Fair Regression under Sample Selection Bias

Wei Du
University of Arkansas
Fayetteville, AR, USA
wd005@uark.edu

Xintao Wu
University of Arkansas
Fayetteville, AR, USA
xintaowu@uark.edu

Hanghang Tong
University of Illinois
Urbana, IL, USA
htong@illinois.edu

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10021107

***Abstract*—Recent research on fair regression focused on developing new fairness notions and approximation methods as target variables and even the sensitive attribute are continuous in the regression setting. However, all previous fair regression research assumed the training data and testing data are drawn from the same distributions. This assumption is often violated in real world due to the sample selection bias between the training and testing data. In this paper, we develop a framework for fair regression under sample selection bias when dependent variable values of a set of samples from the training data are missing as a result of another hidden process. Our framework adopts the classic Heckman model for bias correction and the Lagrange duality to achieve fairness in regression based on a variety of fairness notions. Heckman model describes the sample selection process and uses a derived variable called the Inverse Mills Ratio (IMR) to correct sample selection bias. We use fairness inequality and equality constraints to describe a variety of fairness notions and apply the Lagrange duality theory to transform the primal problem into the dual convex optimization. For the two popular fairness notions, mean difference and mean squared error difference, we derive explicit formulas without iterative optimization, and for Pearson correlation, we derive its conditions of achieving strong duality. We conduct experiments on three real-world datasets and the experimental results demonstrate the approach's effectiveness in terms of both utility and fairness metrics.**

***Index Terms*—sample selection bias, algorithmic fairness, regression analysis**

I. Introduction

Fairness has been an increasingly important topic in machine learning. Fair machine learning models aim to learn a function f for a target variable Y using input features X and a sensitive attribute A (e.g., gender), while ensuring the predicted value $\hat{Y}$ fair with respect to A based on some given fairness criterion. Fair machine learning models can be categorized into pre-processing (modifying training data or learning a new representation such that the information correlated to the sensitive attribute is removed), in-processing (adding fairness penalty to the objective function during training), and post-processing (applying perturbation or transformation to model output to reduce prediction unfairness). Much of existing works has focused on classification. In this paper, we focus on fair regression where the target Y is continuous.

Fair regression can be naturally defined as the task of minimizing the expected loss of real-valued predictions, subject to some fairness constraints. Fairness notions under the regression setting are in principle based on some forms of independence, e.g., the independence of model prediction $\hat{Y}$ and sensitive attribute A, the independence of prediction error $\hat{Y} - Y$ and sensitive attribute A, and the conditional independence of $\hat{Y}$ and A given Y. Different from the classification setting, variables of Y and $\hat{Y}$ (even A) become continuous in the regression setting, which requires new fairness notions and constrained optimization techniques. Researchers have developed quantitative metrics based on moment constraints, such as mean difference [1], mean squared error difference [2], and Pearson correlation [3]. These simplified metrics can be easily calculated but fail to capture subtle effects. For example, the predicted values may have different variances across groups. Recently, researchers started to propose fairness metrics based on distributions/densities instead of simple point estimate [4], and develop approximation methods [5] for achieving fairness in regression. It is imperative to develop a general fair regression framework that enforces a variety of fairness notions and provides efficient implementation and theoretical analysis when dual optimization and approximation are applied. Moreover, all previous fair regression research assumed the training data and testing data are drawn from the same distributions. This assumption is often violated in real world due to the sample selection bias between the training and testing data.

Figure 1 shows an illustrative example of studying the relationship between SAT scores (X) and potential college achievement (Y) of students. The regression model trained on only observed student samples who were already admitted to college (denoted as $\mathcal{D}_s$ and shown as solid data points in Figure 1(a)) would be biased as the fitted model did not consider applicants who could potentially go to college (denoted as $\mathcal{D}_u$ and shown as hollow points in Figure 1(a)). Note that for these applicants who did not go to college, SAT scores (X) are still available although their corresponding college achievements (Y) are missing. Moreover, as shown in Figure 1(b), the fair regression model trained on only admitted college students D_s in fact would be unfair and cannot be adopted for future applicants whose distribution is assumed to resemble the union of D_s and D_u. It is imperative to learn a fair regression model that can incorporate X values of samples from $\mathcal{D}_u$ to both improve model fitness and achieve fairness on population.

In this paper, we propose, *FairLR**, the fair regression framework under sample selection bias when dependent variable values of a set of samples from the training data are

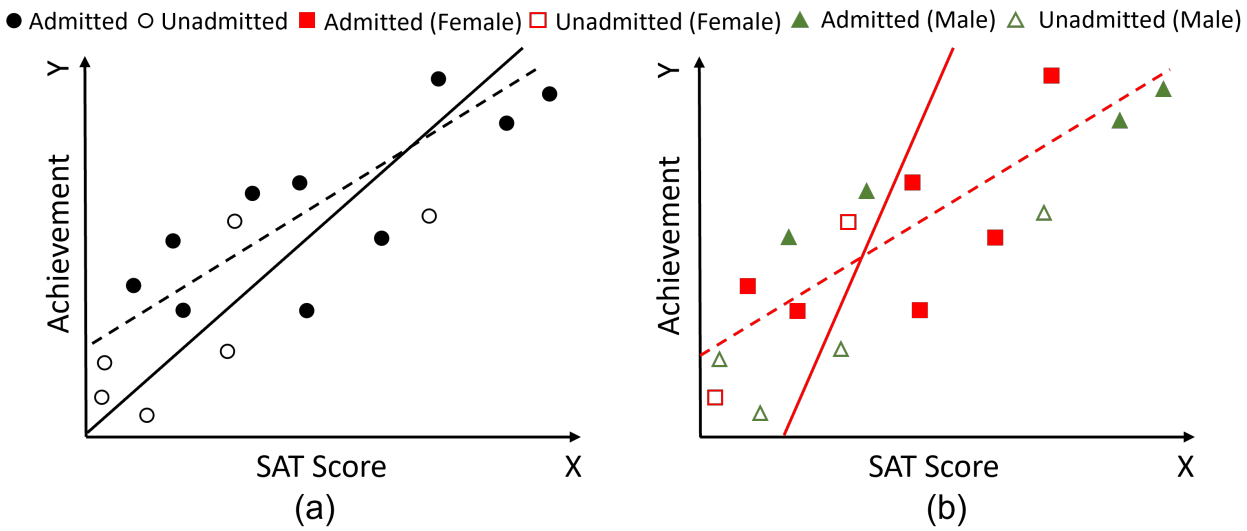

FIG. 1. Illustration for fair regression under sample selection bias. Fitted models with bias correction use feature values of unadmitted students. —— LR w/o correction, - - - LR with correction, —— fair LR w/o correction, - - - fair LR with correction

missing as a result of another hidden selection process. Our *FairLR** adopts the classic Heckman model [6] for bias correction and the Lagrange duality theory [7] to achieve regression fairness based on a variety of fairness notions. Our fair regression framework minimizes the loss function subject to fairness inequality and equality constraints. We apply the Lagrange duality theory to transform the primal problem into a dual convex optimization problem. For the two popular fairness notions, mean difference (MD) and mean squared error difference (MSED), we derive two explicit formulas without optimizing iteratively. For Pearson correlation, we derive its conditions of satisfying the Slater condition, thus achieving strong duality. We conduct experiments on three real-world datasets and the experimental results demonstrate our approach's effectiveness in terms of both utility and fairness.

II. Related Work

Fair Regression. For linear regression $f(\cdot) : X \to Y$ with discrete sensitive attribute A, [1] first introduced mean difference and AUC to measure the unfairness. [8] also used a similar concept termed as group fairness expectation to ensure fair prediction for different groups. For regression with discrete/continuous sensitive attribute, [9] used the Rényi maximum correlation coefficient of prediction and sensitive attribute to describe the fairness penalty. Recently, [4] presented two fairness definitions, statistical parity and bounded group loss. The statistical parity uses the departure of the cumulative distribution function (CDF) of $f(X)$ conditional on $A = a$ from the CDF of $f(X)$. When the departure is close to zero, the prediction is statistically independent of the protected attribute. The bounded group loss requires that the prediction error of any protected group stay below some predetermined thresholds.

