Decomposing Direct and Indirect Biases in Linear Models Under Demographic Parity Constraint

Bertille Tierny[1,2], **Arthur Charpentier**[3], **François Hu**[1]

[1]Milliman France, R&D Department, Paris AI Lab
[2]ENSAE - Institut Polytechnique de Paris
[3]Université du Québec à Montréal
bertille.tierny@milliman.com, charpentier.arthur@uqam.ca, francois.hu@milliman.com

Abstract

Linear models are widely used in high-stakes decision-making due to their simplicity and interpretability. Yet when fairness constraints such as demographic parity are introduced, their effects on model coefficients, and thus on how predictive bias is distributed across features, remain opaque. Existing approaches on linear models often rely on strong and unrealistic assumptions, or overlook the explicit role of the sensitive attribute, limiting their practical utility for fairness assessment. We propose a post-processing framework that can be applied on top of any linear model to decompose the resulting bias into direct (sensitive-attribute) and indirect (correlated-features) components. Our method analytically characterizes how demographic parity reshapes each model coefficient, including those of both sensitive and non-sensitive features. This enables a transparent, feature-level interpretation of fairness interventions and reveals how bias may persist or shift through correlated variables. Our framework requires no retraining and provides actionable insights for model auditing and mitigation. Experiments on both synthetic and real-world datasets demonstrate that our method captures fairness dynamics missed by prior work, offering a practical and interpretable tool for responsible deployment of linear models.

Code — https://github.com/bias-mitigator/interpretable.git

1 Introduction

Linear models remain a foundational tool in statistical learning due to their interpretability, scalability, and simplicity (Hastie et al. 2009). They are widely used in high-stakes domains such as credit scoring, hiring, insurance, and healthcare, where algorithmic decisions have significant consequences and fairness considerations are critical (Obermeyer et al. 2019; Barocas, Hardt, and Narayanan 2023). In these settings, linear models may inadvertently encode or amplify unfair biases. These biases can arise *directly*, through the explicit use of sensitive attributes such as race or gender, or *indirectly*, through features correlated with those attributes (Hajian and Domingo-Ferrer 2012; Nabi and Shpitser 2018; Tang, Zhang, and Zhang 2023). Fairness in machine learning has been extensively studied, with various formal definitions and mitigation strategies proposed (Del Barrio, Gordaliza, and Loubes 2020; Mehrabi et al. 2021; Pessach and Shmueli 2022). One of the most common criteria is *Demographic Parity* (DP), which requires that the predictions be statistically independent of sensitive attributes. Although many methods aim to enforce DP in classification settings (Agarwal et al. 2018; Gaucher, Schreuder, and Chzhen 2023; Hu, Ratz, and Charpentier 2024; Denis et al. 2024), few provide systematic tools to quantify and separate the sources of unfairness, especially in linear models. In particular, existing approaches, such as (Chzhen and Schreuder 2022; Fukuchi and Sakuma 2023), do not provide a systematic decomposition of bias stemming from the sensitive feature versus that induced by correlated non-sensitive features. The absence of a clear decomposition is particularly limiting for linear models: despite their transparency, it remains unclear how fairness constraints modify individual coefficients. Consequently, practitioners lack insight into how these constraints redistribute predictive weight across features or whether indirect biases persist after the removal of sensitive variables.

1.1 Main Contributions

We propose a framework for learning fair linear models, designed to identify and mitigate both indirect and direct biases in linear models. Specifically:

- We introduce a linear modeling framework aligned with standard practices and derive a closed-form solution for the optimal fair regressor. To our knowledge, this is the first solution that remains linear under group-wise feature standardization. In practice, it can be applied on top of any linear model (penalized, with or without intercept) making it broadly compatible and easily deployable.
- Building on this optimal solution, we disentangle the contributions of sensitive and non-sensitive features to fairness violations (see Fig. 1) while providing clear guidance on how to adjust coefficients toward fairness.
- We illustrate the effectiveness of our approach on both synthetic and real-world datasets, demonstrating its ability to produce fair linear models while offering interpretability of both direct and indirect biases.

