FREE: Fairness Regularization with Equalized Error for Regression

by

Weifeng Wang

(Under the Direction of Sheng Li)

Abstract

The prevalence of machine learning applications in decision-making has sparked abundant interest in the fairness of machine learning. Existing notions of fairness are mainly defined upon predictions like equalized odds. This paper characterized the unfairness in regression from a new perspective by inspecting the prediction errors. In particular, we first defined a new fairness measurement with equalized error, which measures the dependence of prediction error on sensitive attributes. We then propose a regularization approach called Fairness Regularization with Equalized Error (*FREE*) which can handle more dimensions of fairness. We conducted two extensive experiments on both simulated datasets and real-world datasets to evaluate our approach's effectiveness in terms of mean square error, Hirschfeld-Gebelein-Rényi (HGR) maximal correlation coefficient, and overlapping index. The results show that our approach reduces unfairness in error more effectively, compared with representative methods.

Index words: [Sensitive attributes, Fairness regularization, Equalized Error, HGR, Overlapping index]

FREE: Fairness Regularization with Equalized Error for Regression

by

Weifeng Wang

M.S., University of Georgia, 2018

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree

Master of Science

Athens, Georgia

2021

FREE: Fairness Regularization with Equalized Error for Regression

by

Weifeng Wang

Major Professor: Sheng Li

Committee: Chenglin Miao
Wenwen Wang

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2021

Acknowledgments

This research work might not have been possible without the support of many people. I want to thank my supervisor Prof. Dr. Sheng Li for invaluable assistance. Also, I would like to thank my parents and wife for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Fairness has become an emerging research topic in machine learning Mehrabi et al., 2019. In general, fair machine learning methods aim to learn a mapping $h(X, S)$ for a target variable Y using input features X and a protected/sensitive attribute S (e.g, race, gender), while ensuring fairness with respect to S Barocas et al., 2017; Chouldechova and Roth, 2018. For instance, when we predict the probability that a defendant will be a recidivist, we would like the algorithm to not unfairly treat the groups with sensitive attribute *Race* Dressel and Farid, 2018. Previous work has shown that standard machine learning methods usually cause unfairness. An intuitive solution is to simply train models by excluding the sensitive attributes, however, such a strategy still could not reduce fairness. One possible reason is that, in most cases, the sensitive attribute S may be correlated with other attributes. For instance, personal income varies significantly by age/race. In addition, the prediction is estimated from independent variables that are usually correlated with sensitive attributes. For example, education level is correlated with race. To obtain a more accurate model, the error between target and prediction need to be reduced, which increases the dependence between prediction and sensitive attribute and renders the prediction biased.

There have been mainly three strategies to improve algorithmic fairness Hajian and Domingo-Ferrer, 2012. The first one is pre-processing approach. It trains the model using a new representation, which removes the information correlated to the sensitive attributes S and obtains the information of X Calmon et al., 2017; Louizos et al., 2015. The second is to add fairness penalty in the objective function at the training time Agarwal et al., 2018; Kamishima et al., 2011. The third one is post-processing approach, which applies transformations to model output and reduce prediction unfairness Hardt et al., 2016; Kamiran et al., 2010. However, much of the work to date has focused on classification with binary targets, where standard fairness notions include equal false positive or negative rates across different populations. Less attention has been paid to fairness in regression, where the target is continuous Agarwal et al., 2019; Berk et al., 2017; Okray et al., 2019. Existing fair regression methods mainly exploit the conditional independence of model prediction $\hat{y}$ and the sensitive attribute S.

In this paper, we identify the source of unfairness in regression from a new perspective based on prediction errors. The idea originated from the assumption that the error term ϵ in a general regression problem (Eq. (1.1)) is independent of covariates X:

$$y = h(X) + \epsilon, \tag{1.1}$$

where y is a continuous target variable. In practice, however, there might be complicated structures in the error term ϵ, e.g., ϵ is a function of S. Figure 1.1 shows that models without considering fairness and models considering only dependence in $(\hat{y}, S)$ cause unfairness in terms of prediction error, as the distributions of prediction errors on two groups (High_AA and Low_AA) are not clearly overlapped. Sensitive attributes are usually correlated with other attributes or the target variable. For example, in the Communities and Crime dataset, the target variable *ViolentCrimesPerPop* and sensitive attribute *racepctblack* are highly correlated with a Pearson's correlation coefficient 0.636. Hence, penalizing on the dependence of prediction and sensitive attributes would suffer an increased loss of accuracy. In addition, for the sake of fair decision making, we should not overestimate or underestimate any groups in the sensitive attribute. This implies that the distribution of prediction error should be independent of sensitive attribute.

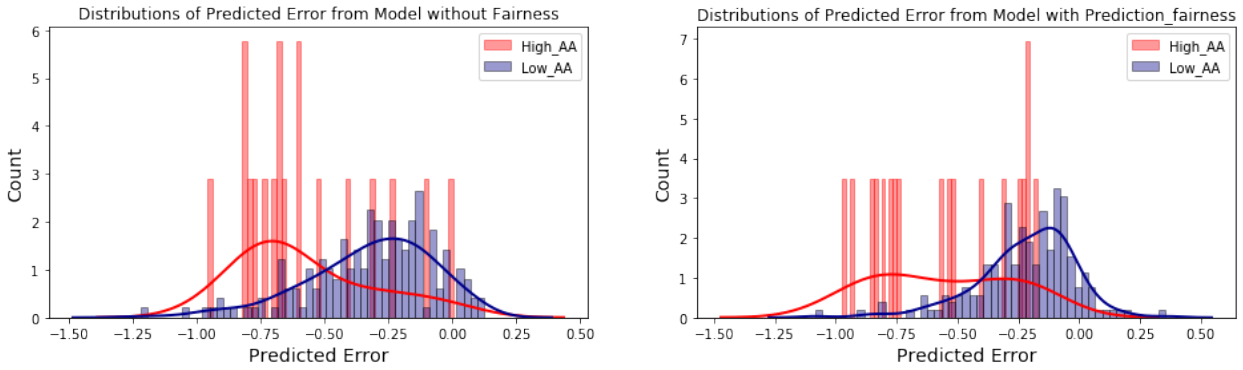

Figure 1.1: Distributions of Prediction Error across Sensitive Groups for the Communities and Crime dataset (i.e., High_AA: ratio of African American Residents $\geq 50\%$ and Low_AA otherwise).

1.2 Contributions

We aim to achieve fairness in regression by fully exploiting the independence across model prediction, prediction errors and sensitive attributes. In particular, we incorporate the fairness awareness into the regression regularization framework, and consider two different cases with continuous sensitive attributes and categorical sensitive attributes. For continuous sensitive attribute, we employ the Hirschfeld-Gebelein-Rényi (HGR) correlation coefficient Mary et al., 2019 of the prediction error distribution and the sensitive

attribute distribution to measure the fairness. For categorical sensitive attributes, we use the HGR correlation coefficients and overlapping index Pastore and Calcagnì, 2019 as fairness measurement. The main contributions of this paper are as follows:

- We identify the unfairness in regression from a new perspective based on prediction error.
- We propose a new fairness regularization approach that employs a new fairness measurement named equalized error. Both continuous and discrete sensitive attributes are modeled by using different metrics.
- We conduct extensive experiments on multiple simulated and real-world datasets, and the results show that our approach reduced more dependence between prediction error and sensitive attribute than baselines and our approach can handle both fairness in error and fairness in prediction. We also show that it is easier to control fairness in error than in prediction.

1.3 Thesis Structure

This thesis is structured as follows: Chapter 2 provides preliminary work about fairness in machine learning. In Chapter 3, we define the new fairness measurement and construct the fair regression model called Fairness Regularization with Equalized Error (FREE). Chapter 4 shows the model performance on simulated datasets and real-world dataset. Finally, we conclude and discuss future work in Chapter 5.

Chapter 2

Fairness in Machine Learning

2.1 Preliminary Work

The above objectives are motivated by the fact that available historic data are usually biased due to discrimination such as race and gender. Simply using the standard approaches in which the sensitive attributes are ignored in training process may still biase the results.

2.1.1 Definitions of Fairness

Numerous definitions of fairness were proposed in recent years. However, there is no clear agreement on which definition is the most appropriate Dwork et al., 2012; Hardt et al., 2016; Zafar, Valera, Gomez Rodriguez, et al., 2017. Two popular fairness definitions are statistical parity and equalized odds.

