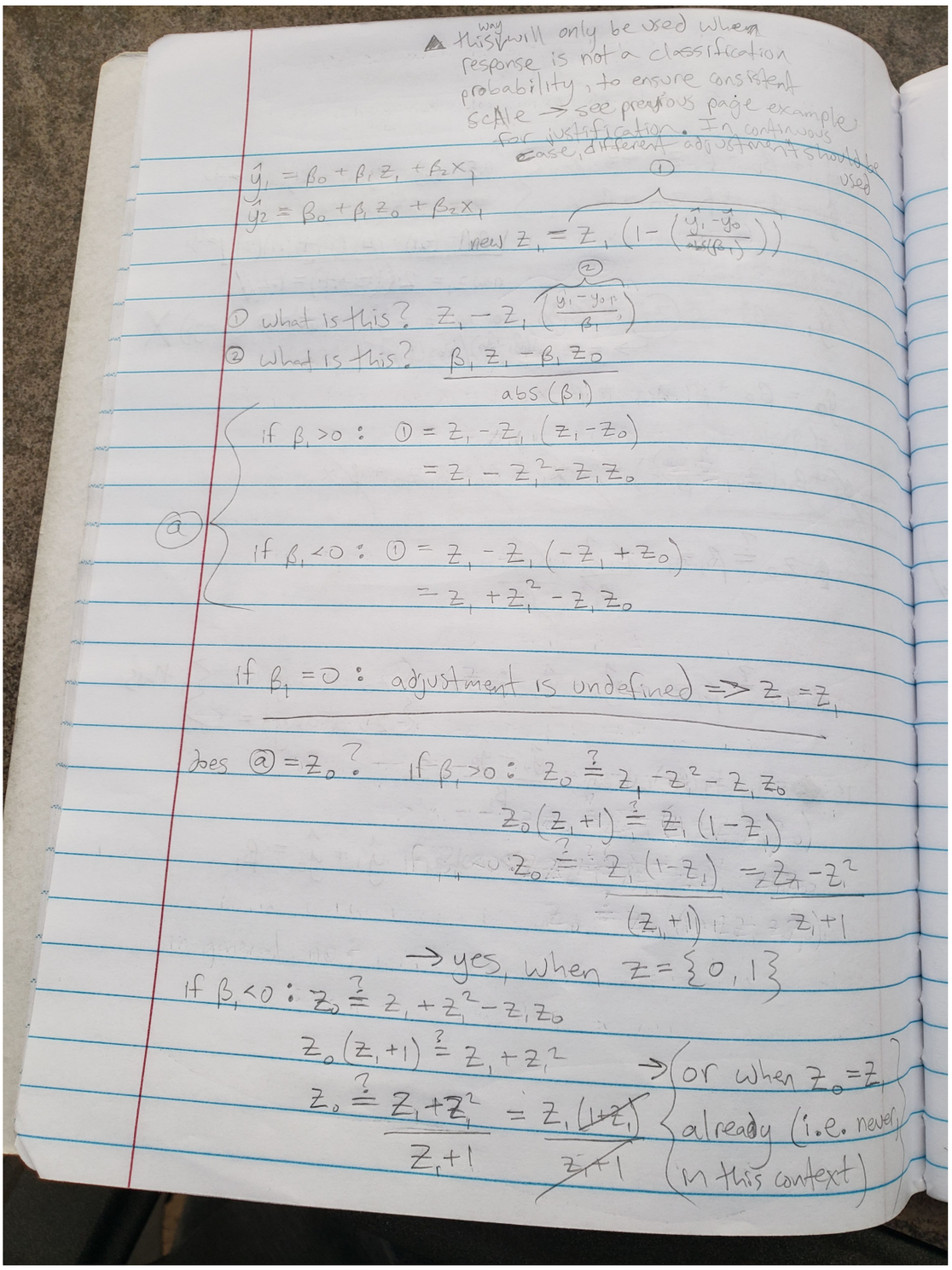
11.1 Results (continued)

```
> xun.diffs
                                [,1]         [,2]
positive (true,false)    -0.007101401  0.007101401
negative (false,true)     0.022018452 -0.022018452
(accuracy,classifError) -0.007766054  0.007766054
(null, AUC)                       NA -0.012103161
> xog.diffs
                                [,1]         [,2]
positive (true,false)     0.000000000  0.000000000
negative (false,true)     0.022018452 -0.022018452
(accuracy,classifError) -0.003143007  0.003143007
(null, AUC)                       NA -0.005437847
> b.diffs
                               [,1]        [,2]
positive (true,false)     0.00000000  0.00000000
negative (false,true)     0.03628542 -0.03628542
(accuracy,classifError) -0.01240501  0.01240501
(null, AUC)                       NA -0.01359721
> a.diffs
                                 [,1]         [,2]
positive (true,false)     0.0000000000 0.0000000000
negative (false,true)    -0.0035900946 0.0035900946
(accuracy,classifError) -0.0006777135 0.0006777135
(null, AUC)                        NA 0.0049788227
> xun.errorStats$white
                               [,1]       [,2]
positive (true,false)    0.99447514 0.005524862
negative (false,true)    0.04635762 0.953642384
(accuracy,classifError) 0.96894410 0.031055901
(null, AUC)                       NA 0.979620211
> xun.errorStats$black
                               [,1]       [,2]
positive (true,false)    0.98737374 0.01262626
negative (false,true)    0.06837607 0.93162393
(accuracy,classifError) 0.96117805 0.03882195
(null, AUC)                       NA 0.96751705
> xog.errorStats$white
                               [,1]       [,2]
positive (true,false)    1.00000000 0.00000000
negative (false,true)    0.04635762 0.95364238
(accuracy,classifError) 0.97101449 0.02898551
(null, AUC)                       NA 0.97643701
```