To address the challenge of estimating information-theoretic divergences between conditional probability density functions, [5] introduced fast approximations of the independence, separation and sufficiency group fairness criteria for regression models from their (conditional) mutual information definitions. [10] focused on demographic parity that requires the distribution of the predicted output independent of the sensitive attribute. They established a connection between fair regression and optimal transport theory and derived a closed form expression for the optimal fair predictor, i.e., the distribution of this optimum is the Wasserstein barycenter of the distributions induced by the standard regression function on the sensitive groups.

Fair Classification under Sample Selection Bias. The sample selection bias causes the training data to be selected non-uniformly from the population to be modeled. Generally there are four types of sample selection bias, missing completely at random, missing at random, missing at random-class, and missing not at random when there is no independence assumption between features, target, and selection. Extensive research has been conducted on classification under sample selection bias (refer to a survey [11]). Some recent research focused on robust classification under sample selection bias and covariate shift. covariate shift is the most commonly studied scenario [12]–[16]. For example, [14], [17] considered covariate shift between the training and testing data and proposed a minimax robust framework that applies Gaussian kernel functions to reweigh the training examples. [16] adopted the reweighing estimation idea for sample selection bias correction and used the minimax robust estimation to achieve robustness on prediction accuracy. However, the reweighing approaches [18] usually assume that target distribution support implies source distribution support, which is not required in our proposed approach. To tackle the unfairness issue under distribution shift, [19]

developed a distributionally robust logistic regression model with an unfairness penalty. They assumed the unknown true test distribution is contained in a Wasserstein ball centered at the empirical distribution on the observed training data. [20] proposed the use of ambiguity set to derive the fair classifier based on the principles of distributional robustness. There are also other related works including fair transfer learning [21], fair federated learning [15] and fair classification with bias in label collection [22], [23].

III. Fair Regression under Sample Selection Bias

A. Problem Formulation

We first define the notations used in this paper. Let $(\boldsymbol{X}, A, Y)$ denote the training data $\mathcal{D}$, where $\boldsymbol{X}$ is the feature space, A is the protected attribute, and Y is the continuous target attribute. $\mathcal{D}$ contains n data samples, among which m samples $(\boldsymbol{x}_i, a_i, y_i)$ are fully observed and the remaining $n-m$ data points have y_i missing. We denote the fully observed part as $\mathcal{D}_s$ and the other part as $\mathcal{D}_u$. The whole training data $\mathcal{D}$ is selected uniformly from the population to be modeled. However, $\mathcal{D}_s$ is non-uniformly selected and the bias of $\mathcal{D}_s$ could depend on both feature vector $\boldsymbol{x}, a$ and target variable y. A regression function $f(\cdot) : X \to Y$ tries to learn optimal parameter $\boldsymbol{w}$. We denote $\hat{y} = f(\boldsymbol{x}; \boldsymbol{w})$.

Problem Statement. Given the training dataset $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_u$, derive regression function $f(\cdot) : X \to Y$ that achieves fairness on population with respect to protected attribute A based on some fairness criterion.

We emphasize the sample selection bias considered in our paper is different from the traditional covariate shift scenario. Although the covariate shift tackles the shift between the training distribution $P_{tr}(X)$ and test distribution $P_{te}(X)$, it usually assumes both a labeled training dataset and an unlabeled testing dataset are available in the training phase. In our work, we do not require the unlabeled testing data in the training phase. Instead, we assume the available training dataset contains a mixture of labeled and unlabeled data points but the labeling process is biased. In our setting, we are able to use the Heckman model to correct the bias with theoretical guarantee as we can compute the conditional unbiased expectation analytically. However, the previous commonly used approaches for covariate shift can only achieve robust estimation within a range but cannot provide theoretical guarantee.

B. Heckman Model Revisited

Heckman model [6] addresses the issue of sample selection bias when the dependent variable in the regression has values that are missing not at random. In the two-step estimation procedure of Heckman model, the first step uses probit regression to model the sample selection process and derives a new variable called the Inverse Mills Ratio (IMR). The second step adds the IMR to the regression analysis as an independent variable and uses ordinary least squares to estimate the regression coefficients. This two-step estimator can perform well when there is no multicollinearity between the IMR and the explanatory variables. We present below the Heckman model formally.

The selection equation of the ith sample is $z_i = \boldsymbol{x}_{1i}\boldsymbol{\gamma} + u_i$ where $\boldsymbol{x}_{1i}$ includes the set of features related to sample selection, $\boldsymbol{\gamma}$ is the set of regression coefficients, and u_i is the error term. The selection index s is defined as:

$$s_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i \leq 0 \end{cases} \tag{1}$$

where $s_i = 1$ indicates that the ith sample is fully observed and $s_i = 0$ indicates its target value y_i is missing. The prediction model is based on linear regression and for the ith sample we have $y_i = \hat{y}_i + \epsilon_i = \boldsymbol{x}_{2i}\boldsymbol{\beta} + \epsilon_i$ where $\hat{y}_i$ is the predicted value, $\boldsymbol{x}_{2i}$ includes the set of features used for prediction, $\boldsymbol{\beta}$ is the set of regression coefficients, and ϵ_i is the error term. Following the default assumptions in the Heckman model, $\boldsymbol{x}_{2i}$ is a subset of $\boldsymbol{x}_{1i}$, indicating that all attributes predicting the outcome of interest can also predict selection equation, and $u_i \sim N(0, 1)$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. Note that if $\boldsymbol{x}_{2i}$ is non-strict subset of $\boldsymbol{x}_{1i}$, a severe collinearity among regressors and large standards errors can be induced [24]. The correlation coefficient of u_i and ϵ_i is denoted by ρ. The prediction outcome based on $\mathcal{D}_s$ alone is biased and we can correct it by computing the conditional means of the prediction outcome as:

$$\begin{aligned} \mathbb{E}(y_i|s_i = 1) = \mathbb{E}(y_i|z_i > 0) &= \mathbb{E}(\boldsymbol{x}_{2i}\boldsymbol{\beta} + \epsilon_i|\boldsymbol{x}_{1i}\boldsymbol{\gamma} + u_i > 0) \\ &= \boldsymbol{x}_{2i}\boldsymbol{\beta} + \mathbb{E}(\epsilon_i|\boldsymbol{x}_{1i}\boldsymbol{\gamma} + u_i > 0) \\ &= \boldsymbol{x}_{2i}\boldsymbol{\beta} + \mathbb{E}(\epsilon_i|u_i > -\boldsymbol{x}_{1i}\boldsymbol{\gamma}) \end{aligned} \tag{2}$$

Because u_i and ϵ_i are correlated, then we have

$$\mathbb{E}(\epsilon_i|u_i > -\boldsymbol{x}_{1i}\boldsymbol{\gamma}) = \alpha_i \rho \sigma_\epsilon \tag{3}$$

where $\alpha_i = \dfrac{\phi(-\boldsymbol{x}_{1i}\boldsymbol{\gamma})}{1 - \Phi(-\boldsymbol{x}_{1i}\boldsymbol{\gamma})}$ is usually termed as IMR. Here $\phi(\cdot)$ denotes the standard normal density function and $\Phi(\cdot)$ denotes the standard cumulative distribution function.

To compute the value of α_i, the first step is to estimate the coefficients $\boldsymbol{\gamma}$. We use the maximum likelihood estimate (MLE) to estimate $\boldsymbol{\gamma}$ by treating the selection equation as a probit classification model and we have

$$P(s_i = 1) = \Phi(\boldsymbol{x}_{1i}\boldsymbol{\gamma}), P(s_i = 0) = 1 - \Phi(\boldsymbol{x}_{1i}\boldsymbol{\gamma}) \tag{4}$$

Then the likelihood of $\mathcal{D}$ is expressed as:

$$LH(\boldsymbol{\gamma}; s_i, \boldsymbol{x}_{1i}) = \prod_{i=1}^{n} \Phi(\boldsymbol{x}_{1i}\boldsymbol{\gamma})^{s_i} (1 - \Phi(\boldsymbol{x}_{1i}\boldsymbol{\gamma}))^{1-s_i} \tag{5}$$

The maximization of Eq. 5 will obtain the estimates of $\boldsymbol{\gamma}$, and thus we can compute α_i for each selected sample (x_i, a_i, y_i) in D_s. With available α_i, we can rewrite Eq. 2 as:

$$\mathbb{E}(y_i|s_i = 1) = \boldsymbol{x}_{2i}\boldsymbol{\beta} + \alpha_i \rho \sigma_\epsilon \tag{6}$$

Then we can estimate the coefficients $\boldsymbol{\beta}$ from Eq. 6, e.g., via the ordinary least squares (OLS) by minimizing $\min_\beta L(\boldsymbol{\beta}) = \sum_{i=1}^{m} (\boldsymbol{x}_{2i}\boldsymbol{\beta} + \alpha_i \rho \sigma_\epsilon - y_i)^2$.