Definition 1 (Statistical Parity) *A predictor $h(X, S)$ for target Y satisfies statistical parity with respect to the sensitive attribute S if $h(X)$ is independent of S.*

Definition 2 (Equalized Odds) *A predictor $h(X, S)$ for target Y satisfies equalized odds with respect to the sensitive attribute S if $\hat{Y}$ is independent of S conditioned on Y.*

In addition, Chouldechova, 2017 proposed a definition of fairness based on false error rates, and they tried to balance the rates across protected and unprotected groups. Treatment equality introduced by Berk et al., 2018 looks at the ratio of errors that the classifier produces rather than its accuracy. A classifier satisfies this definition when both protected and unprotected groups have an equal ratio of false negatives and false positives. To avoid unfairness in classification, Zafar, Valera, Gomez Rodriguez, et al., 2017 and Zafar, Valera, Rogriguez, et al., 2017 introduced the notion of unfairness, disparate mistreatment. Moreover, Kamiran et al., 2010 introduced a fairness-aware decision tree learner by considering fairness gain in its splitting criterion and pruning strategy.

2.2 Fair Regularization

Much work has considered the algorithmic fairness in Machine learning based on regularization approach Kamishima et al., 2011. Recall the objective function of fairness learning:

$$argmin_{h \in H} L(h, X, Y) + \lambda * FP, \tag{2.1}$$

where h is a regressor or classifier from a family of regressors or classifiers H for target variable Y using input X, $L()$ is the loss function, FP is the fairness penalty term.

2.2.1 Fair Regression

For regression problems with discrete sensitive variable, Calders et al., 2013 firstly introduced measures including Mean difference and AUC and imposed them on prediction and residuals. Our method differs from these because we consider the dependence of prediction error and sensitive attribute on distributions/densities instead of just the point estimate mean. And our models can handle both continuous and discrete sensitive attributes in regression problems. More recently, Agarwal et al., 2019 proposed two different constraints based on a relaxation of statistical parity and the bounded group loss criteria onto the objective function for the sake of flexibility.

For regression problems with discrete/continuous sensitive variables, Mary et al., 2019 used the Rényi maximum correlation coefficient of prediction and sensitive attribute to generalize the fairness penalty in regression. Grari et al., 2019 considered the dependence of error and sensitive attribute to study the effects of different approximation methods for Rényi maximum correlation coefficient. Independent of their work, our work focused on the difference between fairness in prediction and fairness in prediction error and provided a more general regression framework to handle more dimensions of fairness. What's more, our initial idea came from the assumption of noise term ϵ in the regression framework. And Narasimhan et al., 2020 introducted the pairwise fairness metrics for ranking and regression problems with discrete and continuous sensitive attributes. Authors in Steinberg et al., 2020 introduced fairness regression by incorporating fast approximations of the independence, separation and sufficiency group fairness criteria based on mutual information. To better understand the optimal solution for fair regression, Chzhen et al., 2020 shows that the relationship between fair regression problems with the Demographic Parity constraint and the problem of Wasserstein barycenters.

In addition, for fairness-aware multi-task regression, Zhao and Chen, 2019 added a non-convex constraint based on the group-wise ranking functions of individuals, by using a rank-based non-parametric independence test of the target variable and protected variables. Pérez-Suay et al., 2017 solved the fairness regression problem in kernel space, and Okray et al., 2019 further extended this method by learning fair feature embeddings in the kernel space. It minimizes prediction loss while additionally penalizing the correlation between the prediction and sensitive attribute in the kernel space.

Existing work on fair regression is mainly based on the independence of model prediction and sensitive attribute. Our approach differs from the existing fair regression methods in that we consider the

dependence of prediction error and sensitive attribute on distributions/densities, instead of just the point estimate. We propose a regularization approach, which can handle more dimensions of fairness. And our approach can handle both continuous and discrete sensitive attributes in regression problems.

Chapter 3

Fairness Regularization with Equalized Error (FREE)

In this chapter, we define a new fairness measurement based on equalized error. Then, by leveraging the regularization approach Kamishima et al., 2011, we propose a fair regression approach called Fairness Regularization with Equalized Error (FREE).

The regularization approach has been widely adopted in existing work in order to achieve algorithmic fairness Kamishima et al., 2011. In general, the fairness regularization framework for regression could be formulated as:

$$\arg\min_{h\in\mathcal{H}} \mathcal{L}(h(X), Y) + \lambda FP, \tag{3.1}$$

where h is a regressor from a family of regressors $\mathcal{H}$ for target variable Y using input X. $\mathcal{L}(\cdot)$ denotes the loss function. FP is a fairness penalty term, and λ is a trade-off parameter. The key idea of fairness regularization is to penalize the dependence of two random variables (one of them is the sensitive attribute), by customizing the fairness penalty term FP.

3.1 Fairness Definition: Equalized Error

We consider a fair regression problem that involves sensitive (i.e., protected) attributes. Let (X, S, Y) denote the training data, where X is an input feature matrix except the sensitive attribute, S is a sensitive attribute, and Y is a continuous target variable. The goal of fair regression is to learn an accurate regressor $h(X)$ from a set of regressors $\mathcal{H}$, such as linear threshold rules or neural networks, while satisfying some fairness constrains.

Recall the regression framework detailed in (Eq. (1.1)) which assumes the error ϵ is independent with (X, S). Then, we define a new fairness measurement, which characterizes the fairness with regards to prediction error.

Definition 3 (Equalized Error) *A regressor h satisfies Equalized Error under a distribution over (X, S, Y) if its prediction error $\epsilon(X, Y) = h(X) - Y$ is independent of the sensitive attribute S, that is,*

$$P[\epsilon(X, Y) \leq z|S = s] = P[\epsilon(X, Y) \leq z],$$

for all $s \in S$ and all $z \in \mathbb{R}$.

Our work is independent with that in paperGrari et al., 2019 which used the same formula to measure fairness of regression. While our initial idea came from the assumption of noise term ϵ in (Eq. (1.1)).

3.2 Fairness Measurements

With the proposed equalized error for fairness measurement, we present two different fairness penalty terms to deal with continuous and categorical sensitive attributes. The first one is based on the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation coefficient Rényi, 1959, which is able to handle both continuous and categorical variables.

Definition 4 (Hirschfeld-Gebelein-Rényi (HGR) Maximal Correlation Coefficient Rényi, 1959) *For random variables $U \in \mathcal{U}$ and $V \in \mathcal{V}$, the Hirschfeld-Gebelein-Rényi(HGR) Maximal Correlation Coefficient is defined as follow,*

$$HGR(U, V) = \sup_{f,g} \rho(f(U), g(V)),$$

where ρ is the Pearson's correlation coefficient and f, g are measurable functions with $E[f^2(U)], E[g^2(V)] < \infty$.

The fairness penalty term FP based on HGR is defined as: $\rho_{\text{HGR}}(\epsilon, S)$, which measures the dependency of prediction error ϵ on the sensitive attribute S.

We have $0 \leq HGR(U, V) \leq 1$. $HGR(U, V)$ closing to 1 means high dependence between U and V. $HGR(U, V) = 0$ iff V and U are independent. We did not use the well known measurement Pearson Correlation Coefficient, because it measures the linear correlation between two variables. In practice, the correlation between target Y and sensitive attribute S may be complicated. And U and V are independent implies that $\rho(U, V) = 0$, but the converse is not true.

The second one is based on the overlapping (OL) index, which can deal with categorical sensitive attributes. In particular, the OL index is defined as $\eta(f, g) = \int_{x \in R} \min[f(x), g(x)]dx$. $\eta(f, g) = 0$ indicates that densities $f(x)$ and $g(x)$ are distinct. It can be used to measure the distribution similarity of prediction error densities across protected groups. Also, a distribution-free approximation of overlapping index has been introduced in Pastore and Calcagnì, 2019. Then, the fairness penalty term FP based on OL is defined as:

$$\rho_{\text{OL}}(\epsilon, S) = \eta(\epsilon_{X_a}, \epsilon_{X_{a^c}}),$$

where X_a is the set of instances with sensitive attribute $S = a$, a is a subgroup in sensitive attribute S, X_{a^c} is the the complement of set X_a, and ϵ_{X_a} is the probability density of prediction error for instances with sensitive attribute $S = a$.

3.3 Final Model

By using the proposed fairness regularization terms based on equalized error, the objective function for fair regression in Equation (3.1) can be rewritten as:

$$\arg\min_{h\in\mathcal{H}} \mathcal{L}(h(X), Y) + \lambda\rho(\hat{\epsilon}, S), \tag{3.2}$$

where $\hat{\epsilon}$ is the estimation of ϵ.

For continuous sensitive attribute S, we adopt the HGR approximation approach Mary et al., 2019 and use the fairness penalty term $\rho_{\text{HGR}}(\hat{\epsilon}, S)$. To estimate the density of probability distribution, we used Gaussian KDE and set the bandwidth based on the Silverman's rule Silverman, 1986. For categorical sensitive attribute, $\rho(\hat{\epsilon}, S)$ could be implemented by either $\rho_{\text{HGR}}(\hat{\epsilon}, S)$ or $\rho_{\text{OL}}(\hat{\epsilon}, S)$. For regression task, a commonly used loss function $\mathcal{L}(h(X), Y)$ is the Mean Squared Error loss function.

