2013 IEEE 13th International Conference on Data Mining

Controlling Attribute Effect in Linear Regression

Toon Calders*, Asim Karim†, Faisal Kamiran‡, Wasif Ali†, and Xiangliang Zhang‡
*Computer and Decision Engineering Dept., Universite Libre de Bruxelles (ULB), Belgium
Email: toon.calders@ulb.ac.be
†Dept. of Computer Science, SBASSE, Lahore University of Management Sciences (LUMS), Pakistan
Email: akarim@lums.edu.pk; xs2wasifali@gmail.com
‡CEMSE Division, King Abdullah University of Science and Technology (KAUST), KSA
Email: faisal.kamiran@gmail.com; xiangliang.zhang@kaust.edu.sa

***Abstract*—In data mining we often have to learn from biased data, because, for instance, data comes from different batches or there was a gender or racial bias in the collection of social data. In some applications it may be necessary to explicitly control this bias in the models we learn from the data. This paper is the first to study learning linear regression models under constraints that control the biasing effect of a given attribute such as gender or batch number. We show how propensity modeling can be used for factoring out the part of the bias that can be justified by externally provided explanatory attributes. Then we analytically derive linear models that minimize squared error while controlling the bias by imposing constraints on the mean outcome or residuals of the models. Experiments with discrimination-aware crime prediction and batch effect normalization tasks show that the proposed techniques are successful in controlling attribute effects in linear regression models.**

***Keywords*—Linear Regression; Fair Data Mining; Batch Effects; Propensity Score**

I. Introduction

In data mining we are often confronted with situations where we have to learn from data that is biased in one way or another. A first potential reason of data bias comes from data being collected from different sources; nowadays, many companies and scientific communities are collecting huge data repositories, opening up unprecedented opportunities for data mining. For instance, by combining data from multiple medical studies one may be able to identify statistically relevant patterns that are not apparent in the individual datasets in isolation. More recently, another type of bias arising from dependence on a socially sensitive attribute was identified in the data mining community [1]–[3]. This type of bias, which in general can be categorized as a measurement-cum-selection bias, can lead to unfair and illegal decisions if not controlled in statistical models. For instance, in a demographic dataset it may be observed that for the same type of work, overall females receive a lower income than males. A classifier trained on such a dataset may pick up this dependency and predict lower wages either directly for females, or indirectly for persons with female characteristics. Depending on the application field, using such a biased classifier may be undesirable, unethical, or even illegal.

From now onward, we will abstract away from the reason of the data bias, and assume that (1) the data can be grouped in one way or another, by batch, gender, ethnicity, or any other grouping attribute, (2) there is a bias with respect to this grouping, and (3) we want to control and remove, at least to some extent, this bias in the models we learn from the data. Specifically, we study how we can learn linear regression models in such a situation by imposing additional constraints on the learning process. There already exist several approaches for removing the bias from the training data [4]. These approaches, however, may fail to identify some specific subgroups of the data where the bias occurs. Our approach is the first one to directly adjust regression models. As such, it is orthogonal to existing bias removal techniques and can be used in combination.

We start by proposing two measures, one based on the mean difference between predictions in one group versus the other, and another one based on the area under the ROC curve (AUC). Then, we introduce a way to deal with the fact that often the bias can partially be justified by some *explanatory* attributes. For example, attributes "number of working hours" and "education level" may explain salary differences between males and females to some extent. We assume that these attributes are externally nominated by the domain experts. To remove this effect of the explanatory attributes and balance the groups, we use a technique from statistics, named propensity modeling. Based on a so-called *propensity score* we divide the data into *strata* such that within the stratum none of the bias can be justified by the explanatory attributes.

Once the explanatory part of the bias has been removed by dividing into strata, we start the modeling process. To compensate for the remaining, unexplainable bias in the input data, we impose additional constraints in the learning process. We study two constraints: the first constraint imposes that the mean prediction needs to be the same for the different batches or groups. Loosely speaking, this constraint expresses that members of the different groups need to be treated similarly even if that is not the case in the input data. The second constraint expresses that the mean residuals needs to be the same; that is, the errors for the different groups need to be balanced. For both constraints we show analytical solutions to derive an optimal linear regression model on the training data under different conditions: one model per stratum versus one general model valid for all strata, and whether or not the group identifier is part of the final model. Depending on the setting the result is hence either one unbiased model per stratum, or one global model that is unbiased conditioned on stratum.

All techniques have been implemented and were tested on two datasets. The experiments show that the techniques are able to satisfy the constraints on the training data and that these results carry over to the test data. They also demonstrate

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.114

the applicability of the techniques for discrimination-aware regression and source rating normalization.

In summary, our contributions in this paper are as follows:

1) For the first time the problem of bias control in regression models is introduced. We present two measures to quantify bias; one based on the mean difference in predictions, and one on the AUC.
2) For the first time the issue of attributes explaining part of the bias is treated in a principled way, based upon the propensity scoring technique from statistics.
3) A new technique for controlling the influence of a categorical grouping attribute, such as batch number or ethnicity, in linear regression models is presented. This technique is based upon the introduction of constraints in the optimization problem at the basis of the ordinary least squares regression modelling. Optimal analytical solutions to the constrained optimization problem are derived.
4) Experiments with two real datasets show the potential of the two step technique consisting of first removing the explanatory bias with the propensity based method, followed by the induction of a regression model under constraints.

II. Related Work and Motivation

Bias in observations and statistical models has been studied for decades with applications in sociology [5], econometrics [6], and biomedicine [7]–[9]. Also attribute bias in regression models appears in many applications. The issue of gender income or wage gap, for instance, has been studied extensively in sociology and economics [10], [11]. Legislation exists in many countries that disallow wage discrimination with respect to gender, with severe penalties stipulated for violating employers. Thus, a discrimination-ignorant approach to wage prediction based on historical data (which is often biased) can lead to violation of laws. Similarly, racial discrimination in criminology and police arrests is a continuing concern [12]–[14]. In this setting, a discrimination-ignorant regression model based on historical data can further exacerbate the racial discrimination. This has been demonstrated recently for crime suspect prediction using real-world data [14].

There already exist quite some works in discrimination-aware classification that deal with social discrimination during learning a classifier. For instance, techniques exist for learning decision trees [15], Bayesian models [16], and logistic regression [2] from biased data. The regression problem we study in this paper, however, is more challenging in the following senses. Firstly, instead of assessing the correlation between two categorical attributes (for instance race and label), we now have to assess the correlation between the categorical sensitive attribute and the continuous target. Secondly, the number of ways to change how a model predicts increases greatly; in classification the only possible modification is the change of one class label into another. For regression tasks, the continuous character of the target allows for a continuous range of potential changes.

Besides the discrimination-aware regression application, attribute bias arises in many other application domains as well. For example, two publishers or evaluators can generate ratings for products/services where the ratings of one publisher are generally higher than that of the other [17], [18]. It is therefore important that this bias is controlled in any regression model learned over data from both publishers. Another application of bias-aware regression is that of observational and experimental studies of data from two or more different sources with different selection and measurement biases [19]. One notorious example of such data collections comes from the field of computational biology, where huge repositories of experimental micro-array data from different studies are being collected such that they can be combined and reused in new studies. Regarding these collections, however, *Chen et al.* [20] state that "*data produced by the thousands of micro-array studies published annually are confounded by "batch effects," the systematic error introduced when samples are processed in multiple batches. Although batch effects can be reduced by careful experimental design, they cannot be eliminated unless the whole study is done in a single batch.*" This type of bias arises in other situations as well, like product ratings being influenced by their source/publisher [18]. It is crucial to either remove such bias before mining, or to explicitly take it into account during the learning process.