This work advances the understanding of fairness in linear models and contributes to the broader literature by providing tools to dissect and interpret bias at the feature level. For

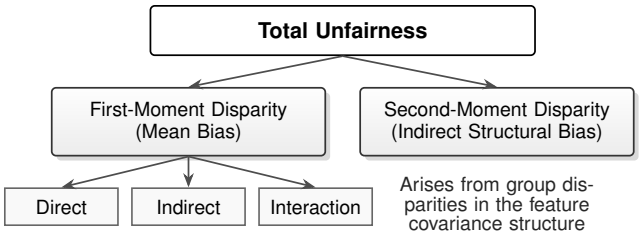

Figure 1: Conceptual decomposition of the total unfairness measure.The unfairness splits into two bias sources: disparities in the **mean** of predictions (First-Moment) and disparities in the **variance** of predictions (Second-Moment).

clarity of presentation, all proofs are provided in the supplementary materials.

1.2 Related Work

The study of fairness constraints in linear regression, particularly under DP, is relatively recent. Most existing methods either focus on model-level fairness objectives or rely on restrictive assumptions that limit their applicability in practice.

(Chzhen and Schreuder 2022) propose a minimax solution for linear regression under DP, deriving a closed-form intercept correction. However, their formulation is based on a strong assumption: the sensitive feature is independent of the other covariates. Therefore, they are omitting completely the indirect biases. This assumption rarely holds in real-world data and significantly restricts both the predictive accuracy of the model and the relevance of its fairness guarantees.

(Fukuchi and Sakuma 2023) extend this line of work by adjusting both intercept and non-sensitive feature coefficients. Although this allows more flexibility, their framework still omits an explicit treatment of the sensitive feature's contribution, which limits bias diagnostics. Moreover, their solution also still builds on simplifying assumptions that may distort the fairness-performance trade-off.

In contrast, our approach explicitly characterizes the effect of DP constraints on all model components, including the sensitive feature. This enables a fine-grained decomposition of direct and indirect biases and provides clearer insights into how fairness interventions affect both predictive behavior and feature-level fairness contributions.

1.3 Outline of the Paper

The remainder of this article is structured as follows: Section 2, introduces the problem setup and the key metrics used throughout the article. Section 3 reviews the limitations of existing fair linear models. Section 4 presents our main contribution: a general framework for learning optimal fair linear models. This is followed in Section 5 by a decomposition of unfairness into direct and indirect biases. Finally, Section 6 details the practical implementation of our methodology and Section 7 presents numerical results comparing our method to state-of-the-art baselines.

	Direct (Mean)	Indirect (Mean)	Interaction	Indirect
[CS22]	✓		✓	
[FS23]	✓	✓	✓	
ours	✓	✓	✓	✓

Table 1: Comparison of bias mitigation methods across linear models proposed by [CS22] (Chzhen and Schreuder 2022), [FS23] (Fukuchi and Sakuma 2023), and our approach. Checkmarks indicate addressed biases.

2 Problem Formulation

Let $(\boldsymbol{X}, S, Y)$ be a random triplet, where $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^d$ is a non-sensitive feature vector, $Y \in \mathcal{Y} \subset \mathbb{R}$ is the target variable, and $S \in \mathcal{S} = [M]$ is a discrete sensitive attribute where $[M] := \{1, \dots, M\}$. We define $p_s = \mathbb{P}(S = s)$ for all $s \in [M]$. Our goal is to find a predictor $f : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ from a set $\mathcal{F}$ that balances predictive utility with fairness. We denote by ν_f the distribution of $f(\boldsymbol{X}, S)$, and by $\nu_{f|s}$ its distribution given $S = s$. We make the following standard assumption.