Compared with existing fair regression methods that focus on fairness in terms of prediction $\hat{y} = h(X)$, our model in Eq. (3.2) deals with potential unfairness related to prediction error ϵ. Moreover, we notice that the potential unfairness in $\hat{y}$ and $\hat{\epsilon}$ are not contradictory. Instead, they could jointly present a comprehensive characterization of unfairness in regression problems. Motivated by this insight, we try to penalize the dependence in $(\hat{y}, S)$ and $(\hat{\epsilon}, S)$ simultaneously, and thus propose a generalized fair regression model as follows:

$$\arg\min_{h\in\mathcal{H}} \mathcal{L}(h(X), Y) + \lambda(\mu \cdot \rho(\hat{y}, S) + (1-\mu) \cdot \rho(\hat{\epsilon}, S)), \tag{3.3}$$

where $\lambda \in [0, 1]$ is a hyper-parameter balancing the trade-off between accuracy and fairness, $\mu \in [0, 1]$ is a hyper-parameter balancing the trade-off between fairness in prediction $h(X)$ and fairness in ϵ, $\hat{\epsilon}$ is the estimation of ϵ, and $\rho(.)$ is a fairness penalty term that could be implemented by HGR, OL, or other metrics.

Many existing fair regression models could be considered as special cases of Eq. (3.3). For instance, by setting μ to 1, the model Eq. (3.3) is conceptually equivalent to existing fair regression models that solely rely on dependence of prediction $\hat{y}$ on sensitive attribute S.

Chapter 4

Experiments

In this chapter, we evaluate the performance of the proposed approaches on both simulated and real-world datasets. Following a similar setting in Mary et al., 2019, we adopt a simple neural network which has two hidden layers (50 neurons in the first layer and 30 neurons in the second layer) and scaled exponential linear unit (SELU). Also, the Adam optimization method is used, and the learning rate is set to 10^{-6}.

4.0.1 Data Simulation

We generate samples based on the dependence of (y, S) and (ϵ, S) to study how different fairness penalties influence the model accuracy and dependence in $(\hat{y}, S)$ and $(\hat{\epsilon}, S)$, where $\hat{\epsilon}$ is the estimation of ϵ. First, we generate $20,000$ samples from normal distributions and construct simulated datasets (X, y_i, S_i) for $i = 1, 2, 3, 4$ in Section 4.1. In Section 4.2, to construct simulated data with binary sensitive attribute, we binarized S_i from Section 4.1 by re-coding S_i as (1: High_AA) if it is ≥ 0.5 and (0: Low_AA) otherwise. Each dataset is split into a training set (70% of samples) and a test set (30% of samples). The details of four simulated datasets are shown as follows.

- ***Data 1-1*** (X, y_1, S_1)*: y correlated with S and error independent of S* (i.e., $y \not\perp S, \epsilon \perp S$). The data generating functions are $y_1 = 10x_1 + error_1, S_1 = x_1 - x_2, error_1 = x_1 + x_2$.

- ***Data 1-2*** (X, y_2, S_2)*: y and error correlated with S* (i.e., $y \not\perp S, \epsilon \not\perp S$). The data generating functions are $y_2 = 3x_1 - 5x_2 + \epsilon_2, S_2 = x_1 - x_2, \epsilon_2 = x_2$.

- ***Data 1-3*** (X, y_3, S_3)*: y and error independent of S* (i.e., $y \perp S, \epsilon \perp S$). The data generating functions are $y_3 = 10x_1 + 10x_2 + \epsilon_3, S_3 = x_1 - x_2, \epsilon_3 = x_1 + x_2$.

- ***Data 1-4*** (X, y_4, S_4)*: y independent of S and error correlated with S* (i.e., $y \perp S, \epsilon \not\perp S$). The data generating functions are $y_4 = 10x_1 + 9.99x_2 + 0.01\epsilon_4, S_4 = x_1 - x_2, \epsilon_4 = x_2$.

4.0.2 Baselines and Settings

We compare our approaches with following two baselines. (1) ***NO_Fair*** model. It is a standard regression model implemented by the aforementioned 2-layer neural networks, which does not consider any fairness constraints. (2) ***Pred_Fair*** model Mary et al., 2019. It is a representative fair regression method in literature, which only considers the dependence in $(\hat{y}, S)$. Our approach ***FREE_Err*** model (Equation (3.2)) only considers dependence in $(\hat{\epsilon}, S)$, while our ***FREE_PE*** model (Equation (3.3)) considers dependence in both $(\hat{y}, S)$ and $(\hat{\epsilon}, S)$. Because the HGR correlation coefficient is in $[0, 1]$, we normalize the input data before model training, which makes MSE and HGR correlation coefficient in a common scale. Moreover, to examine the robustness of the models, we ran each model 100 times and compare the MSE, HGR correlation coefficient between $\hat{y}_i$ and S_i, and HGR correlation coefficient between ϵ_i and S_i for $i = 1, 2, 3, 4$. Hyperparameters in baselines and our approaches (e.g., λ and μ) are determined by cross-validation on the training set.

4.1 Simulation 1: Continuous Sensitive Attribute

4.1.1 Data 1-1: y correlated with S and error independent of S

In the simulated Data 1-1: $y_1 = 10 * x_1 + error_1, S_1 = x_1 - x_2, error_1 = x_1 + x_2$, the HGR correlation coefficient between y and S is 0.586 and HGR correlation coefficient between $error$ and S is 0.028. As shown in Figure 4.1, all models obtained high accuracy on MSE, for the correlation between prediction and sensitive attribute, all models resulted high HGR correlation coefficient value above 0.5; while for the correlation between prediction error and sensitive attribute, only the new method reduced the HGR correlation coefficient to 0.2. These indicate that the new method works better than the other models.

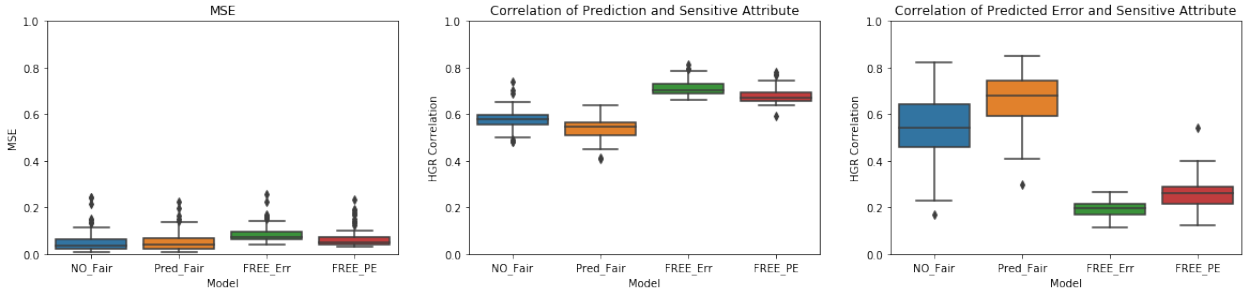

Figure 4.1: Model Comparison (MSE & HGR) for Simulated Data 1-1

4.1.2 Data 1-2: y and error correlated with S

In the simulated Data 1-2: $y_2 = 3 * x_1 - 5 * x_2 + error_2, S_2 = x_1 - x_2, error_2 = x_2$, the HGR correlation coefficient between y and S is 0.951 and HGR correlation coefficient between $error$ and S

is 0.659. As shown in Figure 4.2, all models obtained high accuracy on MSE, for the correlation between prediction and sensitive attribute, all models generated high HGR correlation coefficient values that were above 0.8; while for the correlation between prediction error and sensitive attribute, only the model considering Error fairness reduced the HGR correlation coefficient to below 0.3.