A situation in which the failure to remove bias in an application leads to undesirable outcomes comes from *Sweeney* [21], who analyzed ads provided by Google, and came to the conclusion that for certain searches "*a black-identifying name was 25% more likely to get an ad suggestive of an arrest record*" and raised the question whether "*Google's advertising technology exposes racial bias in society and how ad and search technology can develop to assure racial fairness.*" Discrimination-aware data mining studies the development and application of methods for discovering and preventing discrimination from models learned over discriminatory datasets.

III. Preliminaries

In this section, we introduce the task of controlling the effect of an attribute in a regression model and we define different measures of model unbiasedness.

We consider a dataset $\mathcal{D} = \{\mathbf{x}_i, t_i\}_{i=1}^N$. The vector $\mathbf{x}_i$ represents the input attributes and the scalar t_i represents the target for the ith instance. We will focus our work on learning linear regression models of the form $t(\mathbf{x}) = w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}$. For notational convenience, we take parameter w_0 as the first element of weight vector $\mathbf{w}$ and $x_0 = 1$ as the first element of all input vectors $\mathbf{x}_i$. We denote the length of the vectors $\mathbf{w}$ and $\mathbf{x}_i$ by M. Furthermore we assume that $\mathcal{D}$ can be divided into different groups, based upon a factor, or grouping attribute x_s. For ease of presentation we will restrict our derivations to a factor with two levels, dividing the dataset into two partitions $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$. All results can be generalized to multiple groups. We will use $N^{\uparrow}$ and $N^{\downarrow}$ to respectively denote the number of instances in $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$. In the case of social fairness, for example, x_s could be gender dividing $\mathcal{D}$ into $\mathcal{D}^{\uparrow}$, the males, and $\mathcal{D}^{\downarrow}$, the females. Inspired by the applications in fairness-aware data mining we will often refer to the factor x_s as the *sensitive attribute*.

The specific goal in the paper is to learn the weights $\mathbf{w}$ of a linear regression model such that the sum of squared errors, $SSE := \sum_{i=1}^N (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - t_i)^2$, on $\mathcal{D}$ is minimized, while the

TABLE I. EXAMPLE DATASET FOR A SALARY STUDY. WAGE IS THE TARGET VARIABLE. THE DATA IS BIASED TOWARD LOWER WAGES FOR FEMALES.

#	Gender	Study (Years)	Working Hours	Sector Health?	Wage (K)
1	M	5	40	0	66
2	M	5	40	0	66
3	M	3	40	1	60
4	M	2	30	0	44
5	M	2	40	1	56
6	F	4	40	1	60
7	F	3	40	0	55
8	F	3	30	1	42
9	F	2	30	0	40
10	F	2	30	1	40

direct and indirect influence of the sensitive attribute x_s in the predictions gets controlled.

While presenting our strategy for quantifying and controlling attribute bias in regression models, we will illustrate concepts and techniques by referring to a fictitious employee wage dataset, given in Table I, that is known to be biased or discriminatory with respect to attribute gender, with female employees generally having lower wages than comparable male employees. In this example, education level (expressed as number of years of study), weekly working hours, and whether or not the person works in the health sector are the predictor attributes, wage is the target variable, and gender is the sensitive attribute describing the grouping.

Example 1: When learning a linear regression model on the dataset of Table I, we find the following model (female is encoded as $G = 0$, male as $G = 1$):

$$-4 \times G + 3.3 \times S + 1.3 \times W + 0.16 \times HC - 1.1 \qquad (1)$$

Hence, we can see a strong influence of gender in the predictions. If we remove Gender from the list of attributes, we obtain the following model:

$$3.1 \times S + 1.5 \times W - 0.7 \times HC - 8.9 \qquad (2)$$

Here we can observe the influence of HC (person working in the health care sector) increases as this sector has a high relative number of female employees and is used as a proxy. The mean salary for males under this model is 58.4, and for females is 51.4. This example already shows that simply removing the Gender attribute does not solve the problem.

A. Measuring Imbalance in Data and Models

Intuitively, an attribute effects a target variable if there is a statistical dependency between the attribute and the target. Since the sensitive attribute x_s is binary valued while target variable t is continuous, it is not possible to use typical same-type measures of dependency like correlation coefficient and point-wise mutual information to quantify the statistical dependency between x_s and t. We will use the *Mean Difference (MD)* and the *Area Under the ROC Curve (AUC)* for quantifying the effect of x_s on the target variable.

Definition 1: (MD): The mean difference (MD) of the continuous target variable t in dataset $\mathcal{D}$, partitioned into $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ by a sensitive attribute x_s is given by:

$$MD(t, x_s; \mathcal{D}) = \frac{\sum_{(\mathbf{x},t)\in\mathcal{D}^{\uparrow}} t}{N^{\uparrow}} - \frac{\sum_{(\mathbf{x},t)\in\mathcal{D}^{\downarrow}} t}{N^{\downarrow}} \qquad (3)$$

The mean difference is a real number with a value of zero signifying no dependency or attribute effect.

Definition 2: (AUC): The area under the ROC curve (AUC) of the continuous target variable t in dataset $\mathcal{D}$, partitioned into $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ by a sensitive attribute x_s, is given by:

$$AUC(t, x_s; \mathcal{D}) = \frac{\sum_{(\mathbf{x}_u,t_u)\in\mathcal{D}^{\uparrow}} \sum_{(\mathbf{x}_d,t_d)\in\mathcal{D}^{\downarrow}} I(t_u > t_d)}{N^{\uparrow} \times N^{\downarrow}} \qquad (4)$$

where $I(\cdot)$ is the indicator function that returns 1 when its argument is true and 0 otherwise.

The AUC varies from zero to one, and it is symmetric around 0.5 which represents random predictability or zero attribute effect. The AUC value is a statistically consistent measure of predictive strength [22].

In the definitions above, the effect of the sensitive attribute on the target is measured; however, similar definitions can be applied to regression models $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$, by replacing the true target t with either the prediction $y(\mathbf{x})$, or the residual $y(\mathbf{x}) - t$. In this way, we get the measures MD_o and AUC_o that measure the effect of x_s on the outcome of the regression model, and MD_r and AUC_r that measure the effect of x_s on the residuals of the regression model's predictions.

Example 2: For the example dataset of Table I, $MD(wage, gender) = 58.4 - 47.4 = 11$, and $AUC(wage, gender) = 21/25 = 0.84$. Thus, there is a strong (predictive) dependency between wage and gender. The regression model of Equation (2) makes the following (rounded) predictions for the dataset:

#	1	2	3	4	5	6	7	8	9	10
score	65	65	58	41	55	61	59	44	41	40

The residuals are:

#	1	2	3	4	5	6	7	8	9	10
score	−1	−1	−2	−3	−1	1	4	2	1	0

Hence, the measures for the classifier are as follows:

	prediction	residual
MD	7.85	-3.15
AUC	0.68	0

From this we can conclude that the classifier's predictions are still biased with respect to gender, although the bias is less strong. From the measures applied to the residuals we learn that the errors that the classifier makes (if we consider the target t as the ground truth), are highly biased; salaries of males are systematically underestimated, while those of females are overestimated. Such a strong bias in the residuals may indicate a so-called omitted variable bias; in this case omitting the gender attribute led to the overestimation of the effect of "Sector Health?" attribute and overestimation (in absolute terms, but in negative direction) of the constant factor.