Assumption 1. *For $f \in \mathcal{F}$, measures $(\nu_{f|s})_{s\in[M]}$ are non atomic with finite second moments.*

We evaluate any predictor f along three key and potentially competing dimensions: predictive risk, fairness, and goodness-of-fit. Each is formally defined below.

2.1 Measuring Risk

We measure the predictive performance of a predictor using the classical quadratic risk, defined as:

$$\mathcal{R}(f) = \mathbb{E}\left[(f(\boldsymbol{X}, S) - Y)^2\right].$$

This risk is uniquely minimized by the Bayes optimal predictor $f^*(\boldsymbol{X}, S) = \mathbb{E}[Y \mid \boldsymbol{X}, S]$, recognizing that fairness constraints entail a trade-off with this optimal benchmark.

2.2 Measuring Unfairness

Our work is grounded in the concept of Demographic Parity, which exists in both a weak and a strong form. In particular, a predictor f satisfies *Weak* DP if its expectation is independent of the sensitive attribute. That is,

$$\mathbb{E}[f(\boldsymbol{X}, S) \mid S = s] = \mathbb{E}[f(\boldsymbol{X}, S)], \quad \text{for all } s \in [M],$$

ensuring fairness at the level of the first moment (the mean).

Definition 2 ((Strong) Demographic Parity)**.** *A predictor f satisfies Strong DP if its entire output distribution is independent of the sensitive attribute. That is,*

$$\nu_{f|s} = \nu_f \quad \textit{for all } s \in [M].$$

This is a much stricter criterion, requiring equivalence of all statistical moments.

Unfairness Measure. We quantify unfairness through the lens of Strong DP, using Wasserstein-2 ($\mathcal{W}_2$) to measure distributional dissimilarities. For further details, we refer the reader to (Santambrogio 2015). Specifically, the unfairness of a predictor f is defined as the weighted sum

of $\mathcal{W}_2$ distance between the group-conditional distributions $(\nu_{f|s})_{s\in[M]}$ and their common barycenter:

$$\mathcal{U}(f) = \min_{\nu\in\mathcal{P}_2(\mathbb{R})} \sum_{s=1}^{M} p_s \mathcal{W}_2^2(\nu_{f|s}, \nu) \ . \tag{1}$$

A predictor f is said to be exactly fair, that is, $\mathcal{U}(f) = 0$ *iff* the predictor satisfies Strong DP. Thus, it provides a measure of how far a model is from achieving exact fairness.

2.3 Measuring Goodness-of-fit

Evaluating fair regression models requires more than assessing overall risk and unfairness. A key consideration is the group-conditional adequacy of the model. The classical coefficient of determination defined as $R^2(f) = \mathrm{Var}(f(\boldsymbol{X}, S))/\mathrm{Var}(Y)$ is a standard metric for explained variance, particularly in linear settings. While it provides a familiar baseline, R^2 can obscure performance disparities and fails to capture group-specific *goodness-of-fit*. For example, a linear model may approximate one group well but fit another poorly, a limitation not revealed by R^2.

Group-Weighted Coefficient of Determination (GWR^2). To diagnose this critical issue, we use the *Group-Weighted* R^2 (GWR^2). This metric is the average of the R^2 computed independently within each sensitive group, providing a direct measure of how well a model fits the data, on average, for all populations under consideration. For a predictor f, the definition is:

$$GWR^2(f) := \sum_{s\in\mathcal{S}} p_s R_s^2(f) \ ,$$

where,

$$R_s^2 = 1 - \frac{\mathrm{Var}(Y - f(\boldsymbol{X}, s) \mid S = s)}{\mathrm{Var}(Y \mid S = s)} \ ,$$

The strength of this metric is theoretically grounded in our analysis of the gap between GWR^2 and the global R^2. Divergence between these two metrics indicates model failure to capture group-specific structures. Thus, GWR^2 is a necessary diagnostic to signal structural mismatch that global metrics can obscure.