C. Fair Regression via Heckman Correction

We first present the general fair regression framework that aims to minimize the risk and learns the parameters $\boldsymbol{w} \in \mathcal{W}$ subject to the fairness constraints:

$$\begin{aligned} &\min_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}[l(\hat{y}, y)] = \mathbb{E}[l(f(\boldsymbol{x}; \boldsymbol{w}), y)] \\ &\text{subject to} \quad g_i(\hat{y}, y, a) \leq 0, i = 1, \cdots, p \\ &\qquad\qquad\ \ h_j(\hat{y}, y, a) = 0, j = 1, \cdots, q \end{aligned} \tag{7}$$

where f is the learning model, l is the loss function, g_i are fairness inequality constraints, and h_j are fairness equality constraints.

We then formulate the fair regression under sample selection bias and rewrite Eq. 7 to minimize the empirical loss subject to the fairness constraint:

$$\begin{aligned} &p^* = \min_{\tilde{\beta}} L(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{m} l[(f_h(\tilde{\boldsymbol{x}}_{2i}; \tilde{\boldsymbol{\beta}}), y)] \\ &\text{subject to} \quad g_i(\tilde{\boldsymbol{\beta}}) \leq 0, i = 1, \cdots, p \\ &\qquad\qquad\ \ h_j(\tilde{\boldsymbol{\beta}}) = 0, j = 1, \cdots, q \end{aligned} \tag{8}$$

where $\tilde{\boldsymbol{\beta}} = [\boldsymbol{\beta}, \boldsymbol{\beta}_\alpha]$, $\beta_\alpha = \rho\sigma_\epsilon$, $\tilde{\boldsymbol{x}}_{2i} = [\boldsymbol{x}_{2i}, \alpha_i]$, and $f_h(\tilde{\boldsymbol{x}}_{2i}; \tilde{\boldsymbol{\beta}}) = \tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}}$ is the Heckman prediction function. Note that from Eq. 6 the bias is corrected by α_i which carries the information of $\mathcal{D}_u$. The effect of each α_i on the sample (x_i, a_i, y_i) is quantified by $\rho\sigma_\epsilon$. We can then treat $\rho\sigma_\epsilon$ as one additional dimension β_α of the coefficient vector $\tilde{\boldsymbol{\beta}}$. In addition, the fairness constraint computed based on the corrected $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{x}}_{2i}$ is unbiased.

1) Fairness Notions: Previous works on fair regression developed notions are based on the independence of model prediction (or prediction error) and sensitive attribute. Different from the classification setting, target variable Y is continuous and sensitive attribute A can be either categorical or continuous. Table I summarizes fairness notions including their formula, reference, and applicability in terms of sensitive attribute type. Refer to the appendix for their definitions.

The mean difference (MD), the mean squared error difference (MSED), the statistical parity (SP), and the bounded group loss (BGL) handle categorical sensitive attribute whereas Pearson correlation ($\rho_{\hat{y}a}$) and our introduced partial correlation ($\rho_{\hat{y}a.y}$) handles numerical sensitive attribute. The partial correlation $\rho_{\hat{y}a.y}$ includes both y and a in the condition which is similar to the equalized opportunity [30] in fair classification. MD, Pearson and SP focus on independence of model prediction and sensitive attribute whereas MSED, Partial and BGL consider prediction error. Moreover, SP (BGL) measures the dependence of prediction (prediction error) and sensitive attribute on distributions/densities, in contrary to point estimate of other notions.

In general, we can enforce strict fairness via equality constraints and relaxed fairness via inequality constraints in Eq. 20. For example, we use $h_j(\tilde{\boldsymbol{\beta}}) = 0$ for MD $= 0$, and use $g_i(\tilde{\boldsymbol{\beta}}) - \tau \leq 0$ for MD $\leq \tau$ where τ is a user-specified threshold. One challenge is for SP as the number of constraints is uncountable. We can apply the algorithm developed in [4] that discretizes the real-valued prediction space and reduces the optimization problem to cost-sensitive classification. The cost-sensitive classification is then solved by the reduction approach [31]. We note that our framework can be used to enforce multiple fairness notions at the same time and some notions may be mutually contradictory [32], which can cause vacuous solutions.

2) Dual Formulation: To solve the primal optimal problem (Eq. 20) with a variety of fairness notions, we apply the Lagrange duality theory [7] to relax the primal problem by its constraints. The Lagrangian function is

$$L_c(\tilde{\beta}, \lambda, \upsilon) = L(\tilde{\beta}) + \lambda^T g(\tilde{\beta}) + \upsilon^T h(\tilde{\beta}) \tag{9}$$

where $\lambda \in \mathbb{R}_+^p$ and $\upsilon \in \mathbb{R}^q$ are the Lagrange multiplier vectors (or dual variables) associated with inequality constraints and equality constraints. The dual function hence is defined as $Q(\lambda, \upsilon) = \inf_{\tilde{\beta}} L_c(\tilde{\beta}, \lambda, \upsilon)$. Note that the dual function $Q(\lambda, \upsilon)$ is a pointwise affine function of (λ, υ), it is concave even when the problem (Eq. 20) is non-convex. For each pair (λ, υ), the dual function gives us a lower bound of the optimal value p^*, i.e., $Q(\lambda, \upsilon) \leq p^*$. The best lower bound leads to the Lagrange dual problem:

$$d^* = \max_{\lambda\succeq 0, \upsilon} Q(\lambda, \upsilon) = \max_{\lambda\succeq 0, \upsilon} \min_{\tilde{\beta}} L_c(\tilde{\beta}, \lambda, \upsilon) \tag{10}$$

The Lagrange dual problem is a convex optimization problem because the objective to be maximized is concave and the constraint is convex. We can solve the dual optimization problem by alternating gradient descent steps over the primal variables $\tilde{\beta}$ and dual variables (λ, υ), respectively. In particular, by iteratively executing the following two steps: 1) find $\tilde{\beta}^* \leftarrow argmin_{\tilde{\beta}} L_c(\tilde{\beta}, \lambda, \upsilon)$; 2) compute $\lambda \leftarrow \lambda + \eta \frac{dL_c}{d\lambda}(\tilde{\beta}^*, \lambda, \upsilon)$, $\upsilon \leftarrow \upsilon + \eta \frac{dL_c}{d\upsilon}(\tilde{\beta}^*, \lambda, \upsilon)$, the solution will converge.

Next, we show instantiations of two widely used fairness notions, MD and MSED, by deriving their explicit formulas without iterative optimization. We leave the detailed proofs in the appendix.