IV. ADJUSTING FOR EXPLANATORY ATTRIBUTES: A PROPENSITY SCORE BASED APPROACH

In many cases we cannot directly apply our constraint-based method, because it is unreasonable to assume that the constraint holds on the raw data. In the example of the wages in section III, for instance, the reason for lower wages of females

can partially be attributable to their shorter working hours. To filter out this explainable difference, we propose a propensity score based stratification approach for balancing the dataset.

A. Propensity Modeling

We illustrate the propensity score with an example. The treatment patients receive often depends on their general condition, which in turn has a direct influence on their survival (the target variable). If we want to study the influence of treatment A versus B on survival, we will have to adjust for all parameters that influenced the treatment the patient received, including his or her general health at the start of the treatment. That is, we determine the *propensity* of a patient to receive a certain treatment based on his or her parameters and capture it as a score. This score is usually estimated from the dataset by a logistic regression model. Roughly put, this score expresses how likely the patient was, based on his or her characteristics, to receive treatment A (regardless of the treatment he or she received in the end). An important property of the propensity score is [23]: given the propensity score, the target is independent of the explanatory attributes. Thus, two instances that have the same propensity score should have the same value for the target, as explained by these attributes; if the target values differ, then this represents an unexplained effect (unexplained by the specified explanatory attributes) of the attribute on the target. After obtaining the propensity score, it is either included as part of the prediction model, or the analysis is designed as to ensure that only patients with similar propensity are compared. We will base ourselves on a common technique that partitions the input data into strata according to the propensity score.

B. Accounting for Explainable Bias

In our setting, we assume there are externally provided explanatory attributes $\mathbf{x}_e$, such as the number of working hours and years of study in the wages example, or a set of variables of which we know that they have been varied between different batches of data. The differences in treatment of the groups $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ that is explainable by these variables $\mathbf{x}_e$ can be considered explainable bias, and should not be controlled. To achieve this, we proceed as follows:

1) The propensity score is defined as the conditional probability that a randomly selected instance of $\mathcal{D}$ belongs to $\mathcal{D}^{\uparrow}$, given the explanatory attributes, i.e., $ps = P(\mathbf{x} \in \mathcal{D}^{\uparrow} | \mathbf{x}_e)$.
2) Using the propensity scores the dataset is split into five strata via the propensity score quintiles. That is, each split contains 20% of the data lying between quintile i and quintile $i+1$ of the propensity score ($i = 0 \ldots 5$). Let $\mathcal{S}_i$ ($i = 1, 2, \ldots, 5$) denote the five strata of dataset $\mathcal{D}$. Recall the fact whether or not $\mathbf{x} \in \mathcal{D}^{\uparrow}$ is independent of the variables $\mathbf{x}_e$, conditioned on the ps, and hence within the strata, the dependency between the division into $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ and the explanatory attributes $\mathbf{x}_e$ will be minimal, as all instances in a stratum have similar propensity.

The potential difference in treatment between $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ within a stratum can no longer be attributed to the explanatory attributes. We will use $AUC_i = AUC(t, x_s; \mathcal{S}_i)$ and $MD_i = MD(t, x_s; \mathcal{S}_i)$ to denote the AUC and MD for stratum i. AUC_i or MD_i represent the unexplained effect of attribute x_s on t in stratum i. This is the effect that needs to be controlled or removed in the regression models we learn.

Example 3: Suppose that for the data in Table I, we assume that differences in salary between the genders can partially be attributed to *Study* (numbers of years spent to study) and *Working Hours* (number of working hours per week). Hence, we consider the attributes *Study* and *Working Hours* explanatory attributes. As a first step we will learn the propensity score for Gender given these two attributes. For the data the following logistic regression function for gender is learned:

$$logit(p(Gender = female)) = -0.047 \times S - 0.17 \times W + 6.4$$

Based on this function, the following scores are computed for the different instances. For the males:

#	1	2	3	4	5
score	0.32	0.32	0.34	0.75	0.35

And for the females:

#	6	7	8	9	10
score	0.33	0.34	0.74	0.75	0.75

Given the size of the dataset we will divide the dataset into two strata instead of 5, and compute the MD and AUC for both strata:

	Stratum 1	Stratum 2
Males	1(66), 2(66), 3(60)	4(44), 5(56)
Females	6(60), 7(55)	8(42), 9(40), 10(40)
MD	6.5	9.33
AUC	5/6 = 83%	100%

So, in comparison to the measures on the complete dataset, we can observe that the mean difference decreases while the AUC increases. This reverse relation between the two measures indicates that after filtering out the effect of the explanatory variables, even though the differences in wages become smaller on average, they also become more persistent.

V. Controlling Bias With Constraints

In this section we will concentrate on learning models in which the influence of a sensitive attribute x_s on the target t is controlled by constraints. In the last section we introduced an approach based on propensity scoring to partition the dataset into strata to reduce the biasing effect of a sensitive attribute x_s on the target t that can be justified by a set of explainable attributes $\mathbf{x}_e$. This stratification will be an important first step in our methodology to control the influence of x_s, because obviously we do not want to remove the explainable part of the bias. Therefore, in the learning process, we will deal with the remaining bias in the strata by either:

1) learning different models for every stratum separately, and control the difference in treatment between $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ with a constraint,
2) or, learn one global model for the whole dataset controlling the difference in treatment between $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$ for all strata at the same time with multiple constraints.

We will consider the following two constraints on a model $y(\mathbf{x})$ for a stratum $\mathcal{S}$:

Equal Means: The mean predictions for $\mathcal{S}^{\uparrow} = \mathcal{D}^{\uparrow} \cap \mathcal{S}$ and $\mathcal{S}^{\downarrow} = \mathcal{D}^{\downarrow} \cap \mathcal{S}$ need to be equal. That is:

$$MD_o(y(\mathbf{x}), x_s; \mathcal{S}) \;=\; 0$$

This constraint expresses that future predictions across the two groups should be comparable, regardless if this was the case in the original data or not. Notice that this constraint does not depend on the target. As such, if the (unlabeled) test instances are already known at training time, we could enforce this constraint directly on the test instances.

Balanced Residuals: Sometimes we do have a true difference between two groups that can be reflected in the predictions, as long as there is no bias in the errors made by the predictor. This can be controlled by requiring that the mean residual for both groups is equal:

$$MD_r(y(\mathbf{x}), x_s; \mathcal{S}) = 0$$

It could hence be acceptable that predictions in one group are consistently higher than in the other group, as long as this was the case in the original data, and the effect is not exaggerated.

We start by showing two analytical solutions in which the constraints are strictly satisfied by the model. After showing the solution for building a model for one partition, we extend it to multiple partitions. We end the section with a relaxed version where the degree to which the constraints are satisfied is added as a penalty term.