3 Limitations of Existing Fair Linear Models

The existing literature on fair linear regression provides foundational solutions but often relies on simplifying assumptions about the data-generating process. We review two key works that represent the progression from handling direct bias to incorporating some forms of indirect bias.

Mitigating Direct Bias. (Chzhen and Schreuder 2022) consider a hypothesis where unfairness arises solely from a group-dependent intercept term:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}_{CS22}\rangle + \beta_{0,CS22}^{(s)} + \zeta, \quad \text{where } \zeta \sim \mathcal{N}(0,1), \tag{2}$$

with the key assumption that features are independent of the sensitive group, i.e., $\boldsymbol{X} \perp\!\!\!\perp S$. In this setting, the associated Bayes optimal predictor is $\langle \boldsymbol{X}, \boldsymbol{\beta}_{CS22}\rangle + \beta_{0,CS22}^{(s)}$. The independence assumption eliminates all sources of indirect bias by construction, isolating direct bias as the only source of unfairness. Therefore, achieving fairness is straightforward.

Lemma 3 (Adapted from (Chzhen and Schreuder 2022))**.** *Given the equation in Eq.* (2)*, the optimal DP-fair predictor is obtained by averaging out the group-specific intercepts:*

$$f_{CS22}(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}_{CS22}\rangle + \sum_{s\in[M]} p_s \beta_{0,CS22}^{(s)}.$$

Mitigating Indirect Mean Bias. (Fukuchi and Sakuma 2023) relax the feature independence assumption, allowing for group-dependent feature means and slopes:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}_{FS23}^{(S)}\rangle + \zeta, \quad \text{where } \zeta \sim \mathcal{N}(0,1), \tag{3}$$

where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(s)}, \sigma_X^2 I)$. This structure introduces an indirect bias that results from the differing feature means $\boldsymbol{\mu}^{(s)}$. However, it maintains a restrictive assumption of homoscedastic, uncorrelated features across groups.

Lemma 4 (Adapted from (Fukuchi and Sakuma 2023), Lemma 1)**.** *Given the model in Eq.* (3)*, the optimal DP-fair predictor is:*

$$f_{FS23}(\boldsymbol{x}, s) = \|\boldsymbol{\beta}_{FS23}^{(.)}\| \langle \tilde{\boldsymbol{\beta}}_{FS23}^{(s)}, \boldsymbol{x} - \boldsymbol{\mu}^{(s)}\rangle + \sum_{s'\in[M]} p_{s'} \langle \boldsymbol{\beta}_{FS23}^{(s')}, \boldsymbol{\mu}^{(s')}\rangle,$$

with

$$\|\boldsymbol{\beta}_{FS23}^{(.)}\| = \sum_{s\in[M]} p_s \|\boldsymbol{\beta}_{FS23}^{(s)}\| \quad \textit{and} \quad \tilde{\boldsymbol{\beta}}_{FS23}^{(s)} = \frac{\boldsymbol{\beta}_{FS23}^{(s)}}{\|\boldsymbol{\beta}_{FS23}^{(s)}\|}.$$

Limitations of Prior Work. While these works represent important progress, they rely on restrictive assumptions about the data covariance structure. In particular, they do not address heteroscedasticity, where the feature covariance matrix $\Sigma^{(s)}$ varies across groups. As a result, it overlooks *indirect structural bias* from distributional disparities, highlighting the need for a more general approach.

4 A General Framework for Optimal Fair Regression

We introduce a linear model framework that captures all key sources of bias, enabling us to derive the optimal fair predictor for more complex, group-dependent data structures.