Table 2: Predictive performance on the COMPAS dataset for the approaches described in Section 4.2. W stands for White ethnicity, NW for the remaining ethnic groups. Metrics: Classification Error, AUC, Total positive Rate, Total Negative Rate, False negative rates, False positive rates.

		Unadjusted					Adjusted			
		OR	OR Z	OR Z_0	LASSO(u)	RF(u)	OG	SOG	LASSO	RF
CE	W	0.281	0.285	0.281	0.303	0.285	0.329	0.329	0.344	0.354
	NW	0.318	0.316	0.316	0.321	0.325	0.333	0.333	0.332	0.395
AUC	W	0.734	0.734	0.734	0.715	0.725	0.722	0.722	0.717	0.655
	NW	0.729	0.729	0.729	0.720	0.733	0.731	0.731	0.732	0.651
TPR	W	0.569	0.562	0.569	0.534	0.577	0.453	0.453	0.434	0.444
	NW	0.401	0.408	0.419	0.389	0.428	0.434	0.434	0.447	0.325
TNR	W	0.150	0.153	0.150	0.163	0.138	0.218	0.218	0.221	0.201
	NW	0.282	0.276	0.266	0.290	0.247	0.233	0.233	0.221	0.280
FNR	W	0.045	0.052	0.045	0.080	0.037	0.161	0.161	0.180	0.170
	NW	0.123	0.116	0.105	0.135	0.096	0.089	0.089	0.077	0.199
FPR	W	0.236	0.233	0.236	0.223	0.248	0.168	0.168	0.165	0.185
	NW	0.194	0.200	0.211	0.186	0.229	0.244	0.244	0.255	0.196