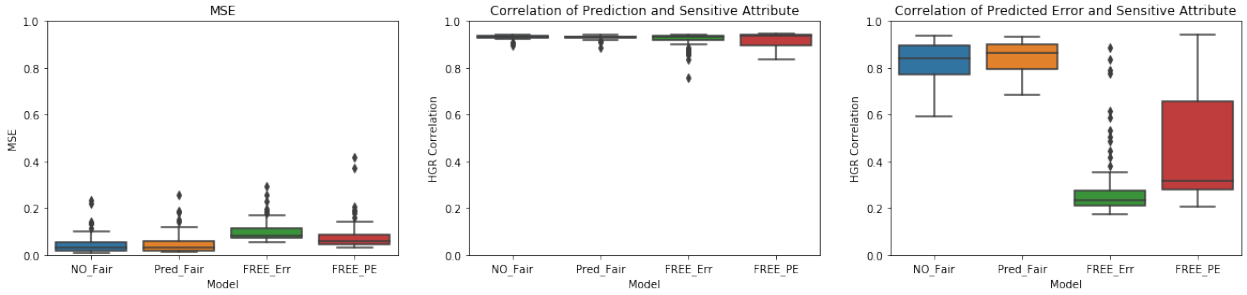

Figure 4.2: Model Comparison (MSE & HGR) for Simulated Data 1-2

4.1.3 Data 1-3: y and error independent of S

In the simulated Data 1-3: $y_3 = 10 * x_1 + 10 * x_2 + error_3$, $S_3 = x_1 - x_2$, $error_3 = x_1 + x_2$, the HGR correlation coefficient between y and S is 0.028 and HGR correlation coefficient between $error$ and S is 0.028. As shown in Figure 4.3, all models obtained high accuracy on MSE, for the correlation between prediction and sensitive attribute, all models had low HGR correlation coefficient value around 0.1; while for the correlation between prediction error and sensitive attribute, only the new method reduced the HGR correlation coefficient to 0.2.

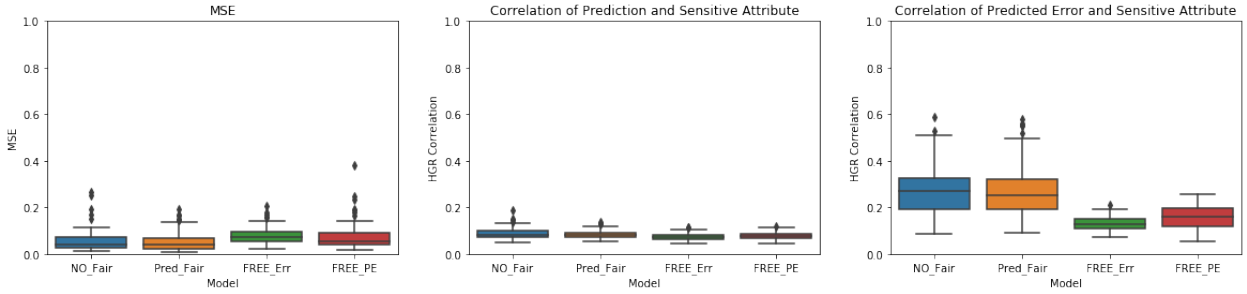

Figure 4.3: Model Comparison (MSE & HGR) for Simulated Data 1-3

4.1.4 Data 1-4: y independent of S and error correlated with S

In the simulated Data 1-4: $y_4 = 10 * x_1 + 9.99 * x_2 + 0.01 * error_4$, $S_4 = x_1 - x_2$, $error_4 = x_2$, the HGR correlation coefficient between y and S is 0.028 and HGR correlation coefficient between $error$

and S is 0.659. As shown in Figure 4.4, all models obtained high accuracy on MSE, for the correlation between prediction and sensitive attribute, all models had low HGR correlation coefficient value to 0.2; while for the correlation between prediction error and sensitive attribute, only the new method reduced the non-significant HGR correlation coefficient to below 0.2.

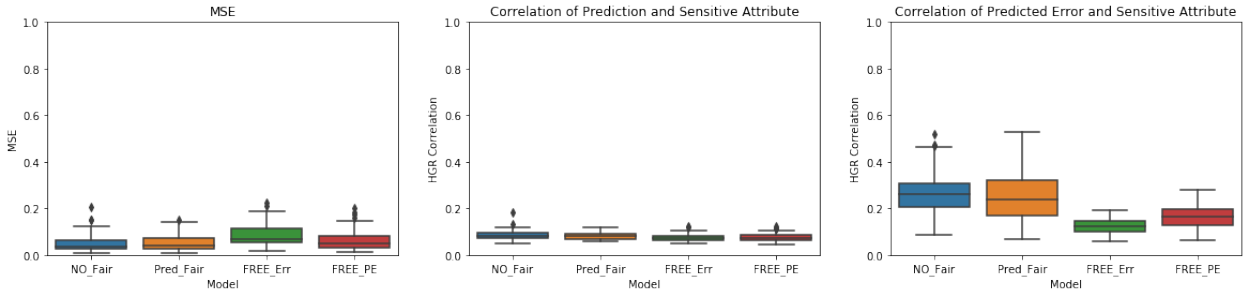

Figure 4.4: Model Comparison (MSE and HGR) for Simulated Data 1-4

Table 4.1 summarized the results for all simulated datasets with continuous sensitive attribute. It shows that, fair regression methods usually obtain slightly larger MSE values, as they sacrifice the model performance due to fairness penalties. At the same accuracy MSE level, our proposed FREE models would obtain lower HGR correlation coefficients of prediction error and sensitive attribute than Pred_Fair model and NO_Fair model without increasing the HGR correlation coefficients of prediction and sensitive attribute. Although Pred_Fair* reduces the fairness to negligible, it increases the MSE by over 12 times. In Data1-1 and Data1-2 in which $y \not\perp S$, all models cause unfairness in $\hat{y}$, because y is highly correlated to S and $\hat{y}$ is close to y. It shows that NO_Fair and Pred_Fair would cause unfairness in prediction error with values of HGR$(\hat{\epsilon}, S)$ above 0.3 even thought when $\epsilon \perp S$. While our models FREE_Err and FREE_PE will control the fairness in error. In Data1-3 and Data1-4 with $y \perp S$, all models provide a non-significant HGR correlation coefficients. Our models would control the fairness in prediction error by reducing the HGR$(\hat{\epsilon}, S)$ below 0.3. However, models NO_Fair and Pred_Fair would still cause unfairness in error with values of HGR$(\hat{\epsilon}, S)$ above 0.3. These results show that it is easier to control fairness in prediction error especially when the response y is highly correlated with sensitive attribute S.

4.2 Simulation 2: Binary Sensitive Attribute

This section explores the performance of our approaches (FREE_Err and FREE_PE) and baselines (Pred_Fair and NO_Fair) on the simulated datasets with binary sensitive attribute. To construct simulated data with binary sensitive attribute, we binarized S_i from Section 4.1 by re-coding S_i as (1: High_AA) if it is ≥ 0.5 and (0: Low_AA) otherwise. In this way, in accordance with Data 1-i ($i = 1, 2, 3, 4$), we obtain four modified simulated datasets Data 2-i ($i = 1, 2, 3, 4$). To evaluate the model performance, we employ the following measurements: MSE, HGR $(\hat{y}, S)$, HGR $(\hat{\epsilon}, S)$ and OL$(\hat{\epsilon}, S)$, where $\hat{\epsilon}$ is the estimation of ϵ. In addition, to illustrate the distribution similarity of prediction error, we visualize the distributions of

Table 4.1: Simulation Results (mean±std) for Continuous Sensitive Attributes.

Scenario	Model	MSE	HGR($\hat{y}, S$)	HGR($\hat{\epsilon}, S$)
Data1-1:	NO_Fair	$.05 \pm .05$	$.58 \pm .04$	$.54 \pm .14$
$y \not\perp S$,	Pred_Fair	$.05 \pm .04$	$.54 \pm .04$	$.67 \pm .11$
	Pred_Fair*	$.62 \pm .36$	**.17** $\pm .04$	$.76 \pm .16$
$\epsilon \perp S$	FREE_Err	$.09 \pm .04$	$.71 \pm .03$	**.19** $\pm .04$
	FREE_PE	$.07 \pm .04$	$.68 \pm .03$	**.26** $\pm .06$
Data1-2:	NO_Fair	$.04 \pm .04$	$.93 \pm .01$	$.83 \pm .08$
$y \not\perp S$,	Pred_Fair	$.05 \pm .05$	$.93 \pm .01$	$.85 \pm .07$
	Pred_Fair*	$.96 \pm .18$	**.18** $\pm .15$	$.91 \pm .07$
$\epsilon \not\perp S$	FREE_Err	$.10 \pm .04$	$.92 \pm .03$	**.28** $\pm .14$
	FREE_PE	$.08 \pm .06$	$.92 \pm .03$	**.45** $\pm .24$
Data1-3:	NO_Fair	$.06 \pm .05$	$.09 \pm .02$	$.27 \pm .10$
$y \perp S$,	Pred_Fair	$.05 \pm .04$	$.08 \pm .02$	$.27 \pm .11$
$\epsilon \perp S$	FREE_Err	$.08 \pm .04$	$.08 \pm .01$	**.13** $\pm .03$
	FREE_PE	$.07 \pm .05$	$.08 \pm .02$	$.16 \pm .05$
Data1-4:	NO_Fair	$.05 \pm .04$	$.09 \pm .02$	$.27 \pm .09$
$y \perp S$,	Pred_Fair	$.05 \pm .03$	$.08 \pm .01$	$.26 \pm .11$
$\epsilon \not\perp S$	FREE_Err	$.09 \pm .03$	$.07 \pm .02$	**.13** $\pm .03$
	FREE_PE	$.06 \pm .04$	$.08 \pm .02$	$.16 \pm .05$

NO_Fair: models without considering any fairness;
Pred_Fair[Mary etal., 2019]: models only considering fairness in prediction;
FREE_Err: FREE models only considering fairness in prediction error;
FREE_PE: FREE models considering both fairness in prediction and fairness in prediction error.

prediction error for two groups and checked the overlapping index of the densities in Figure 4.7, Figure A.2, Figure 4.10 and Figure A.5.