4.1 The General Model

We consider a setting where the outcome Y is generated by:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta}^{\star}\rangle + \gamma^* S + \beta_0^* + \zeta, \tag{4}$$

where the features $\boldsymbol{X} \mid S = s \sim \mathcal{N}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})$ are group-dependent, and the noise $\zeta \sim \mathcal{N}(0,1)$ is independent of S and $\boldsymbol{X}$. This model captures direct bias (γ^*), indirect mean bias ($\boldsymbol{\mu}^{(s)}$), and indirect structural bias ($\boldsymbol{\Sigma}^{(s)}$).

Our goal is to find the optimal predictor within the class of linear models, $\mathcal{F}_{\text{linear}}$, that minimizes the quadratic risk $\mathcal{R}$ subject to Strong DP. Given $(\boldsymbol{x}, s) \in \mathcal{X} \times \mathcal{S}$, the Bayes optimal predictor is $f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}^{\star}\rangle + \gamma^* s + \beta_0^*$.

4.2 The Optimal Risk-Fairness Trade-off

We seek to find the predictor that optimally navigates the trade-off between minimizing risk and ensuring fairness. To formalize this, we adopt the *ε-Relative Fairness Improvement* (ε-RI) constraint from (Chzhen and Schreuder 2022). A predictor f_ε satisfies this constraint if its unfairness is bounded by an ε-fraction of the Bayes-optimal predictor:

$$\mathcal{U}(f_\varepsilon) \leq \varepsilon \cdot \mathcal{U}(f^*) \ .$$

A key result, applicable to our framework, is that the predictor achieving the optimal risk-fairness trade-off under this constraint, *i.e.*, verifying $f_\varepsilon^* \in \arg\min\{\mathcal{R}(f) : \mathcal{U}(f) \leq \varepsilon \cdot \mathcal{U}(f^*)\}$, is a linear interpolation of the Bayes predictor f^* and the optimal fair predictor f_{DP}^*:

$$f_\varepsilon^* = (1-\sqrt{\varepsilon}) f_{DP}^* + \sqrt{\varepsilon} f^* \ .$$

Our main result is to derive the explicit closed-form expression for f_ε^* within our Gaussian linear model framework.

4.3 Characterizing the Optimal Fair Predictor

To state our main result, we first define the group-conditional mean and standard deviation of the Bayes optimal score:

- Group-conditional mean:

$$\mu_{f^*}^{(s)} := \mathbb{E}[f^*(\boldsymbol{X}, S) \mid S = s] = \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle + \gamma^* s + \beta_0^*.$$

- Group-conditional variance:

$$(\sigma_{f^*}^{(s)})^2 := \mathrm{Var}(f^*(\boldsymbol{X}, S) \mid S = s) = (\boldsymbol{\beta^*})^\top \boldsymbol{\Sigma^{(s)}} \boldsymbol{\beta^*}.$$

We also define their population-level averages, weighted by the group prior probabilities p_s:

$$\bar{\mu}_{f^*} = \sum_{s' \in [M]} p_{s'} \mu_{f^*}^{(s')} \quad \text{and} \quad \bar{\sigma}_{f^*} = \sum_{s' \in [M]} p_{s'} \sigma_{f^*}^{(s')} \ .$$

Proposition 5 (Optimal ε-Fair Predictor)**.** *For the model in Eq.* (4)*, the unique predictor f_ε^* that satisfies the ε-RI constraint and minimizes the quadratic risk is given by:*

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle}{\sigma_{f^*}^{(s)}} \right) + \mu_\varepsilon^{(s)} \ , \qquad (5)$$

where the mean and std are convex combinations of the group-specific and population-averaged statistics:

$$\mu_\varepsilon^{(s)} = (1-\sqrt{\varepsilon}) \bar{\mu}_{f^*} + \sqrt{\varepsilon} \mu_{f^*}^{(s)}$$
$$\sigma_\varepsilon^{(s)} = (1-\sqrt{\varepsilon}) \bar{\sigma}_{f^*} + \sqrt{\varepsilon} \sigma_{f^*}^{(s)} \ .$$

The optimal exactly-fair predictor f_{DP}^* is recovered at $\varepsilon = 0$, and the Bayes optimal predictor f^* is recovered at $\varepsilon = 1$.