11.2 Misc Analysis Notes for Adjustment Method



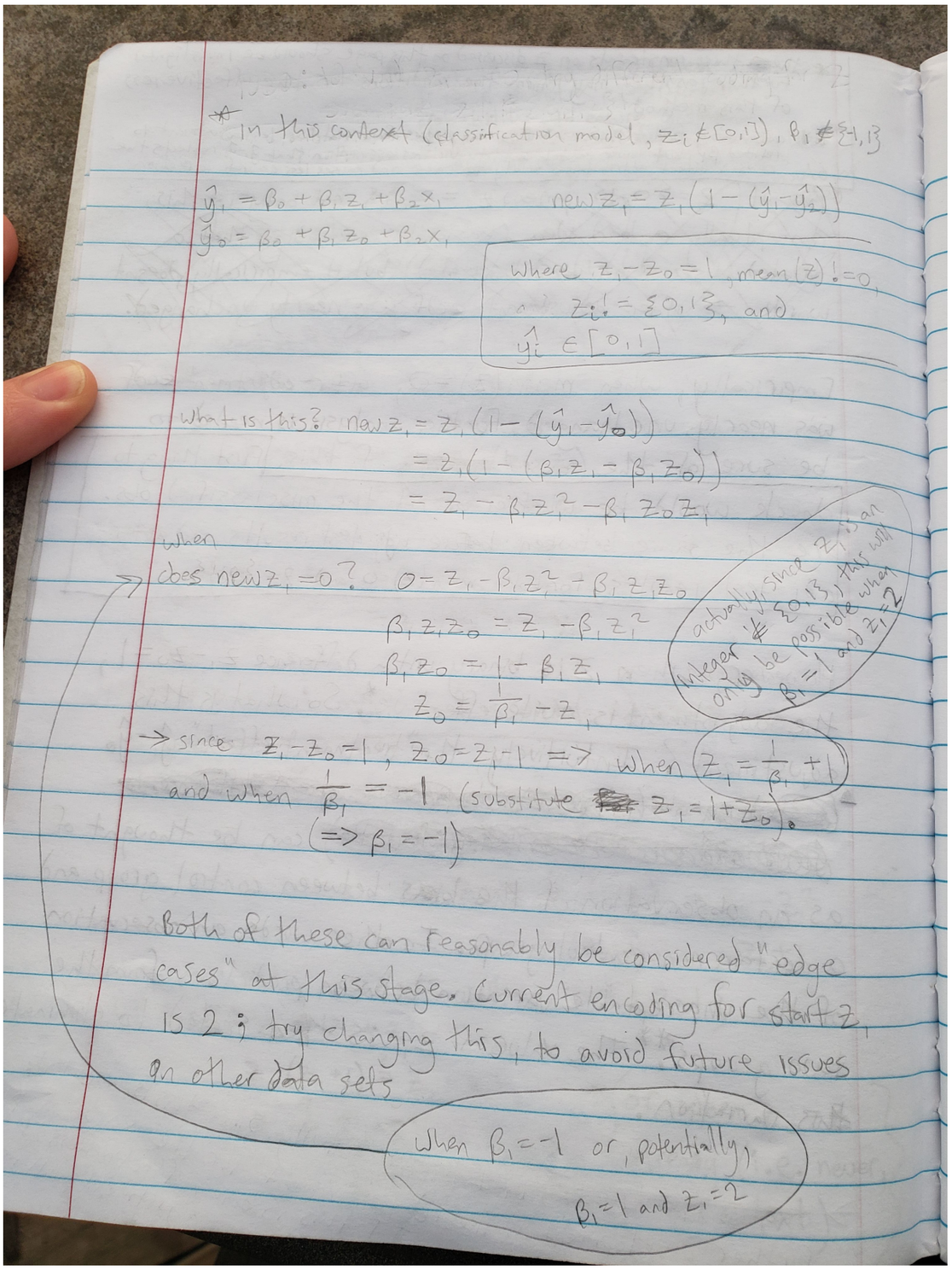
* the three constraints on z discussed on this page should be investigated further, as they may inform the conditions for the effectiveness of this method

** assuming exact distance; ED may produce fewer matches, less adjustment to refitted β, but lower match quality violates assumption that $\hat{y}_1 - \hat{y}_0$ reflects the causal effect attributable solely to race and causes less accurate correction

~~If $z = \{0, 1\}$, new $z_1 = z_0$; not only is this conceptually a bad idea (basically, pretending that a treated observation wasn't treated) but it empirically doesn't work → after-adjustment ecdf was nearly unchanged.~~

Empirically, when mean$(z) = 0$, after-adjustment ecdf was nearly unchanged. Further analysis is needed to be sure of the cause/[redacted] of this. First thing to check would be whether or not the misclassified obs. are the same between before-adj test results and after-adj test results. For both mean$(z) = 0$ and $z = \{0, 1\}$.

However, when z is binary with difference $z_1 - z_0 = 1$, the adjustment is quite effective*. So what is this adjustment? Intuitively, the "treatment effect" $\hat{y}_1 - \hat{y}_0$ [redacted] can be thought of as an observation of the bias between control group and a certain non-control group. In other words, an observation of the unacceptable variation between predictions from the two groups.** The value of z_1 is then scaled to eliminate this variation.