A. Strict Equal Means Constraint

The goal is to learn the coefficients $\mathbf{w}$ of a linear function $\mathbf{w} \cdot \mathbf{x}$ such that the SSE on $\mathcal{S}$ is minimized conditioned to the strict constraint that the mean difference between the predictions on $\mathcal{S}^{\uparrow}$ and $\mathcal{S}^{\downarrow}$ is 0. That is:

$$\textbf{minimize} \sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} (\mathbf{w}\cdot\mathbf{x}_i - t_i)^2$$
$$\textbf{subject to } \frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\uparrow}} \mathbf{w}\cdot\mathbf{x}_i}{N^{\uparrow}_{\mathcal{S}}} = \frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\downarrow}} \mathbf{w}\cdot\mathbf{x}_i}{N^{\downarrow}_{\mathcal{S}}}.$$

We will use $\mathbf{d}$ to denote $\frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\uparrow}} \mathbf{x}_i}{N^{\uparrow}_{\mathcal{S}}} - \frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\downarrow}} \mathbf{x}_i}{N^{\downarrow}_{\mathcal{S}}}$; i.e., $\mathbf{d}$ is the difference between the mean vector of $\mathcal{S}^{\uparrow}$ and the mean vector of $\mathcal{S}^{\downarrow}$. The condition for the mean difference then becomes: $\mathbf{w}\cdot\mathbf{d} = 0$.

We solve this minimization problem using Lagrange multipliers. We use the following Lagrangian for the constrained minimization problem:

$$L \;:=\; \sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} (\mathbf{w}\cdot\mathbf{x}_i - t_i)^2 + 2\lambda \mathbf{w}\cdot\mathbf{d}$$

We take partial derivatives w.r.t. the coefficients $\mathbf{w}_j$:

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{w}_j} &= \sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} 2(\mathbf{w}\cdot\mathbf{x}_i - t_i)\mathbf{x}_{ij} + 2\lambda\mathbf{d}_j \\ &= 2\left(\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} \mathbf{x}_i\mathbf{x}_{ij}\right)\cdot\mathbf{w} - 2\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} t_i\mathbf{x}_{ij} + 2\lambda\mathbf{d}_j\end{aligned}$$

Equating to 0 gives that for all j,

$$\left(\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} \mathbf{x}_i\mathbf{x}_{ij}\right)\cdot\mathbf{w} = \sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}} t_i\mathbf{x}_{ij} - \lambda\mathbf{d}_j$$

Hence, $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \;=\; \mathbf{X}^{\mathrm{T}}\mathbf{t} - \lambda\mathbf{d}$,
and thus, $\mathbf{w} \;=\; (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - \lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{d}$.
Together with the equality $\mathbf{w}\cdot\mathbf{d} = 0$, we can solve for λ:

$$\lambda \;=\; 2\frac{\left((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}\right)\cdot\mathbf{d}}{((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{d})\cdot\mathbf{d}}$$

Reinserting this in the formula for $\mathbf{w}$, we get:

$$\mathbf{w} \;=\; (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}}{\mathbf{d}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\cdot\mathbf{d}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{d}$$

Example 4: Consider again the wages dataset of Table I. If we apply the Equal Means constraint, the optimal linear regression model becomes:

$$7\times G + 3.3\times S + 1.3\times W + 0.16\times HC - 1.1 \qquad (5)$$

In comparison to the model in Eq. 1, the gender attribute is used to give a "bonus" to the females in the dataset. The predictions (rounded) for this new model are:

#	1	2	3	4	5	6	7	8	9	10
score	61	61	54	38	51	65	61	49	45	45

The residuals are:

#	1	2	3	4	5	6	7	8	9	10
score	−5	−5	−6	−6	−5	5	6	7	5	5

The measures for the classifier are as follows:

	prediction	residual
MD	0	-11
AUC	0.48	0

Not surprisingly, the mean difference for the predictions is 0, and for the residuals it is -11. AUC follows the trend.

B. Strict Balanced Residual Constraint

We follow a similar approach as in last subsection. The constraint for balanced residuals, however, is slightly different:

$$\mathbf{w}\cdot\mathbf{d} \;=\; \frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\uparrow}} t_i}{N^{\uparrow}_{\mathcal{S}}} - \frac{\sum_{(\mathbf{x}_i,t_i)\in\mathcal{S}^{\downarrow}} t_i}{N^{\downarrow}_{\mathcal{S}}} \;.$$

We will denote the right-hand side of this equality by b. The constraint thus becomes $\mathbf{w}\cdot\mathbf{d} - b \;=\; 0$. Using the same technique as for Equal Means, we obtain:

$$\mathbf{w} \;=\; (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - \frac{\mathbf{d}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - b}{\mathbf{d}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\cdot\mathbf{d}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{d}$$

Example 5: Consider again the wages dataset of Table I. This time, in contrast to the previous example, we do not need a predictor that assigns equal wages to males and females in our dataset, but we do need the errors between males and females to be balanced. This time a model such as in Eq. (5) is completely unacceptable, as it makes all negative errors on the males and all positive errors on the females. Therefore,

now we apply the Balanced Residuals constraint. Under this constraint, the optimal linear regression model becomes:

$$-4 \times G + 3.3 \times S + 1.3 \times W + 0.16 \times HC - 1.1 \qquad (6)$$

This model is exactly the same as the normal linear regression model in Eq. (1). This is not a coincidence as we will see later, but true in general when the sensitive attribute partitions $\mathcal{D}$ in two parts, and is used as one of the predictor variables. If we omit the sensitive attribute gender, we get the following model:

$$2.7 \times S + 2.1 \times W - 4 \times HC - 31 \qquad (7)$$

This model has a mean difference of 11, and a difference in residual of 0, as expected. AUC follows this trend with 72% for the predictions and 52% for the residuals.

C. Extending to Multiple Partitions and Groups

Recall that in the previous subsections we have dealt with the case of only one constraint. Usually we will have multiple constraints:

1) x_s may have more than 2 values, and thus there are more than 2 groups ($\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$) in the data; that is: $\mathcal{D}$ is partitioned into $\mathcal{D}_1, \ldots, \mathcal{D}_k$ by x_s. In that case the constraints for Strict Equal Means become:

$$\mathbf{w} \cdot \mathbf{d}^{12} = 0, \mathbf{w} \cdot \mathbf{d}^{13} = 0, \ \ldots, \text{ and } \mathbf{w} \cdot \mathbf{d}^{1k} = 0 \ ,$$

where $\mathbf{d}^{1i}$ denotes the difference between the mean of group 1 and group i. Similarly, for the Strict Balanced Residual constraint, we get:

$$\mathbf{w} \cdot \mathbf{d}^{12} = b^{12}, \mathbf{w} \cdot \mathbf{d}^{13} = b^{13}, \ \ldots, \text{ and } \mathbf{w} \cdot \mathbf{d}^{1k} = b^{1k} \ .$$

2) Sometimes it may be useful to build one global model for the complete dataset, instead of different models for each stratum. In such a case we may opt to enforce the Strict Equal Means constraint on all strata at the same time, instead of building separate models for the different strata. Again this will result in a set of similar constraints. For instance, suppose we have two groups $\mathcal{D}^{\downarrow}$ and $\mathcal{D}^{\uparrow}$, and ℓ strata dividing $\mathcal{D}$ into $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$. Let $\mathbf{d}^i$, $i = 1 \ldots \ell$ denote

$$\mathbf{d}^i := \frac{\sum_{(\mathbf{x},t) \in \mathcal{S}_i^{\uparrow}} \mathbf{x}}{|\mathcal{S}_i^{\uparrow}|} - \frac{\sum_{(\mathbf{x},t) \in \mathcal{S}_i^{\downarrow}} \mathbf{x}}{|\mathcal{S}_i^{\downarrow}|} \ .$$

The Strict Equal Means constraint becomes:

$$\mathbf{w} \cdot \mathbf{d}^1 = 0, \mathbf{w} \cdot \mathbf{d}^2 = 0, \ \ldots, \text{ and } \mathbf{w} \cdot \mathbf{d}^{\ell} = 0 \ .$$

We will show how to optimize for multiple constraints in general. For all cases described above, the optimization problem can be casted as follows:

$$\textbf{minimize} \sum_{(\mathbf{x}_i,t_i) \in \mathcal{D}} (\mathbf{x}_i^{\mathrm{T}}\mathbf{w} - t_i)^2$$
$$\textbf{subject to } \boldsymbol{\Delta}^{\mathrm{T}}\mathbf{w} = \boldsymbol{\delta}.$$

$\boldsymbol{\Delta}$ is the matrix containing the $\mathbf{d}^{ij}$'s or $\mathbf{d}^j$ as columns, and $\boldsymbol{\delta}$ is a vector containing the constant factors b^{ij}, or 0.