4.4 Interpreting the Fairness Mechanism

The structure of f_ε^* reveals a clear and tunable mechanism for enforcing fairness, which can be understood from two complementary perspectives.

Perspective 1: Tunable Standardization and Averaging. This perspective views fairness as the controlled shift of group-dependent moments toward global average moments.

1. **Group-wise Standardization:** within each group s, the term $\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle / \sigma_{f^*}^{(s)}$ creates a standardized score (zero mean and unit variance). This procedure simultaneously removes indirect mean and structural biases.
2. **Controlled Re-scaling and Shifting:** This standardized score is then re-scaled by $\sigma_\varepsilon^{(s)}$ and shifted by $\mu_\varepsilon^{(s)}$. These coefficients are a direct interpolation between the group-specific moments $(\mu_{f^*}^{(s)}, \sigma_{f^*}^{(s)})$ and the global averages $(\bar{\mu}_{f^*}, \bar{\sigma}_{f^*})$. The parameter ε directly control this trade-off: at $\varepsilon = 0$, the predictor uses only global averages, eliminating all bias; at $\varepsilon = 1$, it uses only group-specific values, retaining all original bias for maximum accuracy.

Perspective 2: A Group-Conditional Fair Model. Alternatively, we can express the predictor as a linear model,

$$f_\varepsilon^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta}_\varepsilon^{(s)} \rangle + \beta_{0,\varepsilon}^{(s)} \ ,$$

to see how fairness is encoded into the parameters of the model. By rearranging the terms from Proposition 5, we find the effective slope and intercept for each group are:

$$\boldsymbol{\beta}_\varepsilon^{(s)} = \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \boldsymbol{\beta}^*$$
$$\text{and} \ \ \beta_{0,\varepsilon}^{(s)} = \mu_\varepsilon^{(s)} - \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \langle \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle \ .$$

This view highlights that fairness is achieved by constructing a group-aware model with parameters systematically adjusted to counteract group-specific biases. The scaling factor $\sigma_\varepsilon^{(s)} / \sigma_{f^*}^{(s)}$ compensates for the structural bias, while the intercept $\beta_{0,\varepsilon}^{(s)}$ corrects for the mean-based biases.

5 Decomposition of Direct and Indirect Biases Through the Unfairness

In this section, we develop a comprehensive framework for understanding unfairness in linear regression.

5.1 Prediction-level Decomposition of Unfairness

We begin by decomposing our unfairness measure $\mathcal{U}(f)$ for any predictor within the class of linear models, $\mathcal{F}_{\text{linear}}$.

Proposition 6 (Linear Model Bias Decomposition)**.** *For any predictor $f \in \mathcal{F}_{linear}$ with coefficients $(\boldsymbol{\beta}, \gamma, \beta_0)$, its total unfairness $\mathcal{U}(f)$ decomposes into First-Moment Disparity (FMD) and Second-Moment Disparity (SMD):*

$$\mathcal{U}(f) = \underbrace{\mathrm{Var}(\mathbb{E}[f|S])}_{FMD} + \underbrace{\mathrm{Var}(\sqrt{\mathrm{Var}(f|S)})}_{SMD} . \qquad (6)$$

These components further decompose into four bias sources:

$$\mathcal{U}(f) = \underbrace{\gamma^2 \mathrm{Var}(S)}_{Direct\ Mean} + \underbrace{\mathrm{Var}(\langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{Indirect\ Mean} \qquad (7)$$
$$+ \underbrace{2\gamma \mathrm{Cov}(S, \langle \boldsymbol{\mu}^{(S)}, \boldsymbol{\beta} \rangle)}_{Interaction} + \underbrace{\mathrm{Var}\left(\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{(S)} \boldsymbol{\beta}} \right)}_{Indirect\ Structural} .$$