4.2.1 Data 2-1: y correlated with S and error independent of S

When y is correlated with binary S and $error$ is independent of binary attribute S with $HGR(y, S) = 0.623$ and $HGR(\epsilon, S) = 0.018$, Figure 4.5 shows that, at the same accuracy level, all models had significant correlation in $(\hat{y}, S)$ with $HGR(\hat{y}, S)$ around 0.7; while only the new proposed models reduced correlation in $(\hat{\epsilon}, S)$ to non-significant correlation to 0.2. Compared to other methods, our new method is more accurate in predicting errors. Figure 4.7 shows that for models without considering fairness and model considering only the correlation in $(\hat{y}, S)$, there is an obvious difference in the overlap of prediction error densities across the sensitive subgroups; while our new models obtain no significant difference in the overlap of prediction error densities across the sensitive subgroups. The overlapping indexes of model without considering fairness and model considering only the prediction fairness have means 0.267, 0.191 respectively which are much smaller than 0.903 and 0.864 from model considering prediction error fairness.

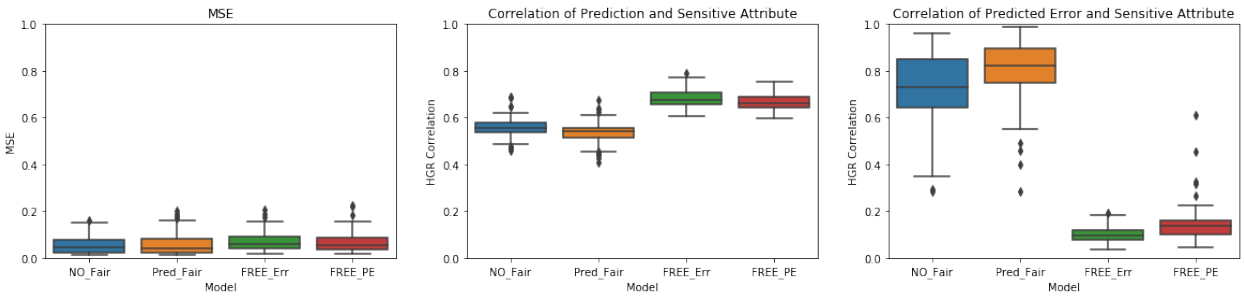

Figure 4.5: Model Comparison (MSE & HGR) for Simulated Data 2-1 with Binary S

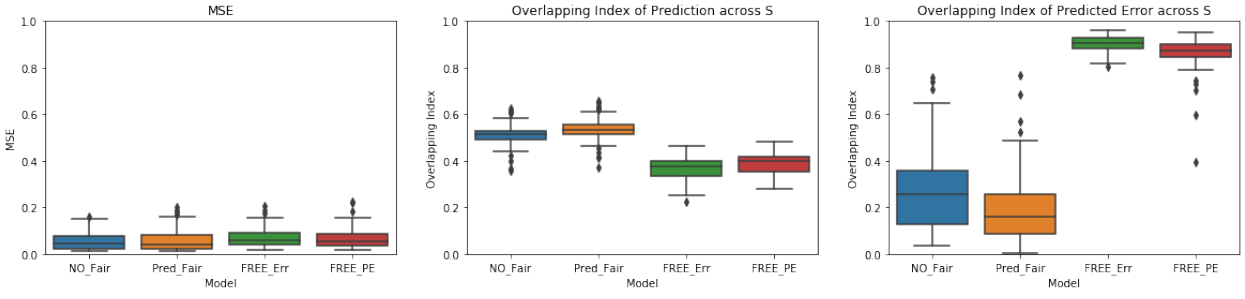

Figure 4.6: Model Comparison (MSE & Overlapping Index) for Simulated Data 2-1 with Binary S

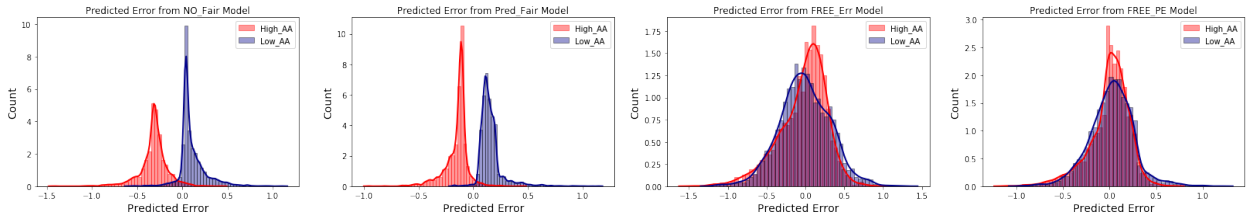

Figure 4.7: Model Comparison (Prediction Error) for Simulated Data 2-1 with Binary S

4.2.2 Data 2-2: y and error correlated with S

When y and $error$ are correlated with binary attribute S with $HGR(y, S) = 0.883$ and $HGR(\epsilon, S) = 0.554$, Figure 4.8 and Figure A.2 gave the same conclusions as that from simulated Data 2-1.

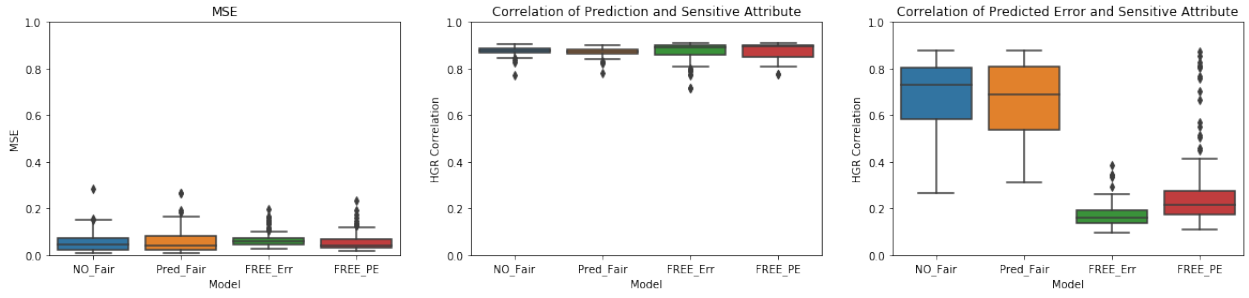

Figure 4.8: Model Comparison (MSE & HGR) for Simulated Data 2-2 with Binary S

4.2.3 Data 2-3: y and error independent of S

When y and $error$ are independent of binary attribute S with $HGR(y, S) = 0.018$ and $HGR(\epsilon, S) = 0.018$, Figure 4.9 shows that, at the same accuracy level, all models had non-significant correlation both in $(\hat{y}, S)$ and $(\hat{\epsilon}, S)$ and the new methods FREE_Err and FREE_PE reduced the HGR correlation coefficient $HGR(\hat{\epsilon}, S)$ to 0.1. Figure A.3 and Figure 4.10 show that all models obtained high overlapping areas in prediction densities and prediction error densities, and FREE_Err model and FREE_PE model had higher overlapping index values than the other models.