Again we can use the same technique via Lagrange multipliers. The Lagrangian for the multiple constraints is:

$$L := \sum_{(\mathbf{x}_i,t_i) \in \mathcal{D}} (\mathbf{x}_i^{\mathrm{T}}\mathbf{w} - t_i)^2 + 2\boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{w} - \boldsymbol{\delta})$$

Equating the gradient with respect to $\mathbf{w}$ to 0, and solving for $\boldsymbol{\lambda}$ gives:

$$\boldsymbol{\lambda} = (\boldsymbol{\Delta}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{\Delta})^{-1}\left(\boldsymbol{\Delta}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - \boldsymbol{\delta}\right) \ .$$

Substituting this into $\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{w} = \boldsymbol{\delta}$ and solving for $\mathbf{w}$ gives the following final formula for $\mathbf{w}$:

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\Big[\mathbf{X}^{\mathrm{T}}\mathbf{t} - \boldsymbol{\Delta}(\boldsymbol{\Delta}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\boldsymbol{\Delta})^{-1}(\boldsymbol{\Delta}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t} - \boldsymbol{\delta})\Big]$$

D. Relaxing the Constraints

In the previous two subsections we have required in our analytical deductions that the constraints should be satisfied exactly. This, however, can lead to degenerate solutions when the number of constraints increases too much. Therefore, in this section we study a somewhat more standard approach in which we incorporate the constraints directly into the objective function as penalty terms. For the "equal means" constraint the objective to be minimized, using the notations introduced before in this section, becomes:

$$obj := \sum_{(\mathbf{x}_i,t_i) \in \mathcal{S}} (\mathbf{w} \cdot \mathbf{x}_i - t_i)^2 + \alpha(\mathbf{w} \cdot \mathbf{d})^2 \ , \qquad (8)$$

where α is a weighing factor controlling the influence of the penalty term. This formula can be optimized by taking partial derivatives for w_j:

$$\begin{aligned}\frac{\partial obj}{\partial w_j} &= 2 \sum_{(\mathbf{x}_i,t_i) \in \mathcal{D}} (\mathbf{w} \cdot \mathbf{x}_i - t_i)x_{ij} + 2\alpha(\mathbf{w} \cdot \mathbf{d})\mathbf{d}_j \\ &= 2\mathbf{w} \cdot \left(\sum_{(\mathbf{x}_i,t_i) \in \mathcal{S}} \mathbf{x}_i x_{ij} + \alpha\mathbf{d}\mathbf{d}_j\right) - 2 \sum_{(\mathbf{x}_i,t_i) \in \mathcal{S}} t_i x_{ij}\end{aligned}$$

Equating all partial derivatives to 0 gives:

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \alpha\mathbf{d}\mathbf{d}^{\mathrm{T}})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{t})$$

Due to space restrictions we do not show the formulas for the other cases, as they can be derived using the same techniques.

E. Relations Between the Approaches

The following theorem relates the different settings of learning linear models with and without constraints to each other. These relations, however, *only* hold when the sensitive attribute x_s splits the data in no more than two parts, *and* is a predictive variable of the prediction model. The theorem can be seen as a sanity check showing that in cases with full control over the sensitive attribute in our models, the optimal solutions under constraints can be obtained with intuitive adaptations to unconstrained optimal models.

Theorem 1: Given a dataset $\mathcal{D}$ and a sensitive attribute x_s that divides the dataset into two partitions $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$. We assume $\mathcal{D}$ has full column rank.

Let $\mathbf{w}$, $\mathbf{w}_{EM}$ and $\mathbf{w}_{BR}$, be the optimal linear models w.r.t. SSE under, respectively, no constraint, the equal means constraint, and the balanced residuals constraint, and

let $\mathbf{w}_{EM}^{approx,\alpha}$ and $\mathbf{w}_{BR}^{approx,\alpha}$ be the optimal linear models w.r.t. the approximate equal means and balanced residuals optimization problems of Subsection V-D. Then the following relations hold:

(1) $\mathbf{w}$ and $\mathbf{w}_{EM}$ only differ in the constant term (w_0) and the coefficient for x_s;

(2) $\mathbf{w} = \mathbf{w}_{BR}$;

(3) $\mathbf{w}_{EM}^{approx,0} = \mathbf{w}_{BR}^{approx,0} = \mathbf{w}$;

(4) $\lim_{\alpha\to\infty} \mathbf{w}_{EM}^{approx,\alpha} = \mathbf{w}_{EM}$; and

(5) $\lim_{\alpha\to\infty} \mathbf{w}_{BR}^{approx,\alpha} = \mathbf{w}_{BR}$.

Proof: We first prove (2). This statement comes down to showing that the optimal linear model without constraints already has a mean residual of 0 in both $\mathcal{D}^{\uparrow}$ ($x_s = 1$) and $\mathcal{D}^{\downarrow}$ ($x_s = 0$): Suppose the average residual on $\mathcal{D}^{\uparrow}$ is $m^{\uparrow}$ and on $\mathcal{D}^{\downarrow}$ is $m^{\downarrow}$. Let $\mathbf{w}^*$ be the following weight vector: take the weight vector $\mathbf{w}$ and subtract $m^{\downarrow}$ from the intercept, and add $m^{\downarrow}-m^{\uparrow}$ to the coefficient for x_s. This will shift the predictions for $\mathcal{D}^{\uparrow}$ by $-m^{\uparrow}$ and those for $\mathcal{D}^{\downarrow}$ by $-m^{\downarrow}$, making the mean residual in both groups 0. In this way the SSE will decrease by $(N^{\uparrow}m^{\uparrow})^2 + (N^{\downarrow}m^{\downarrow})^2$. Since $\mathbf{w}$ was optimal this sum must be 0 which is only possible if $m^{\downarrow} = m^{\uparrow} = 0$.