Result 1. *For fair regression with the mean squared loss and MD$(\hat{y}, a) = 0$, we have the closed solution*

$$\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}_2^T \tilde{\boldsymbol{X}}_2)^{-1}(\tilde{\boldsymbol{X}}_2^T \boldsymbol{y} - \frac{\boldsymbol{d}^T(\tilde{\boldsymbol{X}}_2^T \tilde{\boldsymbol{X}}_2)^{-1}\tilde{\boldsymbol{X}}_2^T \boldsymbol{y}}{\boldsymbol{d}^T(\tilde{\boldsymbol{X}}_2^T \tilde{\boldsymbol{X}}_2)^{-1}\boldsymbol{d}}\boldsymbol{d}) \tag{11}$$

Proof Sketch. The dual optimization form is:

$$L(\tilde{\boldsymbol{\beta}}) = \min \sum_{i=1}^{m} (\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_{2i} - y_i)^2 + 2\lambda \boldsymbol{d}^T \tilde{\boldsymbol{\beta}} \tag{12}$$

where $\boldsymbol{d} = \frac{1}{m_0}\sum_{i\in\mathcal{D}_0} \tilde{\boldsymbol{x}}_{2i} - \frac{1}{m_1}\sum_{i\in\mathcal{D}_1} \tilde{\boldsymbol{x}}_{2i}$, m_0 (m_1) is the number of data in $\mathcal{D}_s$ with $a = 0$ (1). By setting the derivative of $L(\tilde{\boldsymbol{\beta}})$ with respect to $\tilde{\boldsymbol{\beta}}$ be zero, we get $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}_2^T \tilde{\boldsymbol{X}}_2)^{-1}(\tilde{\boldsymbol{X}}_2^T \boldsymbol{y} - \lambda \boldsymbol{d})$, where $\tilde{\boldsymbol{X}}_2$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m]$ and $\boldsymbol{y}$ is the vector form of $y_i, i \in [m]$. Using the fairness constraint, we get the closed solution of λ and then $\tilde{\boldsymbol{\beta}}$ as Eq. 11. □

TABLE I. Fairness Notions for Regression

Definition	Reference	Equation	Categorical	Numeric
MD	[1], [2], [25], [26]	$\text{MD}(\hat{y}, a) = \mathbb{E}(\hat{y}\|a=0) - \mathbb{E}(\hat{y}\|a=1)$	✓	×
MSED	[1], [2]	$\text{MSED}(\hat{y}, a) = \mathbb{E}[(y-\hat{y})^2\|a=0] - \mathbb{E}[(y-\hat{y})^2\|a=1]$	✓	×
Pearson	[3]	$\rho_{\hat{y}a} = \frac{\mathbb{E}[(\hat{y}-\mu_{\hat{y}})(a-\mu_a)]}{\sigma_{\hat{y}}\sigma_s}$	×	✓
Partial	ours	$\rho_{\hat{y}a.y} = \frac{\rho_{\hat{y}a}-\rho_{\hat{y}y}\rho_{ay}}{\sqrt{1-\rho_{\hat{y}y}^2}\sqrt{1-\rho_{ay}^2}}$	×	✓
SP	[4], [27]–[29]	$\text{SP} = \mathbb{P}[f(X) \geq z\|A=a] - \mathbb{P}[f(X) \geq z]$	✓	×
BGL	[4]	$\text{BGL} = \mathbb{E}[l(f(X), Y)\|A=a]$	✓	×

Result 2. *For fair regression with the mean squared loss and MSED$(\hat{y}, a) = 0$, we have the closed solution*

$$\begin{aligned}\tilde{\boldsymbol{\beta}} =&(\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2 + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\tilde{\boldsymbol{X}}_2^0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\tilde{\boldsymbol{X}}_2^1)^{-1}\\ &(\tilde{\boldsymbol{X}}_2^T\boldsymbol{y} + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\boldsymbol{y}_0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\boldsymbol{y}_1)\end{aligned} \tag{13}$$

Proof Sketch. Similar to Result 1, we write its Lagrange dual form of the MSED fairness constraint. Then we set the derivative of $\tilde{\boldsymbol{\beta}}$ to be zero, and compute the solution of λ and $\tilde{\boldsymbol{\beta}}$. Note that $\tilde{\boldsymbol{X}}_2^0$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m_0]$, $\tilde{\boldsymbol{X}}_2^1$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m_1]$, $\boldsymbol{y}_0$ is the vector form of $y_i, i \in [m_0]$, and $\boldsymbol{y}_1$ is the vector form of $y_i, i \in [m_1]$. □

3) Duality Gap Analysis: The optimal value d^* of the Lagrange dual problem, by definition, is the best lower bound on p^* that can be obtained from the Lagrange dual function. The difference $p^* - d^*$, which is always nonnegative, is the optimal duality gap of the original problem. One theoretical question is whether and under what conditions we can achieve zero duality gap (i.e., the optimal values of the primal and dual problems are equal) in our fair regression framework.

Result 3. *For fair regression with the convex loss function and the fairness inequality constraints (i.e., less than a user-specified threshold τ), the strong duality holds for Pearson correlation if the linear relationship exists between x and a.*

Proof. Our proof is based on strong duality via Slater condition. The Slater condition states that if a convex optimization problem has a feasible point $\tilde{\boldsymbol{\beta}}_0$ in the relative interior of the problem domain and every inequality constraint $g_i(\tilde{\boldsymbol{\beta}}) \leq 0$ is strict at $\tilde{\boldsymbol{\beta}}_0$, i.e., $g_i(\tilde{\boldsymbol{\beta}}_0) < 0$, then strong duality holds.

The correlation usually exists between $\tilde{\boldsymbol{x}}_2$ and a, and then the prediction based on $\tilde{\boldsymbol{x}}_2$ has disparate impact. We can remove the correlation between $\tilde{\boldsymbol{x}}_2$ and a through the following regression:

$$\hat{\boldsymbol{B}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\tilde{\boldsymbol{X}}_2, \boldsymbol{U} = \tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{B}}\boldsymbol{A} \tag{14}$$

where $\boldsymbol{A} = (a_1, a_2, \cdots, a_n)$ and we define $\boldsymbol{u}_i$ as the i-th datapoint of $\boldsymbol{U}$. With the assumption that the linear relationship exists between $\tilde{\boldsymbol{x}}_2$ and a, it was proved by [3] that ($\boldsymbol{u}$, a) has the same information with ($\tilde{\boldsymbol{x}}_2$, a) and the correlation between $\boldsymbol{u}$ and a is $O(\frac{1}{\sqrt{n}})$.

Suppose the prediction outcome $\hat{y}$ is expressed as the following:

$$\hat{y} = a\tilde{\boldsymbol{\beta}}_a + \boldsymbol{u}\tilde{\boldsymbol{\beta}}_u \tag{15}$$

Then we can compute the Pearson coefficient as the following:

$$\rho(\hat{y}, a) = \frac{Cov(\hat{y}, a)}{\sqrt{Var(\hat{y})Var(a)}} \tag{16}$$

where $Cov(\hat{y}, a)$ is the correlation between $\hat{y}$ and a, $Var(\hat{y})$ is the variance of $\hat{y}$, and $Var(a)$ is the variance of a. $Cov(\hat{y}, a)$ is calculated as:

$$\begin{aligned}Cov(\hat{y}, a) &= Cov(a\tilde{\boldsymbol{\beta}}_a + \boldsymbol{u}\tilde{\boldsymbol{\beta}}_u, a)\\ &= Cov(a\tilde{\boldsymbol{\beta}}_a, a) + Cov(\boldsymbol{u}\tilde{\boldsymbol{\beta}}_u, a)\\ &= \tilde{\boldsymbol{\beta}}_a Var(a) + 0 = \tilde{\boldsymbol{\beta}}_a Var(a)\end{aligned} \tag{17}$$

The variance of $\hat{y}$ is computed as:

$$\begin{aligned}Var(\hat{y}) &= Var(a\tilde{\boldsymbol{\beta}}_a + \boldsymbol{u}\tilde{\boldsymbol{\beta}}_u) = Var(a\tilde{\boldsymbol{\beta}}_a) + Var(\boldsymbol{u}\tilde{\boldsymbol{\beta}}_u)\\ &= \tilde{\boldsymbol{\beta}}_a^2 Var(a) + \tilde{\boldsymbol{\beta}}_u^T\boldsymbol{V}_u\tilde{\boldsymbol{\beta}}_u\end{aligned} \tag{18}$$

where $\boldsymbol{V}_u$ is the covariances of $\boldsymbol{u}$. Thus Eq. 16 can be written as:

$$\begin{aligned}\rho(\hat{y}, a) &= \frac{Cov(\hat{y}, a)}{\sqrt{Var(\hat{y})Var(a)}}\\ &= \frac{\tilde{\boldsymbol{\beta}}_a Var(a)}{\sqrt{(\tilde{\boldsymbol{\beta}}_a^2 Var(a) + \tilde{\boldsymbol{\beta}}_u^T\boldsymbol{V}_u\tilde{\boldsymbol{\beta}}_u)Var(a)}}\\ &= \frac{\tilde{\boldsymbol{\beta}}_a\sqrt{Var(a)}}{\sqrt{\tilde{\boldsymbol{\beta}}_a^2 Var(a) + \tilde{\boldsymbol{\beta}}_u^T\boldsymbol{V}_u\tilde{\boldsymbol{\beta}}_u}}\end{aligned} \tag{19}$$