This decomposition formalizes the conditions required to achieve Strong DP, showing that fairness in this stronger sense necessitates mitigating bias at two distinct levels:

- The **First-Moment Disparity** $\mathrm{Var}(\mathbb{E}[f \mid S])$ captures unfairness in average predictions. It arises from direct dependence on the sensitive attribute (Direct Mean Bias, related to Weak DP) or from correlations between group membership and feature means (Indirect Mean Bias).
- The **Second-Moment Disparity** $\mathrm{Var}(\sqrt{\mathrm{Var}(f \mid S)})$ captures a more subtle form of unfairness (Indirect Structural Bias) where predictive certainty differs across groups due to variations in feature covariance $\mathbf{\Sigma}^{(s)}$.

This decomposition reveals that a model can satisfy Weak DP (without FMD) while remaining unfair under Strong DP. The following corollary demonstrates a key advantage of our optimal ε-fair predictor:

Corollary 7 (Residual Unfairness of our method)**.** *The total unfairness of our predictor f^*_ε, (see Prop. 5), is exactly:*

$$\mathcal{U}(f^*_\varepsilon) = \varepsilon \cdot \mathrm{Var}(\mathbb{E}[f^* \mid S]) + \varepsilon \cdot \mathrm{Var}(\sqrt{\mathrm{Var}(f^* \mid S)}) \ .$$

This corollary highlights a direct, analytical link between a single control parameter (ε) and the total amount of multi-source unfairness, a property not available in prior models.

5.2 Feature-level Decomposition of Unfairness via Approximation

While the prediction-level decomposition quantifies total unfairness, practical intervention requires attributing this unfairness to individual features. A fully additive decomposition is challenging due to the nonlinearity introduced by the square root in the structural bias. To enable interpretability, we apply a first-order Taylor expansion to linearize this term, yielding a tractable and accurate additive approximation.

The Additive Case: Uncorrelated Features. We consider a simplified setting where features are mutually uncorrelated within each group ($\mathbf{\Sigma}^{(s)}$ are diagonal matrices). In this case, the total indirect unfairness of any linear model decomposes into a sum of marginal contributions from each feature.

Proposition 8 (Additive Feature-Level Decomposition)**.** *Given $f \in \mathcal{F}_{linear}$ with coefficients $(\boldsymbol{\beta}, \gamma)$, let its indirect unfairness be $\mathcal{U}_{\mathrm{indirect}}(f) = \mathcal{U}(f) - \gamma^2 \mathrm{Var}(S)$. If all $\Sigma^{(s)}$ are diagonal, then this unfairness can be approximated by an additive sum:*

$$\mathcal{U}_{\mathrm{indirect}}(f) \approx \sum_{j=1}^{d} \mathcal{U}_j^{approx}(f),$$

with the approximate main contribution from feature X_j is:

$$\mathcal{U}_j^{approx}(f) = \underbrace{(\beta_j)^2 \mathrm{Var}(\mu_j^{(S)})}_{\text{Mean}} + \frac{1}{4\bar{V}} \underbrace{(\beta_j)^4 \mathrm{Var}((\sigma_j^{(S)})^2)}_{\text{Structural}} + \underbrace{2\gamma\beta_j \mathrm{Cov}(S, \mu_j^{(S)})}_{\text{Interaction}},$$

where $\mu_j^{(s)} = \mathbb{E}[X_j | S = s]$ and $(\sigma_j^{(s)})^2 = \mathrm{Var}(X_j | S = s)$. Here, $\bar{V} = \mathbb{E}[\mathrm{Var}(f|S)]$ is the average conditional score variance.

This proposition attributes model unfairness to individual features via three pathways: (1) mean disparity, (2) variance disparity (structural bias), and (3) interaction with direct bias. The term $1/(4\bar{V})$ indicates that structural bias diminishes as predictive variance increases.