For (1), we will apply a similar construction with shifting the regression lines. Suppose that the mean residual for the weight vector $\mathbf{w}_{EM}$ is $\epsilon^{\uparrow}$ for $\mathcal{D}^{\uparrow}$, and $\epsilon^{\downarrow}$ for $\mathcal{D}^{\downarrow}$. Construct $\mathbf{w}_{EM}^*$ by shifting $\mathbf{w}_{EM}$ to reduce the residuals in both groups to 0 as above. The mean difference in outcome for $\mathbf{w}_{EM}^*$ will increase, due to this shift, to $\epsilon^{\uparrow}-\epsilon^{\downarrow}$, and the SSE will decrease by $(N^{\uparrow}\epsilon^{\uparrow})^2+(N^{\downarrow}\epsilon^{\downarrow})^2$. Since the residual of $\mathbf{w}_{EM}^*$ is 0 in both $\mathcal{D}^{\uparrow}$ and $\mathcal{D}^{\downarrow}$, the mean outcome for $\mathcal{D}^{\uparrow}$ (resp. $\mathcal{D}^{\downarrow}$) by $\mathbf{w}_{EM}^*$ is equal to the mean target value for $\mathcal{D}^{\uparrow}$ (resp. $\mathcal{D}^{\downarrow}$). Hence, we also have that $\epsilon^{\uparrow} - \epsilon^{\downarrow}$ equals the difference between the mean target value in $\mathcal{D}^{\uparrow}$ and the mean target value in $\mathcal{D}^{\downarrow}$. Suppose that the mean difference between the groups of the outcomes by the optimal unconstrained model with weights $\mathbf{w}$ is δ. Since $\mathbf{w}$ has a mean residual of 0, δ must be equal to the difference between the mean target value for $\mathcal{D}^{\uparrow}$ and the mean target value for $\mathcal{D}^{\downarrow}$. Hence, $\delta = \epsilon^{\uparrow} - \epsilon^{\downarrow}$. Construct the weight vector $\mathbf{w}'$ such that the instances in $\mathcal{D}^{\uparrow}$ are shifted by $shift_1 := -\delta\frac{(N^{\downarrow})^2}{(N^{\uparrow})^2+(N^{\downarrow})^2}$, and those in $\mathcal{D}^{\downarrow}$ by $shift_2 = -\delta\frac{(N^{\uparrow})^2}{(N^{\uparrow})^2+(N^{\downarrow})^2}$. The model based on vector $\mathbf{w}'$ will have equal means in both groups, since the sum of these two shifts is $-\delta$, compensating for the mean difference δ of $\mathbf{w}$. The SSE of $\mathbf{w}'$ is

$$\begin{aligned} &SSE(\mathbf{w}) + (shift_1 N^{\uparrow})^2 + (shift_2 N^{\downarrow})^2 \\ &\quad = SSE(\mathbf{w}) + (shift_1 N^{\uparrow})^2 + ((-\delta - shift_1)N^{\downarrow})^2 \end{aligned} \quad (9)$$

Since $\mathbf{w}$ is the optimal weight vector for the unconstrained optimization problem, and $\mathbf{w}_{EM}$ the optimal vector for the equal means constraint, we get the following equations:

$$\begin{aligned} SSE(\mathbf{w}) &\leq SSE(\mathbf{w}_{EM}^*) \\ &= SSE(\mathbf{w}_{EM}) - (N^{\uparrow}\epsilon^{\uparrow})^2 - (N^{\downarrow}\epsilon^{\downarrow})^2 \\ SSE(\mathbf{w}_{EM}) &\leq SSE(\mathbf{w}') \\ &= SSE(\mathbf{w}) + (shift_1 N^{\uparrow})^2 \\ &\quad +((-\delta - shift_1)N^{\downarrow})^2 \end{aligned}$$

TABLE II. KEY CHARACTERISTICS OF CRIME AND WINE DATASETS

	Crime	Wine
N	1994	6497
M	99	11
t	Crime Rate	Rating
x_s	Race ∈ {black, non-black}	Type ∈ {white, red}
$N^{\uparrow}$, $N^{\downarrow}$	970, 1024	4898, 1599
MD_O	0.22	0.94
AUC_O	0.80	0.76

Combining the two equations leads to:

$$\begin{aligned} &(shift_1 N^{\uparrow})^2 + ((-\delta - shift_1)N^{\downarrow})^2 \\ &\quad \geq (N^{\uparrow}\epsilon^{\uparrow})^2 + (N^{\downarrow}\epsilon^{\downarrow})^2 \\ &\quad = (N^{\uparrow}\epsilon^{\uparrow})^2 + (N^{\downarrow}(-\delta - \epsilon^{\uparrow}))^2 \ . \end{aligned}$$

Since $shift_1$ is the unique minimum of the function $(xN^{\uparrow})^2 + ((-\delta - x)N^{\downarrow})^2$, we can conclude from this that $shift_1 = \epsilon^{\uparrow}$ and $shift_2 = \epsilon^{\downarrow}$, and hence $SSE(\mathbf{w}) = SSE(\mathbf{w}_{EM}^*)$. Since $\mathcal{D}$ has full column rank, the optimal weights minimizing the SSE is unique, and hence $\mathbf{w}$ must be equal to $\mathbf{w}_{EM}^*$. This concludes the proof, since $\mathbf{w}_{EM}^*$ was obtained from $\mathbf{w}_{EM}$ by changing the weights for the constant factor and the coefficient for x_s only.

The proofs of (3), (4), and (5) are straightforward and have been omitted due to space restrictions. ■

VI. EXPERIMENTS AND RESULTS

We evaluate the proposed techniques for controlling attribute effects in linear regression models using two real datasets. These datasets reflect different application settings and problem domains. The first dataset, Communities and Crime (Crime), is suitable for discrimination-aware regression where a learned regression model is required to not discriminate among communities of different races. The second dataset, Wine Quality (Wine), is appropriate for batch effect control where ratings for red and white wine are required to be normalized. We discuss our techniques for each dataset in the following sections.

A. Communities and Crime

The Communities and Crime (Crime) dataset[1] contains socio-economic information of communities and their crime rates. The goal with this dataset is to learn a model for the crime rate given communities' socio-economic information. Moreover, to prevent discrimination, it is required by law that predictions are not discriminatory between favored and deprived communities based on the majority race of the communities.

For our experiments, we preprocess the dataset by removing attributes with many missing values. We create two groups of the dataset: one group comprises of communities with a majority black population while the other group contains communities with a majority non-black population. All the attributes are standardized to zero mean and unit variance. In the end, the Crime dataset contains 1994 instances (970 and 1024 instances, respectively, for black and non-black

[1] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

TABLE III. CONSOLIDATED OVERALL DATASET RESULTS FOR CRIME AND WINE DATASETS

	Crime					Wine				
	MD_R	AUC_R	MD_O	AUC_O	$RMSE$	MD_R	AUC_R	MD_O	AUC_O	$RMSE$
Data	–	–	0.22	0.80	–	–	– –	0.94	0.76	–
OLS-S	-0.02	0.49	0.20	0.81	0.14	-0.05	0.48	0.90	0.89	0.83
SEM-S	-0.21	0.16	0.01	0.50	0.20	-0.94	0.19	0.00	0.51	0.93
SBR-S	0.00	0.52	0.22	0.82	0.14	0.00	0.50	0.94	0.90	0.83
SEM-MP	-0.05	0.41	0.17	0.76	0.14	-0.46	0.35	0.48	0.82	0.94
OLS-M	-0.01	0.49	0.21	0.80	0.15	-0.04	0.48	0.90	0.89	0.83
SEM-M	-0.04	0.43	0.18	0.75	0.16	-0.14	0.44	0.80	0.85	0.83
SBR-M	0.00	0.51	0.22	0.81	0.15	0.00	0.50	0.94	0.90	0.83