Up to now, we can write down the fairness regression subject to the fairness constraint of Pearson coefficient:

$$\begin{aligned}&\min_{\tilde{\boldsymbol{\beta}}} L(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{m} l[(f_h(\tilde{\boldsymbol{x}}_{2i}; \tilde{\boldsymbol{\beta}}), y)]\\ &\text{subject to } \rho^2(\hat{y}, a) \leq \epsilon\end{aligned} \tag{20}$$

where ϵ is the threshold of the fairness metric. The fairness constraint $\rho^2(\hat{y}, a) \leq \epsilon$ is equivalent to:

$$(1-\epsilon)\tilde{\boldsymbol{\beta}}_a^2 Var(a) - \epsilon\tilde{\boldsymbol{\beta}}_u^T\boldsymbol{V}_u\tilde{\boldsymbol{\beta}}_u \leq 0 \tag{21}$$

The Slater condition requires that $\{(\tilde{\boldsymbol{\beta}}_a, \tilde{\boldsymbol{\beta}}_u) : (1-\epsilon)\tilde{\boldsymbol{\beta}}_a^2 Var(a) - \epsilon\tilde{\boldsymbol{\beta}}_u^T \boldsymbol{V}_u \tilde{\boldsymbol{\beta}}_u < 0\} \neq \emptyset$. It can be easily verified that Slater condition holds. For example, we can set $\tilde{\boldsymbol{\beta}}_a$ to be zero. Since $\boldsymbol{V}_u$ is symmetry and we can apply diagonal decomposition for $\boldsymbol{V}_u$ and the eigenvalues of $\boldsymbol{V}_u$ cannot be all zero. Suppose the jth eigenvalue of $\boldsymbol{V}_u$ is non-zero, and we can set the corresponding jth component of $\tilde{\boldsymbol{\beta}}_u$ to be same sign with the jth eigenvalue, and set all other components of $\tilde{\boldsymbol{\beta}}_u$ to be zero, so that the the Slater condition holds. □

Remarks. Note that we do not need to conduct the duality gap analysis for the mean difference (MD) and the mean squared error difference (MSED) as Results 1 and 2 have already given the explicit formulas for the primal optimization. For Partial, SP, BGL and other potential fairness notions, we leave their analysis in our future work. Moreover, when there is no sample selection bias, our Results 1-3 naturally hold by removing the tilde from those tilde symbols (e.g., $\tilde{\boldsymbol{\beta}}$).

IV. Experiments

A. Experiment Setting

Datasets. We conduct our experiment on three real-world datasets that are widely used to evaluate fair machine learning models. For each dataset, we choose 70% of data as training data $\mathcal{D}$ and leave the rest as testing data. To create the sample selection bias, we follow the procedure in [33] by splitting $\mathcal{D}$ into $\mathcal{D}_s$ (samples with fully observed features X and target Y) and $\mathcal{D}_u$ (samples with missing Y) according to some specific features. We show the characteristics of three datasets including protected attribute A, target Y, sizes of $\mathcal{D}_s$, $\mathcal{D}_u$, and testing data in Table II and show the attribute lists used in selection/prediction in Table III.

CRIME dataset [34] was collected from the 1990 US Census and contains socio-economic data of 1994 communities. The task is to predict the crime rate of a given community based on its socio-economic information. We choose the African American Population Ratio (AAPR) as the sensitive attribute and label a community as protected if its AAPR is greater than 50% and non-protected otherwise. In total, we have 219 protected communities and 1775 non-protected communities. In our experiments, we remove attributes with missing values and standardize all attributes to have zero mean value and unit variance. We include samples to $\mathcal{D}_s$ if the ratio of people under the poverty level in a community is less than 0.05, and samples to $\mathcal{D}_u$ otherwise.

LAW dataset [35] was collected from the Law School Admissions Council's National Longitudinal Bar Passage Study and consists of personal records of law students who went on to take the bar exam, including LSAT score, age, race and so forth. The task is to predict the GPA of a student based on other attributes. We choose race as the sensitive attribute and treat black as protected. The dataset contains a total of 20649 records and we randomly select 2700 records, including 700 protected samples and 2000 non-protected samples. We include samples to $\mathcal{D}_s$ if the year of birth is after 1950, and samples to $\mathcal{D}_u$ otherwise.

COMPAS dataset [36] consists of a collection of data from criminal defenders from Florida in 2013-2014. Each data sample is associated with personal information, including race, gender, age, prior criminal history, and so forth. The task is to predict the risk level of a defender based on other attributes. We choose race as the sensitive attribute and treat black defenders as protected. After removing the duplicated data samples, we have a total of 4397 data samples, including 2694 protected samples and 1703 non-protected samples. We include samples to $\mathcal{D}_s$ if the year of decile score is less than 10, and to $\mathcal{D}_u$ otherwise.

Baseline Models and Metrics. We choose linear regression with the standard loss function, mean squared loss, in our proposed framework *FairLR**. We adopt each of four fairness metrics, MD, MSED, Pearson coefficient and Partial coefficient, with equality constraint forms. We consider the following baseline models: (a) Linear regression (*LR*) without fairness constraint; (b) Linear regression with Heckman correction (*Heckman*) from [6] (c) Linear regression with each fairness constraint (*FairLR*), including MD [1], MSED [2], Pearson coefficient [3], and Partial coefficient. We evaluate the performance of the proposed framework based on prediction accuracy and fairness. We use the mean squared error (MSE) to measure prediction accuracy. For fairness, we use MD and MSED in the binary sensitive attribute setting and Pearson coefficient and Partial coefficient in the numerical sensitive attribute setting. As the goal of fair regression is to achieve good accuracy and fairness on population, we use MSE and fairness calculated from testing data to compare different models. For a comprehensive comparison, we also report those values calculated from $\mathcal{D}_s$. Our experiments were carried out on the Dell PowerEdge C4130 with 2 Nvidia Tesla M10 GPU.

B. Evaluation on Binary Protected Attribute

We report in Figure 2 our main comparison results on three datasets. Y-axis is MSE to reflect prediction accuracy and X-axis is based on the fairness metric chosen in fair regression models (*FairLR* and *FairLR**). In particular, the three plots in the first row of Figure 2 report MD whereas those on the second row report MSED. In each plot, we have eight markers with different shape and color, each of which reflects the MSE and fairness metric for one of the four compared models on either $\mathcal{D}_s$ or testing data. Throughout this section, we use $\circ$, $\triangle$, $\diamond$, and $\star$ to denote *LR*, *Heckman*, *FairLR* and *FairLR**, and use hollow (solid) marker to represent results on $\mathcal{D}_s$ (testing data). In general, markers in bottom-left region (close to origin) indicate good performance of corresponding methods as we want to achieve both low MSE for prediction and low MD/MSED for fairness.

We focus on the main results of comparing four methods on testing data, reflected by four solid markers in each plot [1]. We clearly see that markers of *Heckman* always locate below that of *LR*, indicating *Heckman* successfully corrects the sample

[1]Due to space limit, we skip the comparison results of four methods on $\mathcal{D}_s$ reflected by four hollow markers in each plot.