4.2.4 Data 2-4: y independent of S and error correlated with S

When y is independent of binary S and $error$ is correlated with binary attribute S with $HGR(y, S) = 0.623$ and $HGR(\epsilon, S) = 0.554$, Figure 4.11 shows that, all models obtained high accuracy on MSE, for

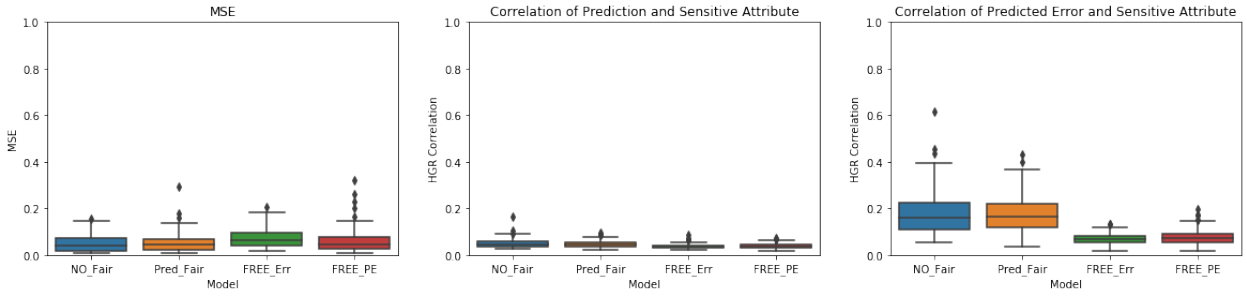

Figure 4.9: Model Comparison (MSE & HGR) for Simulated Data 2-3 with Binary S

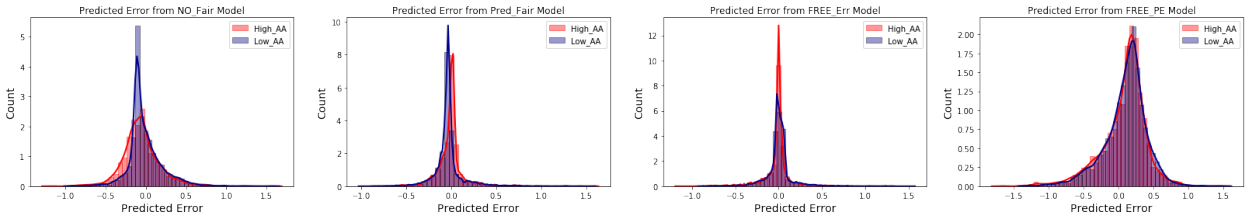

Figure 4.10: Model Comparison (Prediction Error) for Simulated Data 2-3 with Binary S

the correlation between prediction and sensitive attribute, all models had non-significant HGR correlation coefficient value to below 0.1; while for the correlation between prediction error and sensitive attribute, the new method reduced the non-significant HGR correlation coefficient to 0.1. Figure A.5 shows that models considering fairness have more overlap in prediction error densities across the sensitive subgroups. Thus, our new method perform better than the others on the fairness in the prediction error.

Table 4.2 summarizes the results of our approaches and baselines on the simulated datasets with binary sensitive attribute. In the Data 2-1, y is correlated with binary S and ϵ is independent of binary attribute S, $\text{HGR}(y, S) = 0.62$, and $\text{HGR}(\epsilon, S) = 0.02$. Experimental results show that all the compared methods obtain low values of MSE. Moreover, at the same level of MSE, all models have significant correlation in terms of $(\hat{y}, S)$, with $\text{HGR}(\hat{y}, S)$ around 0.7. The baselines NO_Fair and Pred_Fair obtain much higher values of $\text{HGR}(\hat{\epsilon}, S)$ (around 0.80) than our approaches FREE_Err and FREE_PE. Our approaches could reduce the $\text{HGR}(\hat{\epsilon}, S)$ to 0.10. The results imply that the baseline methods cannot satisfy the fairness measurement based on equalized error. As for the metric $\text{OL}(\hat{\epsilon}, S)$, a higher value indicates a stronger overlap, and our approaches obtain much higher results than the baselines.

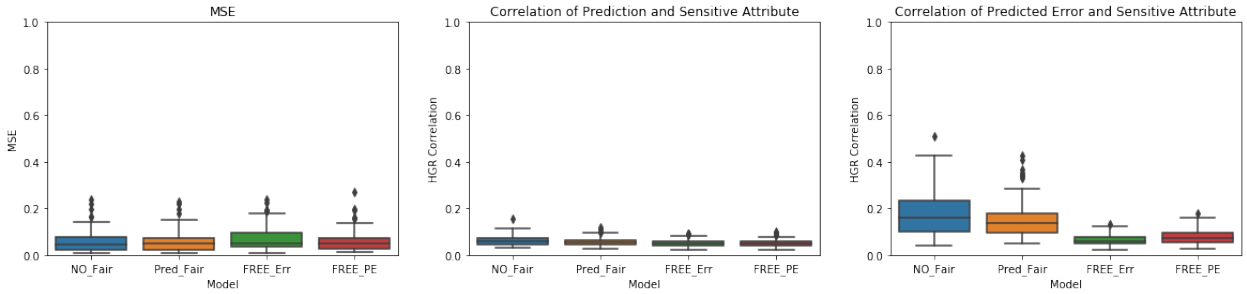

Figure 4.11: Model Comparison (MSE & HGR) for Simulated Data 2-4 with Binary S

4.3 Real-world Data

In this section, we implement the proposed models on the Communities and Crime (C&C) dataset Dua and Graff, 2017 which is a commonly used benchmark for evaluating regression models.

4.3.1 Dataset and Settings

The C&C dataset contains the socio-economic data and crime data on communities in the United States. In this dataset, each data point represents a community. The task is to predict the violent crime rate of community and the sensitive attribute is the ratio of African American Residents. Both the target violent crime rate and the sensitive attribute the ratio of African American people are continuous. To test the model performance for binary sensitive attribute, we binarized the original ratio of African American Residents into (1: High_AA) if it is $\geq 50\%$ and (0: Low_AA) otherwise.

4.3.2 Results and Analysis

C&C dataset with Continuous Sensitive Attribute

When the sensitive attribute is continuous, the results in Table 4.3 and Figure 4.12 show that: (1) All models obtain low MSE values, and the models considering fairness do not harm the model performance too much; (2) NO_Fair models cause unfairness in prediction and error, Pred_Fair models cause unfairness in error, and FREE_Err models cause unfairness in prediction, while FREE_PE models reduce the values of $\text{HGR}(\hat{y}, S)$ and values of $\text{HGR}(\hat{\epsilon}, S)$ below 0.3 which means the correlation coefficient is negligible.(3) For the correlation between Prediction and sensitive attribute, our approaches obtain comparable results than baselines; (4) For the correlation between prediction error and sensitive feature, our approaches could reduce the correlation to non-significant level with HGR values around 0.2.

Table 4.2: Simulation Results ($mean \pm std$) for Binary Sensitive Attributes.

Scenario	Model	MSE	HGR($\hat{y}, S$)	HGR($\hat{\epsilon}, S$)	OL($\hat{\epsilon}$)
Data2-1:	NO_Fair	$.05 \pm .03$	$.56 \pm .04$	$.73 \pm .15$	$.27 \pm .17$
$y \not\perp S$	Pred_Fair	$.06 \pm .04$	$.53 \pm .04$	$.80 \pm .13$	$.19 \pm .15$
$\epsilon \perp S$	FREE_Err	$.07 \pm .04$	$.68 \pm .04$	**.10** $\pm .03$	**.90** $\pm .03$
	FREE_PE	$.07 \pm .04$	$.67 \pm .04$	$.15 \pm .07$	$.86 \pm .07$
Data2-2:	NO_Fair	$.06 \pm .04$	$.88 \pm .02$	$.68 \pm .15$	$.30 \pm .17$
$y \not\perp S$	Pred_Fair	$.06 \pm .05$	$.87 \pm .02$	$.66 \pm .17$	$.33 \pm .20$
$\epsilon \not\perp S$	FREE_Err	$.07 \pm .03$	$.87 \pm .04$	**.17** $\pm .05$	**.85** $\pm .05$
	FREE_PE	$.06 \pm .04$	$.88 \pm .03$	$.29 \pm .19$	$.73 \pm .20$
Data2-3:	NO_Fair	$.05 \pm .03$	$.05 \pm .02$	$.18 \pm .10$	$.82 \pm .09$
$y \perp S$	Pred_Fair	$.05 \pm .04$	$.05 \pm .01$	$.17 \pm .08$	$.83 \pm .08$
$\epsilon \perp S$	FREE_Err	$.07 \pm .04$	$.04 \pm .01$	**.07** $\pm .02$	**.93** $\pm .03$
	FREE_PE	$.06 \pm .05$	$.04 \pm .01$	$.08 \pm .03$	$.91 \pm .04$
Data2-4:	NO_Fair	$.06 \pm .05$	$.06 \pm .02$	$.18 \pm .09$	$.83 \pm .09$
$y \perp S$	Pred_Fair	$.06 \pm .05$	$.06 \pm .02$	$.15 \pm .08$	$.85 \pm .08$
$\epsilon \not\perp S$	FREE_Err	$.07 \pm .05$	$.05 \pm .01$	**.07** $\pm .02$	**.93** $\pm .03$
	FREE_PE	$.06 \pm .05$	$.05 \pm .01$	$.08 \pm .03$	$.92 \pm .03$

NO_Fair: models without considering any fairness;
Pred_Fair[Mary etal., 2019]: models only considering fairness in prediction;
FREE_Err: FREE models only considering fairness in prediction error;
FREE_PE: FREE models considering both fairness in prediction and fairness in prediction error.