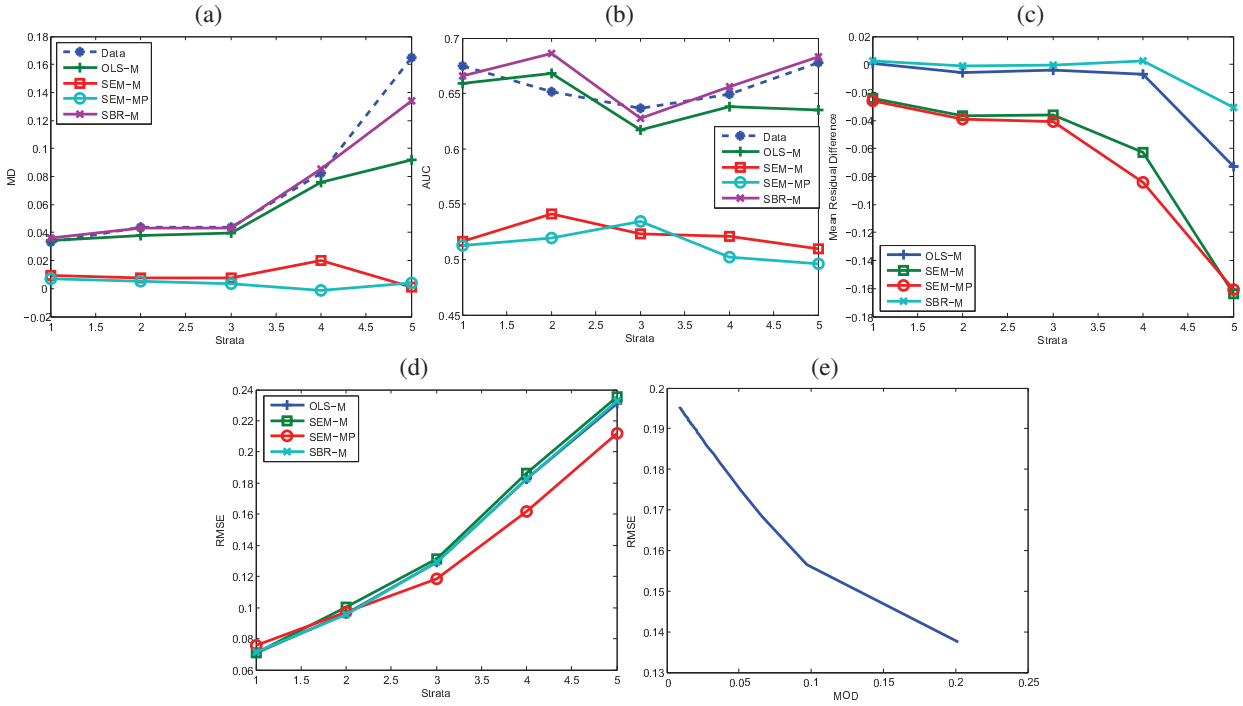

Fig. 1. Experimental results for the Crimes dataset: (a) Mean outcome difference in different strata, (b) Mean residual difference in different strata, (c) Mean residual difference in different strata, (d) RMSE in different strata, (e) Mean difference versus RMSE curve for regularized least squares

communities) described by 99 attributes. Table II shows key characteristics of the Crime dataset.

1) Discrimination Analysis: The Crime dataset shows a strong dependency between the target (Crime Rate) and the sensitive attribute (Race). Communities with a majority black population have a mean crime rate of 0.35 in comparison to a mean crime rate of 0.13 for communities with a majority non-black population ($MD_o = 0.22$ and $AUC_o = 0.80$). However, part of this dependency can be explained by some of the attributes like female divorce percentage and number of illegal immigrants. Such explanatory or confounding attributes can be identified by dependency analysis whereby attributes that are highly correlated to both the target and the sensitive attribute represent potential confounders. Domain background knowledge may be required to filter out attributes that do not contain any objective information for the target.

For our experiments, we identify and utilize four explanatory attributes. Propensity score analysis with quintile stratification reveals that the Crime Rate-Race dependency within each stratum is significantly lower than that in the entire dataset. Figures 1 (a) and 1 (b) show mean outcome difference (MD_o) and AUC value for outcome-sensitive attribute (AUC_o) in each stratum of the dataset (the dashed lines). The average MD_o and AUC_o values over all strata is 0.07 and 0.66, respectively. Thus, part of the dependency or discrimination (i.e. $0.80 - 0.66 = 0.14$ in AUC value) is explainable by the differences in distribution of the explanatory attributes in the two groups. A regression model that eliminates all dependency between Crime Rate and Race can be challenged as unfair and will lead to reverse discrimination in the predictions.

2) Discrimination Control: Here, we present and discuss the results of our attribute effect control techniques when applied to the Crime dataset. All results are obtained from 10-fold cross-validation of the dataset. Furthermore, as required by discrimination law in many regions, the sensitive attribute is not used in the prediction models.

Figure 1 (a), (b), (c), and (d) shows respectively stratum-wise mean outcome difference (MD_o), AUC value for outcome and sensitive attribute (AUC_o), mean residual difference

(MD_r), and root mean square error ($RMSE$) for techniques based on stratification. All techniques, but Strict Equal Means in Multiple Partitions (SEM-MP), involve a separate regression model for each stratum. Single and multiple strata versions of techniques are differentiated by a '-S' or a '-M' at the end of the technique's acronym (SEM-MP has a single model version only).

These results show the effectiveness of the equal outcome mean constraint in ensuring that prediction-sensitive attribute dependency in each stratum is reduced to zero (Figures 1 (a) and 1 (b)). The SEM-M and SEM-MP techniques provide good control as compared to the ordinary least squares technique (OLS-M). It is worth emphasizing that SEM-M and SEM-MP remove the unexplained discrimination only rather than removing all dependency between Crime Rate predictions and Race.

It is interesting to note that the $RMSE$s of the different techniques (Figure 1 (d)) are not significantly different in each stratum. Normally, it is expected that $RMSE$ is higher when enforcing constraints in addition to requiring squared error minimization. However, this behavior is not very obvious on this dataset. The SEM-MP technique has a lower $RMSE$ because it is a single model for the entire dataset rather than different models for each stratum.

The SBR-M technique provides greater control over the residual in the two groups (Figure 1 (d)) but does not reduce discrimination significantly and often performs similarly to the OLS-M technique. Figure 1 (c) also shows how the residuals are impacted when SEM-M and SEM-MP try to balance the predictions in the two groups. The discrimination effect is stronger in the higher strata where probability of black community instances is higher, indicating that greater correction is needed in these strata.

Table III gives the performance of the various techniques when calculated over the entire dataset. These techniques include both single models (like OLS-S, SEM-S, and SEM-MP) and per-stratum models (like SEM-M). The OLS-S technique represents the standard discrimination-ignorant linear model which is observed to magnify discrimination slightly. The SEM-S removes all dependency in the entire dataset, irrespective of whether it is explainable or not, but pays in terms of higher $RMSE$. Notice that the AUC_o values of SEM-M and SEM-MP are 0.76 and 0.75, respectively, even though these techniques reduce the unexplainable dependency in each stratum to near-zero. This is because of the explanatory attributes' influence on the target in the overall dataset (i.e. the overall dataset is not balanced w.r.t. these attributes).

To obtain a better control over the discrimination in linear regression models, a regularized approach can be adopted. Figure 1 (e) shows the MD_o and $RMSE$ plot for different values of the regularization parameter α. The ordinary least squares solution is given by the right most point of the curve where $\alpha = 0$. The value of α can be selected based on the desired level of discrimination over a validation dataset.