TABLE II. Characteristics of datasets

Dataset	Protected A	Target Y	$\lvert\mathcal{D}_s\rvert$	$\lvert\mathcal{D}_u\rvert$	Testing
CRIME	AAPR	Crime Rate	976	419	599
LAW	Black/Non-black	GPA	1323	567	810
COMPAS	Black/Non-black	Risk Score	2153	924	1320

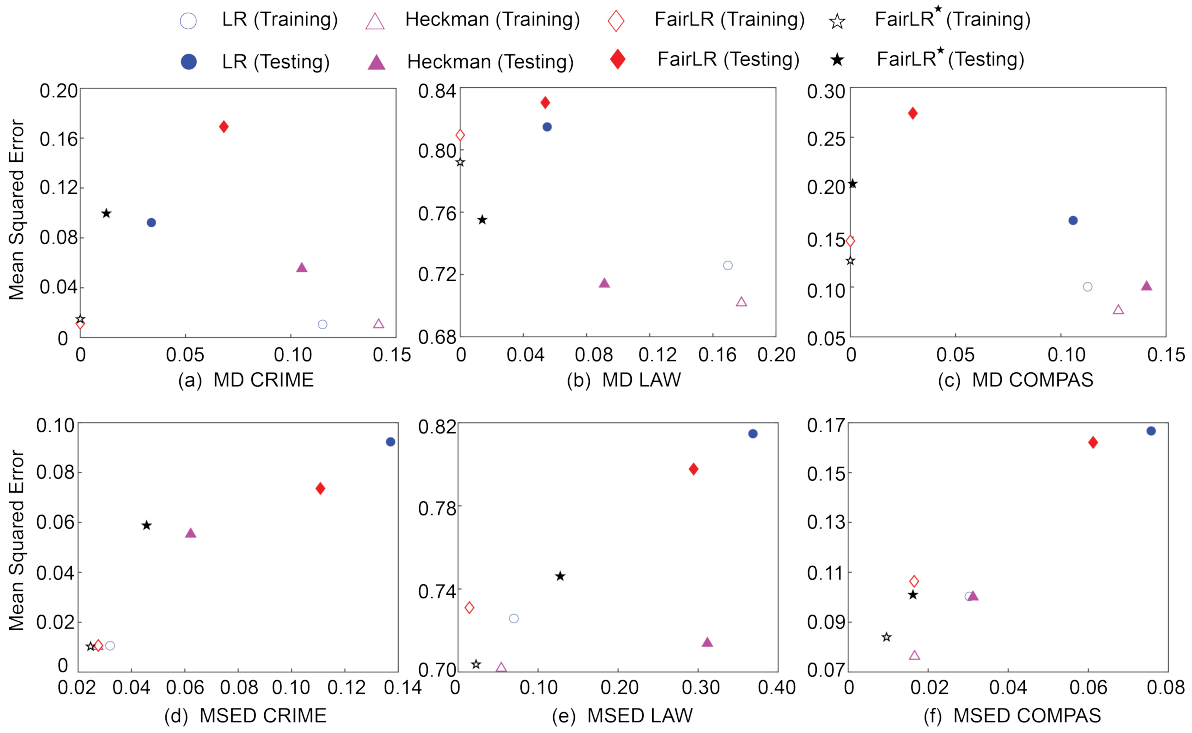

FIG. 2. Performance evaluation of binary protected attribute on CRIME, LAW, and COMPAS. The closer to the origin, the better the accuracy-fairness trade-off.

TABLE III. Attributes used for selection/prediction. Those with italic font are for prediction and those with either regular or italic font are for selection.

Dataset	Attribute
CRIME	population, householdsize, racepctblack, racePctWhite, racePctAsian, *racePctHisp*, *agePct12t21*, *agePct12t29*, *agePct16t24*, *agePct65up*, *numbUrban*, *pctUrban*, *medIncome*, *pctWWage*, *pctWFarmSelf*, *pctWInvInc*, *pctWSocSec*, *pctWPubAsst*, *pctWRetire*, *medFamInc*
LAW	cluster, lsat, ugpa, zgpa, *fulltime*, *fam_inc*, *age*, *gender*, *pass*
COMPAS	decile_score.1, age_cat_25-45, age_cat_45+, age_cat_25-, c_charge_degree_F, c_charge_degree_M, *sex*, *age*, *juv_fel_count*, *juv_misd_count*, *juv_other_count*, *priors_count*, *two_year_recid*

selection bias and reduces prediction error on testing data. Taking Figure 2 (a) as an example, the MSE of *Heckman* is 0.0553 whereas the MSE of *LR* is 0.0923. Similarly, as *FairLR* does not do bias correction, the solid marker of *FairLR* is also higher than that of *FairLR** for all three datasets. This demonstrates the effectiveness of Heckman model for correcting sample selection bias. Moreover, the solid marker of *FairLR* is always on the right side of *FairLR**, reflecting that *FairLR* simply trained on $\mathcal{D}_s$ without bias correction fails to achieve fairness on testing data. For example, in Figure 2 (f), the MSED of *FairLR* is 0.0612 whereas that of *FairLR** is 0.0162. To conclude, our proposed *FairLR** achieves the best trade-off between fairness and regression accuracy on the testing data.

It is also interesting to compare each model's performance between the training data and testing data. For our *FairLR**, we can see its hollow marker and solid marker are close to each other horizontally, indicating that the fairness achieved on $\mathcal{D}_s$ can also guarantee the testing fairness. However, the hollow marker and solid marker of *FairLR* are separate, indicating the sample selection bias can incur unfairness in the testing data although *FairLR* achieved training fairness.

C. Evaluation on Numerical Protected Attribute

We conduct experiments on CRIME by using the original numerical attribute AAPR as sensitive attribute. We use Pearson coefficient to measure the independence between $\hat{Y}$ and A, and use Partial coefficient to measure the conditional independence of $\hat{Y}$ and A given the true target value Y. Figure 3 shows the comparison results of four models. We have similar observations as the binary protected attribute setting. First, for results on $\mathcal{D}_s$ reflected by hollow markers,

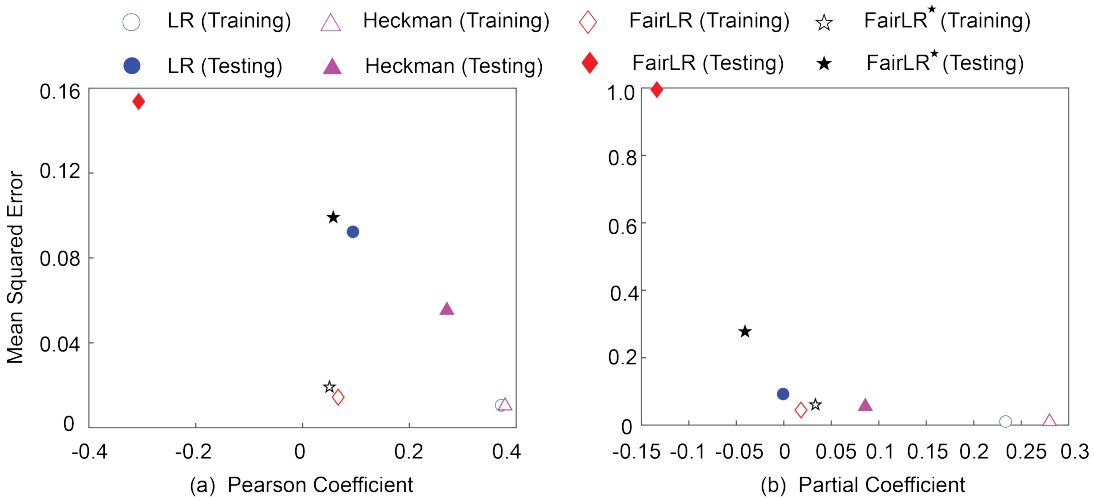

FIG. 3. Performance evaluation of numerical protected attribute on CRIME. The closer to the origin, the better the accuracy-fairness trade-off.