* In this context (classification model, $z_i \in [0,1]$), $\beta_1 \neq \{-1, 1\}$
$\hat{y}_1 = \beta_0 + \beta_1 z_1 + \beta_2 x_1$
$\hat{y}_0 = \beta_0 + \beta_1 z_0 + \beta_2 x_1$
new $z_1 = z_1(1 - (\hat{y}_1 - \hat{y}_0))$
where $z_1 - z_0 = 1$, mean$(z) \neq 0$, $z_i \neq \{0,1\}$, and $\hat{y}_i \in [0,1]$
what is this? new $z_1 = z_1(1 - (\hat{y}_1 - \hat{y}_0))$
$= z_1(1 - (\beta_1 z_1 - \beta_1 z_0))$
$= z_1 - \beta_1 z_1^2 - \beta_1 z_0 z_1$
when does new $z_1 = 0$? $0 = z_1 - \beta_1 z_1^2 - \beta_1 z_1 z_0$
$\beta_1 z_1 z_0 = z_1 - \beta_1 z_1^2$
$\beta_1 z_0 = 1 - \beta_1 z_1$
$z_0 = \frac{1}{\beta_1} - z_1$
actually since z_1 is an integer $\notin \{0,1\}$, this will only be possible when $\beta_1 = 1$ and $z_1 = 2$
→ since $z_1 - z_0 = 1$, $z_0 = z_1 - 1$ => when $z_1 = \frac{1}{\beta_1} + 1$
and when $\frac{1}{\beta_1} = -1$ (substitute $z_1 = 1 + z_0$).
($\Rightarrow \beta_1 = -1$)
Both of these can reasonably be considered "edge cases" at this stage. Current encoding for start z_1 is 2; try changing this, to avoid future issues on other data sets
when $\beta_1 = -1$ or, potentially, $\beta_1 = 1$ and $z_1 = 2$

only when $\beta_1 = -1$ & $z_0 = -2$

this can also happen when scaling down by a factor of $y_1 - y_0$ leads to it "naturally"; show empirically that this isn't always what happens, but also clarify conceptually why this isn't a problem (e.g. diabetes example doesn't hold)

When does $\text{new}z_1 = z_0$?

$$z_0 = z_1 - \beta_1 z_1^2 - \beta_1 z_1 z_0$$
$$z_0(1+\beta_1 z_1) = z_1 - \beta_1 z_1^2$$
$$z_0(1+\beta_1 z_1) = z_1(1-\beta_1 z_1)$$
$$z_0 = \frac{z_1(1-\beta_1 z_1)}{(1+\beta_1 z_1)}$$

→ since $z_1 - z_0 = 1$, $z_0 = z_1 - 1$ ① and $z_1 = 1 + z_0$ ②:

① $z_1 = \frac{z_1(1-\beta_1 z_1)}{1+\beta_1 z_1} + 1$; $z_1(1+\beta_1 z_1) = z_1(1-\beta_1 z_1) + 1 + \beta_1 z_1$

$\Rightarrow z_1 + \beta_1 z_1^2 = z_1 - \beta_1 z_1^2 + 1 + \beta_1 z_1$

$\Rightarrow 2\beta_1 z_1^2 = 1 + \beta_1 z_1 \Rightarrow 1 = \beta_1(2z_1^2 - z_1)$

$\frac{1}{\beta_1} = 2z_1^2 - z_1$

→ since $z_1 \notin \{0, 1\}$ and is an integer, this never happens;
$\beta_1 \in \{-1, 1\}$ are the only possibilities, but no valid z_1 satisfies this