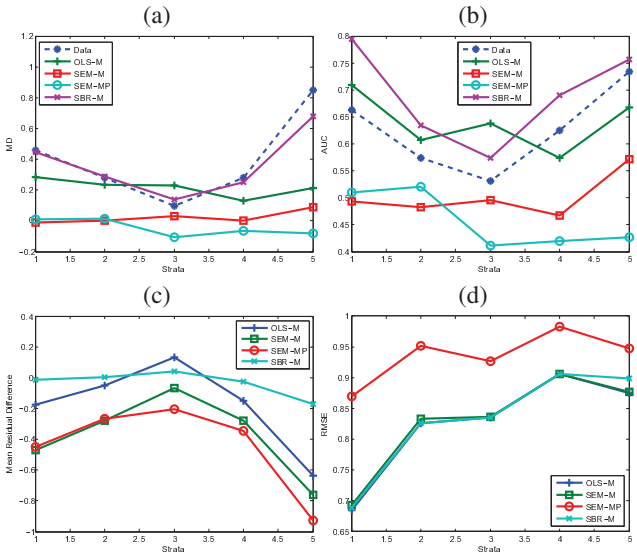

Fig. 2. Experimental results for the Wine Quality dataset: (a) Mean outcome difference in different strata, (b) Mean residual difference in different strata, (c) Mean residual difference in different strata, (d) RMSE in different strata

B. Wine Quality

The Wine Quality (Wine) dataset[2] contains descriptions of red and white wines and their ratings. The wines are described by physical characteristics (such as alcohol content), while the ratings range from 1 to 10. The goal with this dataset is to predict the rating of a wine given its characteristics irrespective of whether it belongs to the red or white wine type. In fact, since the models for the two types of wine are not known and normalized predictions are required without knowledge of type, a rating-normalized linear regression model (without the type attribute) is desired.

The original dataset has a small mean rating difference between the two types of wines. For our experiments, we increase the rating of 70% (randomly selected) white wines by one (with constraint that ratings cannot be greater than 10). The mean rating difference of this modified dataset is 0.94 with the corresponding AUC value of 0.76. The key characteristics of this dataset are given in Table II.

1) Normalized Rating Predictions: We apply our attribute effect control techniques to learn linear regression models that make rating predictions without bias for red or white wine. All results are based on 10-fold cross-validation of the dataset, and wine type is not used in prediction models.

First, we identify the attributes that can potentially explain the dependency between the ratings and types of wine. We select two attributes (volatile acidity and chlorides) as the confounding attributes that need to be adjusted for in the prediction models. Both these attributes have strong correlation with rating and wine type. Propensity score analysis with stratification reveals that the dependency between rating and wine type varies somewhat across the different strata (unlike in the Crime dataset). Thus, wine type effect control over local regions is more appropriate for this dataset.

[2]http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Second, we apply our various attribute effect control techniques for normalized rating predictions. Figures 2 (a), (b), (c), and (d) show the strata-wise performance of the techniques on this dataset. These results follow the general trends observed in the Crime dataset. Both SEM-M and SEM-MP techniques provide good control over the predictions between the two groups as compared to OLS-M and SBR-M. Unlike in the Crime dataset, SEM-M and SEM-MP techniques produce higher $RMSE$ over the biased dataset. An important observation from these results is that SBR-M is more appropriate than OLS-M when attribute effect exaggeration needs to be controlled without requiring effect reduction.

The results for the entire dataset are given in Table III. It is observed that the OLS-S technique significantly exaggerates the ratings difference between the two types of wine. Thus, this technique is not recommended for normalized rating predictions. The SEM-M technique performs the best producing low $RMSE$ while maintaining rating normalization within each stratum. Again, notice that the overall effect in the overall dataset is not reduced to zero even though effects within strata are close to zero for the SEM-M and SEM-MP techniques.

VII. Conclusion

In this paper, we provided a systematic treatment of attribute effect in regression problems. We discussed the nature and motivated the appearance of attribute effects in regression problems, with specific focus on applications in discrimination-aware regression. A key contribution of this work is the introduction of propensity score analysis from statistics for filtering out explainable effects. This strategy can handle multiple explanatory attributes in contrast to previously proposed strategies in the discrimination-aware data mining community. We also defined two measures for quantifying attribute effects in regression problems (mean difference and AUC). We then developed constrained linear regression models for controlling the effect of an attribute on the predictions. Analytical solutions were derived that satisfy equal outcome means and balanced residuals in addition to minimizing the squared error. These techniques allow easy and principled applicability of linear models to control attribute effects. We conducted experiments on two real-world datasets: one dataset represents a discrimination-aware regression problem and the other represents a rating normalization problem. The results show that our techniques are able to achieve perfect control over the training data and this control generalizes well to test data.

To the best of our knowledge, this is the first paper that presents solutions to the discrimination-aware regression problem. Previous works in discrimination-aware data mining were mostly restricted to classification problems.

References

[1] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 560–568.

[2] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," *Machine Learning and Knowledge Discovery in Databases, LNCS*, vol. 7524, pp. 35–50, 2012.

[3] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *IEEE 12th International Conference on Data Mining (ICDM)*. IEEE, 2012, pp. 924–929.

[4] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[5] R. A. Berk, "An introduction to sample selection bias in sociological data," *American Sociological Review*, pp. 386–398, 1983.

[6] L.-F. Lee, "Some approaches to the correction of selectivity bias," *The Review of Economic Studies*, vol. 49, no. 3, pp. 355–372, 1982.

[7] W. Cochran and D. Rubin, "Controlling bias in observational studies: A review," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 417–446, 1973.

[8] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[9] J. Bailar and D. Hoaglin, *Medical uses of statistics*. Wiley, 2012.

[10] T. Petersen and L. A. Morgan, "Separate and unequal: Occupation-establishment sex segregation and the gender wage gap," *American Journal of Sociology*, pp. 329–365, 1995.

[11] D. Weichselbaumer and R. Winter-Ebmer, "A meta-analysis of the international gender wage gap," *Journal of Economic Surveys*, vol. 19, no. 3, pp. 479–511, 2005.

[12] D. A. Smith, C. A. Visher, and L. A. Davidson, "Equity and discretionary justice: The influence of race on police arrest decisions," *J. of Criminal Law & Criminology*, vol. 75, p. 234, 1984.

[13] R. Weitzer and S. A. Tuch, "Race, class, and perceptions of discrimination by the police," *Crime & Delinquency*, vol. 45, no. 4, pp. 494–507, 1999.

[14] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan, "Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset," in *IEEE 12th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2012, pp. 370–377.

[15] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE 10th International Conference on Data Mining (ICDM)*. IEEE, 2010, pp. 869–874.

[16] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[17] C. D. Anderson, J. L. Warner, and C. C. Spencer, "Inflation bias in self-assessment examinations: Implications for valid employee selection." *Journal of Applied Psychology*, vol. 69, no. 4, p. 574, 1984.

[18] V. Skreta and L. Veldkamp, "Ratings shopping and asset complexity: A theory of ratings inflation," *Journal of Monetary Economics*, vol. 56, no. 5, pp. 678–695, 2009.

[19] P. Whiting, A. W. Rutjes, J. B. Reitsma, A. S. Glas, P. M. Bossuyt, J. Kleijnen *et al.*, "Sources of variation and bias in studies of diagnostic accuracy. a systematic review," *Annals of Internal Medicine*, vol. 140, no. 3, pp. 189–202, 2004.

[20] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods," *PloS One*, vol. 6, no. 2, p. e17238, 2011.

[21] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013.

[22] C. Ling, J. Huang, and H. Zhang, "AUC: a statistically consistent and more discriminating measure than accuracy," in *International Joint Conference on Artificial Intelligence*, vol. 18, 2003, pp. 519–526.

[23] P. Rosenbaum and D. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.