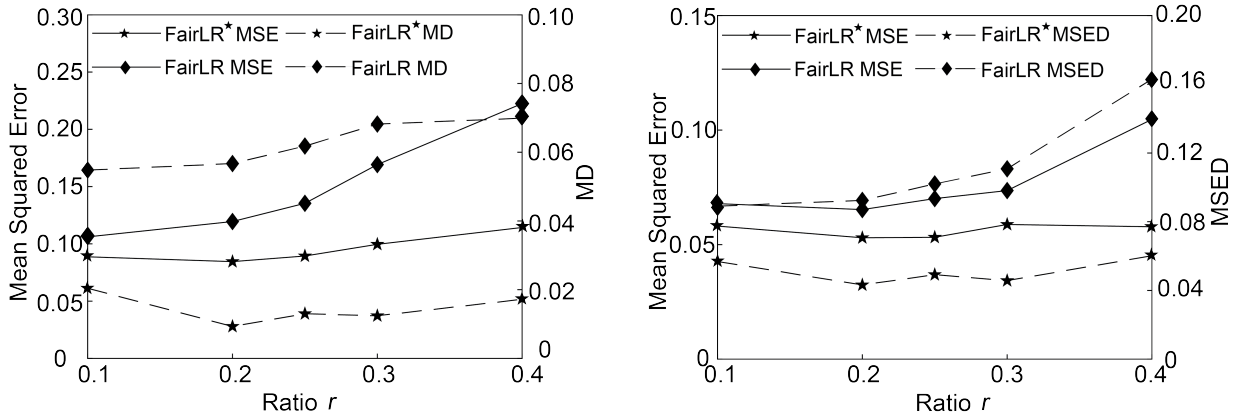

FIG. 4. Effects of Ratio $r = |\mathcal{D}_u|/|\mathcal{D}|$

all of the hollow markers are in bottom region with low MSE, and both *FairLR* and *FairLR** can achieve training fairness in terms of both Pearson coefficient and Partial coefficient. Second, for results on testing data reflected by solid markers, *Heckman* (*FairLR**) achieves lower MSE than *FairLR* as the former model considers the sample bias selection. We also see that *FairLR** is able to achieve testing fairness given the fairness threshold. However, the traditional *LR* achieves better testing accuracy-fairness trade-off than *FairLR** when Partial Coefficient fairness is enforced. This could be due to the potential loss during the iterative optimization.

D. Performance Evaluation on Biased Ratio

In this section, we evaluate how ratio $r = |\mathcal{D}_u|/|\mathcal{D}|$ would affect the performance of our *FairLR** and baseline *FairLR* on the testing data. Note that larger r indicates more bias in sample selection. We conduct experiments on CRIME. Figure 4 plots results of MD and MSED. In both plots, X-axis shows the varied r values from 0.1 to 0.4, the left Y-axis shows MSE, and the right Y-axis shows the fairness metric (MD or MSED). Correspondingly, we use solid lines to represent MSE values and dashed lines to represent fairness values. It is unsurprising to see that *FairLR** always achieves better performance (smaller MSE and smaller MD or MSED) than *FairLR*, as demonstrated in Figure 4 that lines with symbol $\star$ locate below those with symbol $\diamond$. More importantly, for our *FairLR**, the fairness value (MD or MSED) and prediction error (MSE) are stable when r increases, demonstrating the robustness of our *FairLR** against sample selection bias. On the contrary, for *FairLR*, both the unfairness and prediction error on the testing data increase when r increases.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have developed a framework for fair regression under sample selection bias when dependent variable values of a set of samples are missing. The framework adopts the classic Heckman model to correct sample selection bias and captures a variety of fairness notions via inequality and equality constraints. We applied the Lagrange duality theory to derive the dual convex optimization and showed the conditions

of achieving strong duality for Pearson correlation. For the two popular fairness notions, mean difference and mean squared error difference, we further derived explicit formulas without optimizing iteratively. Experimental results on three real-world datasets demonstrated our approach's effectiveness.

In our future work, we will conduct theoretical analysis and empirical evaluation of density based fairness notions, e.g., SP and BGL, and notions for multiple sensitive attributes. Some recent work [37] proposed to use Hirschfeld-Gebelein-Rényi Maximum (HGR) correlation coefficient as a regression fairness notion to evaluate the independence between prediction and sensitive attributes. However, it is quite challenging to compute HGR. We can only get analytical solution for some certain distributions, e.g., jointly Gaussian distribution [38], or apply approximation approaches. In our future work, we will study HGR in our framework. We will also study improved estimators [39] that address the limitations of Heckman estimator, e.g., sensitivity of estimated coefficients with respect to the distributional assumptions on the error terms, and extend to nonlinear cases, e.g., kernel regression, in our fair regression.

Reproducibility. All source code and datasets can be downloaded at https://tinyurl.com/2p9f36tb

ACKNOWLEDGMENT

This work was supported in part by NSF 1920920, 1939725, 1946391, 2137335 and 2147375.

REFERENCES

[1] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 71–80.

[2] K. D. Johnson, D. P. Foster, and R. A. Stine, "Impartial predictive modeling: Ensuring group fairness in arbitrary models," *arXiv e-prints*, pp. arXiv–1608, 2016.

[3] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2737–2746.

[4] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Proceedings of the 36th International Conference on Machine Learning ICML*, vol. 97. PMLR, 2019, pp. 120–129.

[5] D. Steinberg, A. Reid, S. O'Callaghan, F. Lattimore, L. McCalman, and T. S. Caetano, "Fast fair regression via efficient approximations of mutual information," *CoRR*, vol. abs/2002.06200, 2020.

[6] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica: Journal of the Econometric Society*, pp. 153–161, 1979.

[7] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[8] J. Fitzsimons, A. Al Ali, M. Osborne, and S. Roberts, "A general framework for fair regression," *Entropy*, vol. 21, no. 8, p. 741, 2019.

[9] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4382–4391.

[10] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression with wasserstein barycenters," *arXiv preprint arXiv:2006.07286*, 2020.

[11] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[12] W. Hu, G. Niu, I. Sato, and M. Sugiyama, "Does distributionally robust supervised learning give robust classifiers?" in *ICML*, 2018.

[13] A. Liu and B. Ziebart, "Robust classification under sample selection bias," in *NeurIPS*, 2014.

[14] J. Wen, C.-N. Yu, and R. Greiner, "Robust learning under uncertain test distributions: Relating covariate shift to model misspecification." in *ICML*, 2014.

[15] W. Du, D. Xu, X. Wu, and H. Tong, "Fairness-aware agnostic federated learning," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021, pp. 181–189.

[16] W. Du and X. Wu, "Fair and robust classification under sample selection bias," in *Proceedings of the 2021 ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2021.

[17] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample selection bias correction theory," in *International Conference on Algorithmic Learning Theory*. Springer, 2008, pp. 38–53.

[18] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, "Correcting sample selection bias by unlabeled data," *Advances in Neural Information Processing Systems*, vol. 19, 2006.

[19] B. Taskesen, V. A. Nguyen, D. Kuhn, and J. Blanchet, "A distributionally robust approach to fair classification," *arXiv preprint arXiv:2007.09530*, 2020.

[20] A. Rezaei, R. Fathony, O. Memarrast, and B. D. Ziebart, "Fairness for robust log loss classification," in *AAAI*, 2020.

[21] C. Schumann, X. Wang, A. Beutel, J. Chen, H. Qian, and E. H. Chi, "Transfer of machine learning fairness across domains," *arXiv preprint arXiv:1906.09688*, 2019.

[22] N. Kallus and A. Zhou, "Residual unfairness in fair machine learning from prejudiced data," in *ICML*, 2018.

[23] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *AISTATS*, 2020.

[24] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010.

[25] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," in *FAT ML*, 2018.

[26] C. Zhao and F. Chen, "Unfairness discovery and prevention for few-shot regression," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE, 2020, pp. 137–144.

[27] H. Narasimhan, A. Cotter, M. Gupta, and S. Wang, "Pairwise fairness for ranking and regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[28] T. Le Gouic and J.-M. Loubes, "Computing the price for fairness in a regression framework," *arXiv preprint arXiv:2005.11720*, 2020.

[29] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression via plug-in estimator and recalibration with statistical guarantees," in *Advances in Neural Information Processing Systems*, 2020.

[30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[31] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[32] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[33] P. Laforgue and S. Clémençon, "Statistical learning from biased training samples," *arXiv preprint arXiv:1906.12304*, 2019.

[34] http://archive.ics.uci.edu/ml/datasets/communities+and+crime, 2009.

[35] L. F. Wightman, "Lsac national longitudinal bar passage study. Lsac Research Report Series." 1998.

[36] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "Compas dataset," https://github.com/propublica/compas-analysis, 2017.

[37] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fairness-aware neural renyi minimization for continuous features," in *IJCAI*, 2020.

[38] S. Asoodeh, F. Alajaji, and T. Linder, "On maximal correlation, mutual information and data privacy," in *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*. IEEE, 2015, pp. 27–31.

[39] P. Puhani, "The heckman correction for sample selection and its critique," *Journal of Economic Surveys*, vol. 14, no. 1, pp. 53–68, 2000.