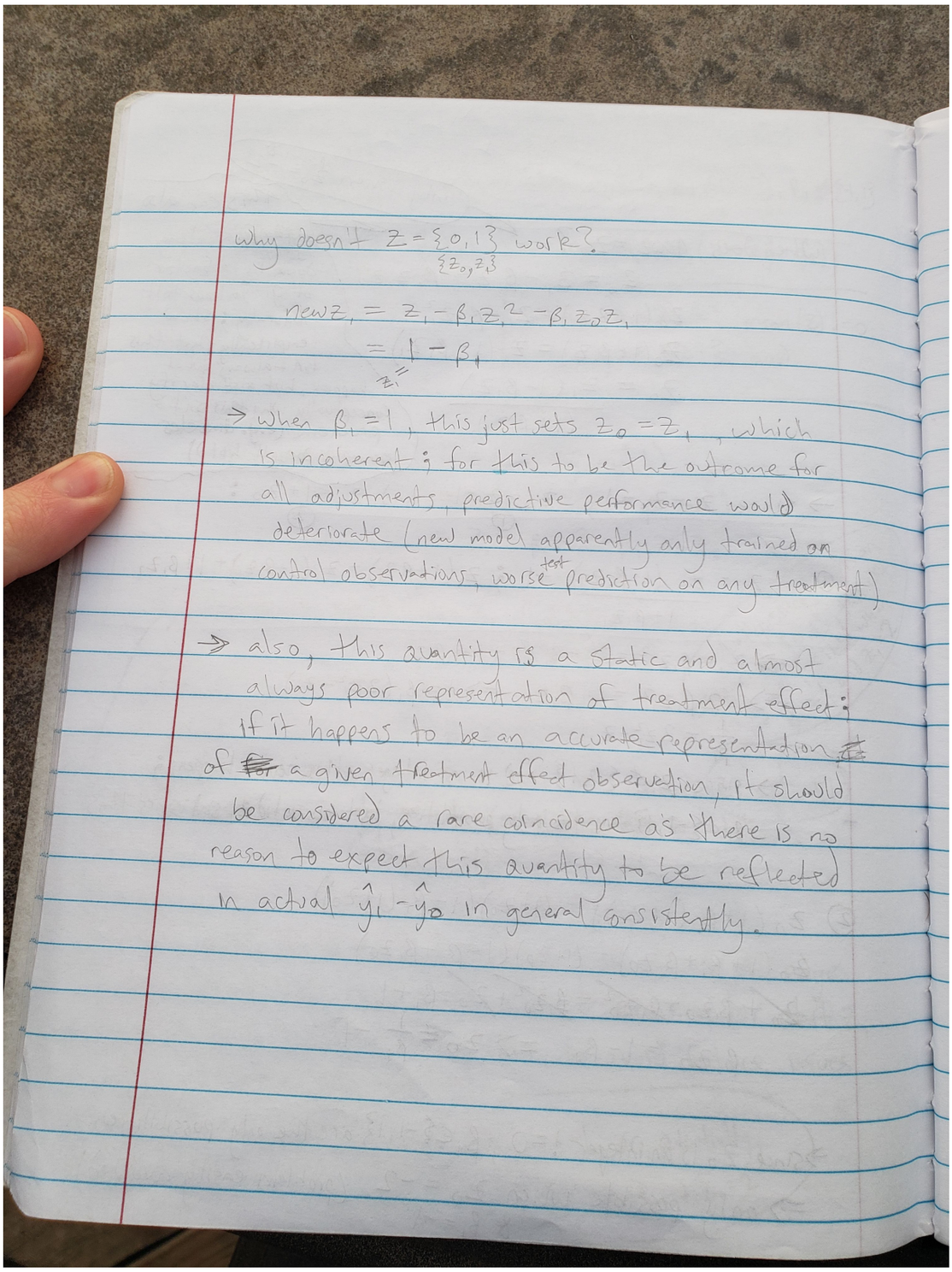
② $z_0(1+\beta_1(1+z_0)) = (1+z_0)(1-\beta_1(1+z_0))$

$z_0(1+\beta_1+\beta_1 z_0) = (1+z_0)(1-\beta_1+\beta_1 z_0)$

$z_0 + \beta_1 z_0 + \beta_1 z_0^2 = \beta_1 z_0^2 + z_0 - \beta_1 + 1$

$\beta_1 z_0 = 1 - \beta_1 \Rightarrow z_0 = \frac{1}{\beta_1} - 1$

→ since z_0 is an integer $\neq 0$, $\beta_1 \in \{-1, 1\}$ are the only possibilities
$\Rightarrow$ only possible when $z_0 = -2$ (problem easily avoided)
& $\beta_1 = -1$


why doesn't $Z = \{0,1\}$ work?
$\{z_0, z_1\}$
$newz_1 = z_1 - \beta_1 z_1^2 - \beta_1 z_0 z_1$
$= 1 - \beta_1$
z_1
→ when $\beta_1 = 1$, this just sets $z_0 = z_1$, which is incoherent; for this to be the outcome for all adjustments, predictive performance would deteriorate (new model apparently only trained on control observations, worse test prediction on any treatment)
→ also, this quantity is a static and almost always poor representation of treatment effect; if it happens to be an accurate representation of a given treatment effect observation, it should be considered a rare coincidence as there is no reason to expect this quantity to be reflected in actual $\hat{y}_1 - \hat{y}_0$ in general consistently.