C&C dataset with Binary Sensitive Attribute

When the sensitive attribute is binary, Table 4.4, Figure 4.13 and Figure 4.15 show that: (1) There is a trade-off between prediction fairness and prediction error fairness; (2) For the correlation between prediction error and sensitive feature, the proposed FREE approaches produce much lower HGR correlation coefficients than two baselines NO_Fair and Pred_Fair. And Pred_Fair model increases the correlation between prediction error and sensitive variable, which is unexpected; (3) FREE_PE model can reduce the dependence both in prediction and prediction error, which provides more unbiased results; (4) For the overlapping index of the distributions of prediction error for subgroups High_AA and Low_AA, the proposed FREE approaches generate bigger overlapping area than baselines, which means our FREE regression methods obtain similar prediction errors across sensitive groups.

4.4 Discussions

We summarize our observations in experiments as follows: 1) Across the simulated datasets and the real-world C&C dataset, there is a trade-off among the predictive power (by MSE), fairness in prediction,

Table 4.3: Results on C&C dataset with Continuous Sensitive Attribute (mean±std of MSE and HGR). In FREE_PE model, we set $\lambda = 1.0$ and $\mu = 0.4$.

Model	MSE	HGR$(\hat{y}, S)$	HGR$(\hat{\epsilon}, S)$
NO_Fair	$.11 \pm .05$	$.27 \pm .09$	$.29 \pm .08$
Pred_Fair	$.15 \pm .05$	$\mathbf{.12} \pm .03$	$.36 \pm .05$
FREE_Err	$.15 \pm .06$	$.31 \pm .07$	$\mathbf{.15} \pm .05$
FREE_PE	$.14 \pm .05$	$.24 \pm .07$	$\mathbf{.23} \pm .08$

NO_Fair: models without considering any fairness;
Pred_Fair[Mary etal., 2019]: models only considering fairness in prediction;
FREE_Err: FREE models only considering fairness in prediction error;
FREE_PE: FREE models considering both fairness in prediction and fairness in prediction error. In FREE_PE Model, we set $\lambda = 1.0$ and $\mu = 0.4$.

Table 4.4: Results on C&C dataset with Binary Sensitive Attribute (mean±std of MSE, HGR and Overlapping Index (OL)). In FREE_PE model, we set $\lambda = 1.0$ and $\mu = 0.4$.

Model	MSE	HGR$(\hat{y}, S)$	HGR$(\hat{\epsilon}, S)$	OL$(\hat{\epsilon}, S)$
NO_Fair	$.12 \pm .05$	$.11 \pm .06$	$.12 \pm .04$	$.57 \pm .09$
Pred_Fair	$.15 \pm .04$	$\mathbf{.04} \pm .02$	$.14 \pm .03$	$.48 \pm .07$
FREE_Err	$.15 \pm .06$	$.13 \pm .05$	$\mathbf{.06} \pm .03$	$\mathbf{.72} \pm .09$
FREE_PE	$.14 \pm .05$	$.09 \pm .03$	$\mathbf{.10} \pm .04$	$\mathbf{.62} \pm .10$

NO_Fair: models without considering any fairness;
Pred_Fair[Mary etal., 2019]: models only considering fairness in prediction;
FREE_Err: FREE models only considering fairness in prediction error;
FREE_PE: FREE models considering both fairness in prediction and fairness in prediction error. In FREE_PE Model, we set $\lambda = 1.0$ and $\mu = 0.4$.

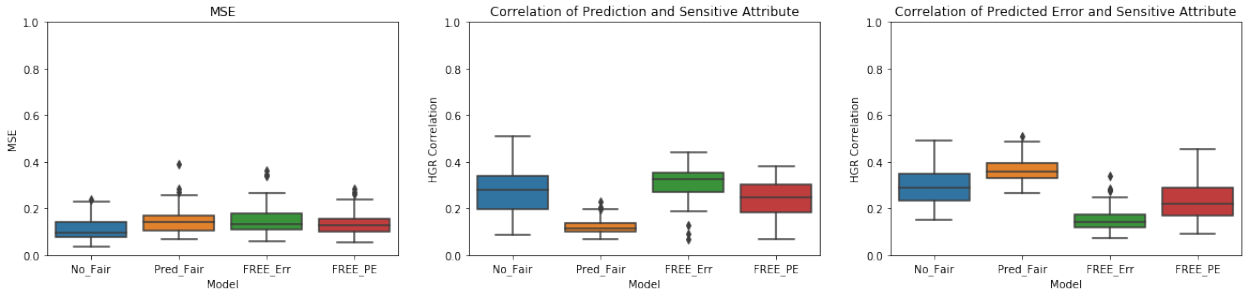

Figure 4.12: Model Comparison (HGR) for C&C Dataset with Continuous Sensitive Attribute

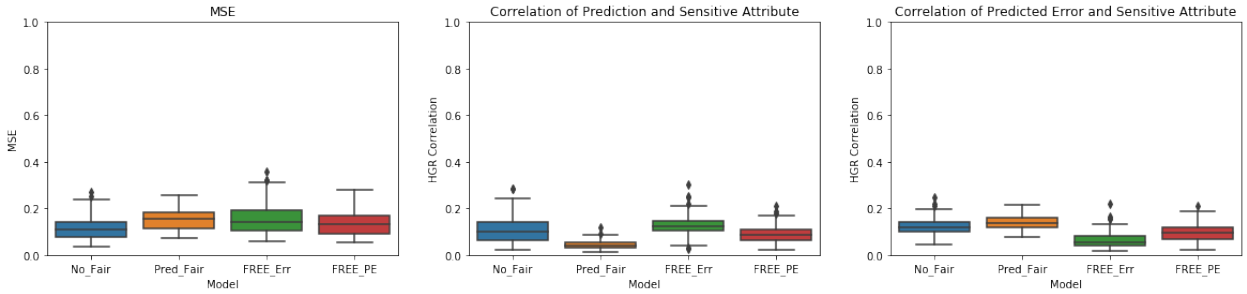

Figure 4.13: Model Comparison (HGR) for C&C Dataset with Binary Sensitive Attribute

and fairness in prediction error; 2) Archiving fairness on error is easier and costs less accuracy than on prediction especially when the response y is highly correlated with sensitive attribute S. 3) The proposed FREE_PE approach which constrain fairness in a more general framework than baselines can handle the trade-off between fairness in prediction and prediction error; 4) Results on both simulated and real-world datasets show that the proposed approaches could reduce more dependence between prediction error and sensitive attribute than baselines; 5) By changing the hyper-parameters of fairness penalty terms, the proposed FREE_PE approach can reduce both the unfairness in prediction and unfairness in prediction error to non-significant level.

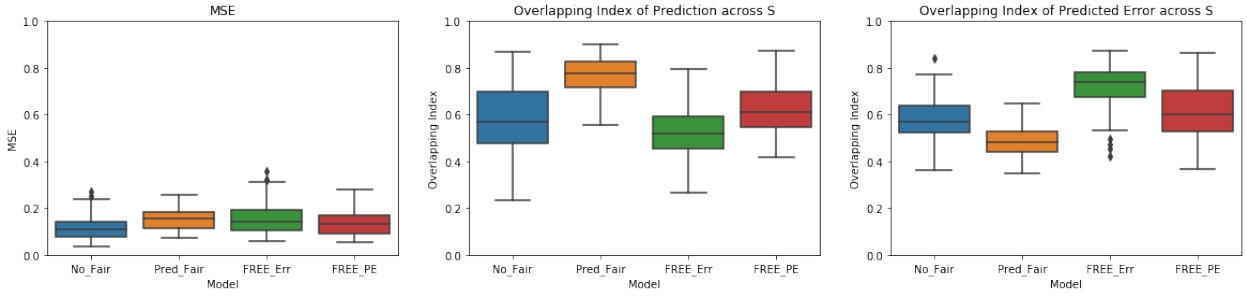

Figure 4.14: Model Comparison (Overlapping Index) for C&C Dataset with Binary Sensitive Attribute

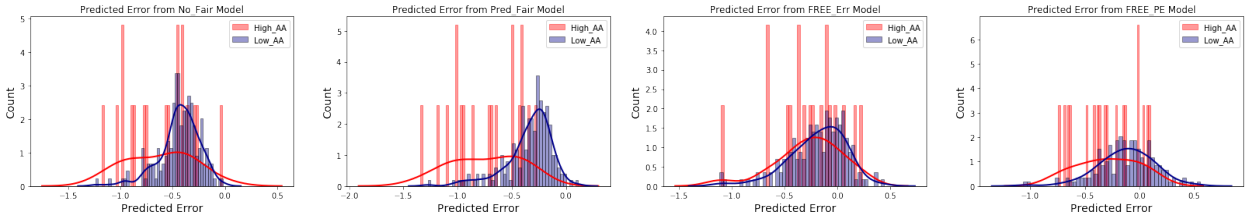

Figure 4.15: Model Comparison (Prediction Error) for C&C Dataset with Binary Sensitive Attribute

Chapter 5

Conclusion

In this paper, we characterize the unfairness in regression problems by defining a new fairness measurement based on equalized error. Furthermore, we propose a regularization approach named fairness regularization with equalized error (FREE) and design two fairness penalty terms for fair regression. Controlling fairness penalty terms, the proposed approach can reduce both the unfairness in prediction and unfairness in prediction error to non-significant level. Extensive experimental results on both simulated and real-world datasets show that the proposed approaches could effectively improve the fairness in prediction error and they can balance the fairness in prediction and fairness in prediction error. And we show that penalizing on fairness in prediction error would cost less accuracy than penalizing on fairness in prediction.