APPENDIX

A. Fairness Metric

Definition 1. *The mean difference (MD) of numeric prediction $\hat{y}$ in $\mathcal{D}$ by a binary protected attribute a is defined as* $MD(\hat{y}, a) = \mathbb{E}(\hat{y}|a = 0) - \mathbb{E}(\hat{y}|a = 1)$.

Definition 2. *The mean squared error difference (MSED) of numeric prediction $\hat{y}$ in $\mathcal{D}$ by a binary protected attribute a is defined as MSED*$(\hat{y}, a) = \mathbb{E}[(y-\hat{y})^2|a=0] - \mathbb{E}[(y-\hat{y})^2|a=1]$.

Definition 3. *The correlation coefficient of numeric prediction $\hat{y}$ and numeric protected attribute a is defined as* $\rho_{\hat{y}a} = \frac{\mathbb{E}[(\hat{y}-\mu_{\hat{y}})(a-\mu_a)]}{\sigma_{\hat{y}}\sigma_s}$.

Definition 4. *The partial correlation coefficient of numeric prediction $\hat{y}$ and numeric protected attribute a given y is defined as* $\rho_{\hat{y}a.y} = \frac{\rho_{\hat{y}a} - \rho_{\hat{y}y}\rho_{ay}}{\sqrt{1-\rho_{\hat{y}y}^2}\sqrt{1-\rho_{ay}^2}}$.

Definition 5. *The statistical parity (SP) is defined as SP* $= \mathbb{P}[f(X) \geq z|A=a] - \mathbb{P}[f(X) \geq z]$ *for all $a \in \mathcal{A}$ and $z \in [0,1]$.*

Definition 6. *The bounded group loss (BGL) is defined as BGL* $= \mathbb{E}[l(f(X), Y)|A=a]$ *for all $a \in \mathcal{A}$.*

[4] presented two fairness definitions, statistical parity and bounded group loss. The statistical parity uses the departure of the CDF of $f(X)$ conditional on $A=a$ from the CDF of $f(X)$. When the departure is close to zero, the prediction is statistically independent of the protected attribute. The bounded group loss which asks that the prediction error of any protected group stay below some pre-determined threshold.

B. Proof of RESULT 1

Proof. The fair Heckman prediction model can be described as:

$$\begin{aligned} &\min L(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{m} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 \\ &\text{subject to} \quad \frac{1}{m_0}\sum_{i\in\mathcal{D}_0} \tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} = \frac{1}{m_1}\sum_{i\in\mathcal{D}_1} \tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} \end{aligned} \tag{22}$$

where m_0 is the number of data in $\mathcal{D}_s$ with $a=0$, m_1 is the number of data with $a=1$, and $m = m_0 + m_1$.

We solve this optimization problem Eq. 22 using Lagrange multipliers. For convenience, $\frac{1}{m_0}\sum_{i\in\mathcal{D}_0} \tilde{\boldsymbol{x}}_{2i} - \frac{1}{m_1}\sum_{i\in\mathcal{D}_1} \tilde{\boldsymbol{x}}_{2i}$ is denoted as $\boldsymbol{d}$. Then we can rewrite Eq. 22 as the following constrained minimization problem:

$$L(\tilde{\beta}) = \min \sum_{i=1}^{m} (\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_{2i} - y_i)^2 + 2\lambda \boldsymbol{d}^T\tilde{\boldsymbol{\beta}} \tag{23}$$

where λ is the Lagrange multiplier.

By taking the partial derivatives of jth coefficient $\tilde{\beta}_j$ of $\tilde{\boldsymbol{\beta}}$:

$$\frac{\partial L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\beta}_j} = \sum_{i=1}^{m} 2(\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)\tilde{\boldsymbol{x}}_{2ij} + 2\lambda d_j \tag{24}$$

where $\tilde{\boldsymbol{x}}_{2ij}$ is the jth component of $\tilde{\boldsymbol{x}}_{2i}$ and d_j is the jth component of $\boldsymbol{d}$. By setting the derivative to be zero for all j, we can get:

$$(\sum_{i=1}^{m} \tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{x}}_{2ij})\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{m} y_i\tilde{\boldsymbol{x}}_{2ij} - \lambda d_j \tag{25}$$

Thus we can rewrite Eq. 25 with matrix form:

$$\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}_2^T\boldsymbol{y} - \lambda\boldsymbol{d} \tag{26}$$

where $\tilde{\boldsymbol{X}}_2$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m]$ and $\boldsymbol{y}$ is the vector form of $y_i, i \in [m]$. Therefore, we have:

$$\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2)^{-1}(\tilde{\boldsymbol{X}}_2^T\boldsymbol{y} - \lambda\boldsymbol{d}) \tag{27}$$

We can also get solution of λ using the fairness constraint $\boldsymbol{d}^T\tilde{\boldsymbol{\beta}} = 0$:

$$\lambda = \frac{\boldsymbol{d}^T(\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2)^{-1}\tilde{\boldsymbol{X}}_2^T\boldsymbol{y}}{\boldsymbol{d}^T(\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2)^{-1}\boldsymbol{d}} \tag{28}$$

By substituting λ into Eq. 27, we have the closed solution. □

C. Proof of RESULT 2

Proof. The fair Heckman prediction model can be described as:

$$\begin{aligned} &\min L(\boldsymbol{\beta}) = \sum_{i=1}^{m} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 \\ &\text{subject to} \quad \frac{1}{m_0}\sum_{i\in\mathcal{D}_0} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 = \frac{1}{m_1}\sum_{i\in\mathcal{D}_1} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 \end{aligned} \tag{29}$$

We use the same notations as above and apply the Lagrange multipliers:

$$\begin{aligned} L(\tilde{\boldsymbol{\beta}}) = \min &\sum_{i=1}^{m} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 + \\ &\lambda(\frac{1}{m_0}\sum_{i\in\mathcal{D}_0} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2 - \frac{1}{m_1}\sum_{i\in\mathcal{D}_1} (\tilde{\boldsymbol{x}}_{2i}\tilde{\boldsymbol{\beta}} - y_i)^2) \end{aligned} \tag{30}$$

We can compute the derivatives of $\tilde{\boldsymbol{\beta}}$ with the matrix form and set it to be zero:

$$\begin{aligned} 2\tilde{\boldsymbol{X}}_2^T(\tilde{\boldsymbol{X}}_2\tilde{\boldsymbol{\beta}} - \boldsymbol{y}) + \frac{2\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T(\tilde{\boldsymbol{X}}_2^0\tilde{\boldsymbol{\beta}} - \boldsymbol{y}_0) \\ - \frac{2\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T(\tilde{\boldsymbol{X}}_2^1\tilde{\boldsymbol{\beta}} - \boldsymbol{y}_1) = 0 \end{aligned} \tag{31}$$

where $\tilde{\boldsymbol{X}}_2^0$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m_0]$, $\tilde{\boldsymbol{X}}_2^1$ is the matrix form of $\tilde{\boldsymbol{x}}_{2i}, i \in [m_1]$, $\boldsymbol{y}_0$ is the vector form of $y_i, i \in [m_0]$, $\boldsymbol{y}_1$ is the vector form of $y_i, i \in [m_1]$, and $\boldsymbol{y}$ is the vector form of $y_i, i \in [m]$. Then we can get:

$$\begin{aligned} (\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2 + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\tilde{\boldsymbol{X}}_2^0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\tilde{\boldsymbol{X}}_2^1)\tilde{\boldsymbol{\beta}} \\ = \tilde{\boldsymbol{X}}_2^T\boldsymbol{y} + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\boldsymbol{y}_0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\boldsymbol{y}_1 \end{aligned} \tag{32}$$

Therefore, the solution of $\tilde{\boldsymbol{\beta}}$ is:

$$\begin{aligned} \tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}_2^T\tilde{\boldsymbol{X}}_2 + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\tilde{\boldsymbol{X}}_2^0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\tilde{\boldsymbol{X}}_2^1)^{-1} \\ \cdot (\tilde{\boldsymbol{X}}_2^T\boldsymbol{y} + \frac{\lambda}{m_0}(\tilde{\boldsymbol{X}}_2^0)^T\boldsymbol{y}_0 - \frac{\lambda}{m_1}(\tilde{\boldsymbol{X}}_2^1)^T\boldsymbol{y}_1) \end{aligned} \tag{33}$$

□