Appendix A

Experiments

A.1 Additional Results on Simulated Data with Binary Sensitive Attributes

Figures show the detailed model comparisons on the simulated Data 2-1, 2-2, 2-3 and 2-4 in terms of MSE, HGR, Overlapping Index, and Prediction Error.

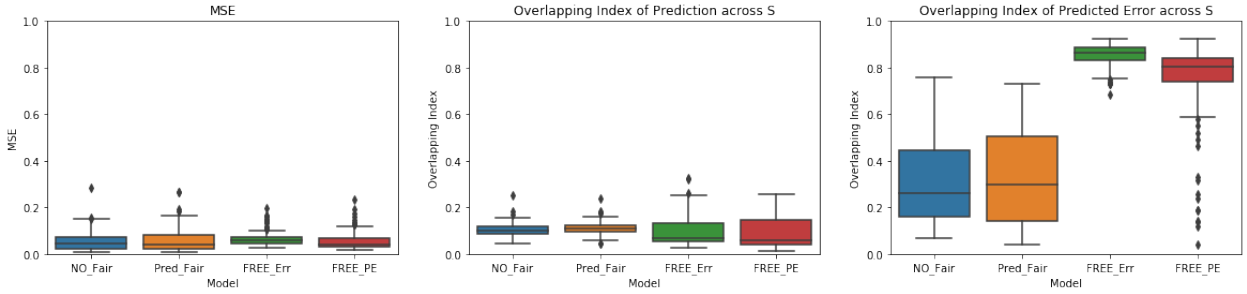

Figure A.1: Model Comparison (MSE & Overlapping Index) for Simulated Data 2-2 with Binary S

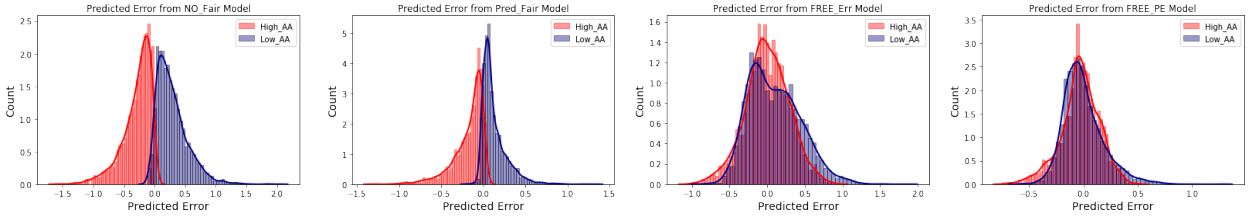

Figure A.2: Model Comparison (Prediction Error) for Simulated Data 2-2 with Binary S

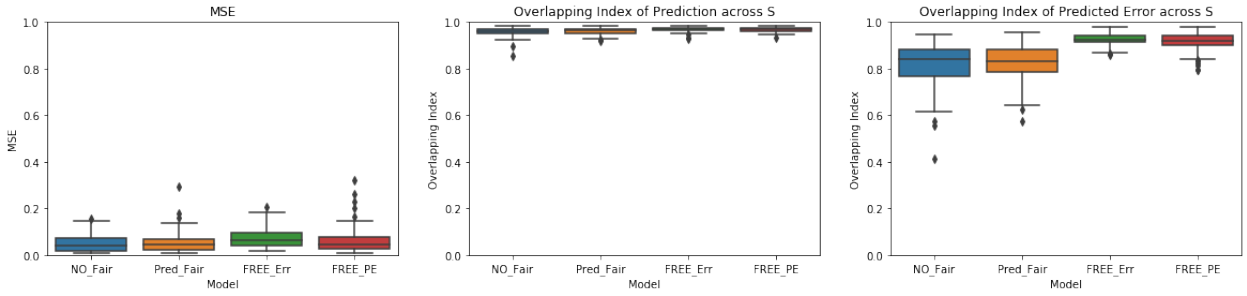

Figure A.3: Model Comparison (MSE & Overlapping Index) for Simulated Data 2-3 with Binary S

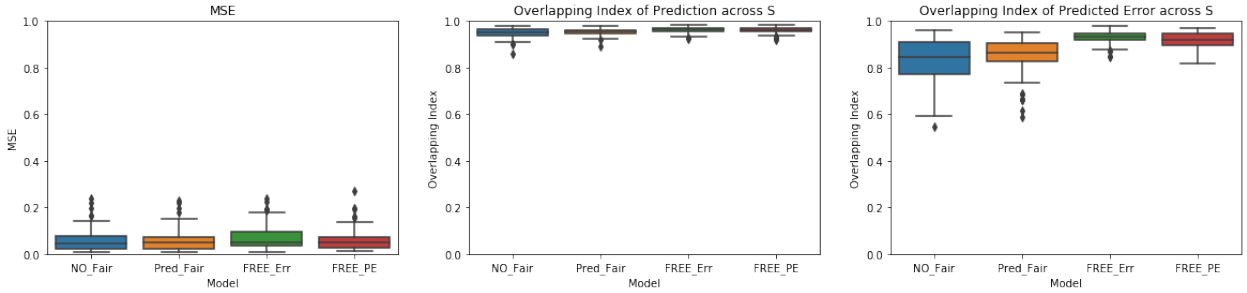

Figure A.4: Model Comparison (MSE & Overlapping Index) for Simulated Data 2-4 with Binary S

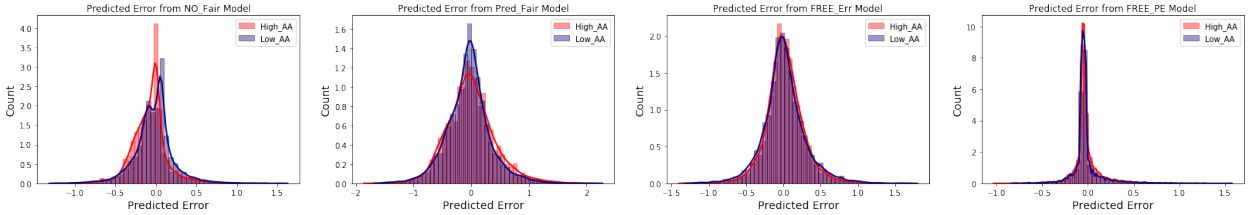

Figure A.5: Model Comparison (Prediction Error) for Simulated Data 2-4 with Binary S

Bibliography

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv:1905.12843*.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *NIPS Tutorial*, *1*.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. *2013 IEEE 13th International Conference on Data Mining*, 71–80. https://doi.org/10.1109/ICDM.2013.114

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized data pre-processing for discrimination prevention. *arXiv preprint arXiv:1704.03354*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression with wasserstein barycenters. *arXiv preprint arXiv:2006.07286*.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Dua, D., & Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Grari, V., Ruf, B., Lamprier, S., & Detyniecki, M. (2019). Fairness-aware neural r\'eyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*.

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, *25*(7), 1445–1459.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information*

processing systems (pp. 3315–3323). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *2010 IEEE International Conference on Data Mining*, 869–874. https://doi.org/10.1109/ICDM.2010.50

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650.

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Mary, J., Calauzénes, C., & El Karoui, N. (2019). Fairness-aware learning for continuous attributes and treatments. *International Conference on Machine Learning*, 4382–4391.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Narasimhan, H., Cotter, A., Gupta, M., & Wang, S. (2020). Pairwise fairness for ranking and regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5248–5255. https://doi.org/10.1609/aaai.v34i04.5970

Okray, A., Hu, H., & Lan, C. (2019). Fair kernel regression via fair feature embedding in kernel space. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 1417–1421.

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, *10*, 1089. https://doi.org/10.3389/fpsyg.2019.01089

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., & Camps-Valls, G. (2017). Fair kernel learning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 339–355.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, *10*(3-4), 441–451.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., & Caetano, T. (2020). Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*. https://doi.org/10.1145/3038912.3052660

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970.

Zhao, C., & Chen, F. (2019). Rank-based multi-task learning for fair regression. *2019 IEEE International Conference on Data Mining (ICDM)*, 916–925.