The General Case: Interactional Unfairness. When features are correlated, the decomposition becomes more complex due to cross-terms capturing *interactional unfairness*. This includes: (1) the compounding of mean biases through correlated feature means, and (2) a deeper structural effect, which we term *Covariance Disparity*, driven by group-level differences in feature correlations.

This analysis provides both practical and comprehensive insight. The additive decomposition highlights features with primary unfairness, while the general case reveals how feature correlations amplify or mitigate these effects.

6 Practical Implementation and Estimation

To apply our framework in practice, the optimal fair predictor must be estimated from finite data, since the population parameters $(\boldsymbol{\beta}^*, \gamma^*, \boldsymbol{\mu}^{(s)}, \mathbf{\Sigma}^{(s)})$ are unknown.

The Plug-in Estimator. The plug-in estimator $\hat{f}_\varepsilon$ of f^*_ε is constructed by replacing all quantities in Prop. 5 with their empirical estimates.

1. **Estimate Model Parameters**. We estimate the base model parameters $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\beta}_0)$. Our framework is agnostic to the fitting procedure; any standard method, such as OLS or penalized version (Ridge, Lasso), is applicable.
2. **Estimating Group Statistics**. For each s, we compute the standard estimates for the group proportions $\hat{p}_s$, feature means $\hat{\boldsymbol{\mu}}^{(s)}$, and feature covariance matrices $\hat{\mathbf{\Sigma}}^{(s)}$.
3. **Assemble the Fair Predictor.** Finally, these empirical components are used to construct the plug-in versions of the conditional score moments ($\hat{\mu}_f^{(s)}$, $\hat{\sigma}_f^{(s)}$) and their population averages ($\hat{\bar{\mu}}_f$, $\hat{\bar{\sigma}}_f$). These are then combined according to Prop. 5 to form the final estimator.

Evaluation Metrics. We evaluate all models on a held-out test set using empirical estimators of our three key metrics. For both the Risk and GWR^2, we consider their empirical counterparts, denoted $\hat{\mathcal{R}}$ (mean squared error) and $\widehat{GWR^2}$, respectively, where:

$$\widehat{GWR^2}(f) = \sum_{s \in [M]} \hat{p}_s \left(1 - \frac{\widehat{\mathrm{Var}}(Y - f \mid S = s)}{\widehat{\mathrm{Var}}(Y \mid S = s)}\right).$$

We quantify the unfairness using the Kolmogorov-Smirnov (KS) test, as it is model-agnostic and does not rely on structural assumptions.

$$\hat{\mathcal{U}}_{\mathrm{KS}}(f) = \max_{s_j, s_k \in [M]} D_{\mathrm{KS}}(\hat{F}_{f|s_j}, \hat{F}_{f|s_k}).$$

Here, $\hat{F}_{f|s}$ is the empirical CDF of scores for group s.

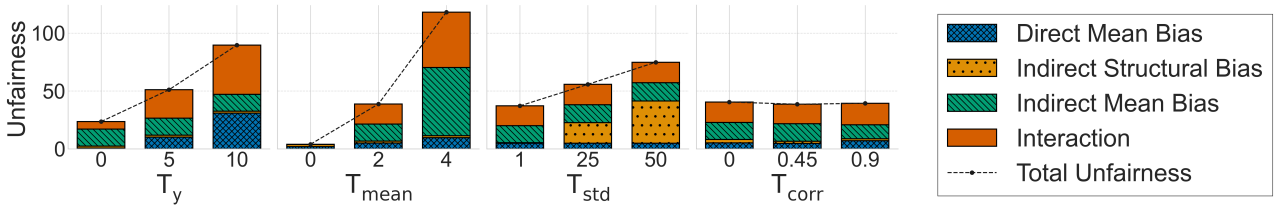

Figure 2: Bias decomposition (see Prop. 6) of a base linear model on synthetic data using by default $\mathrm{T} = (3, 2, 3, 0.7)$.

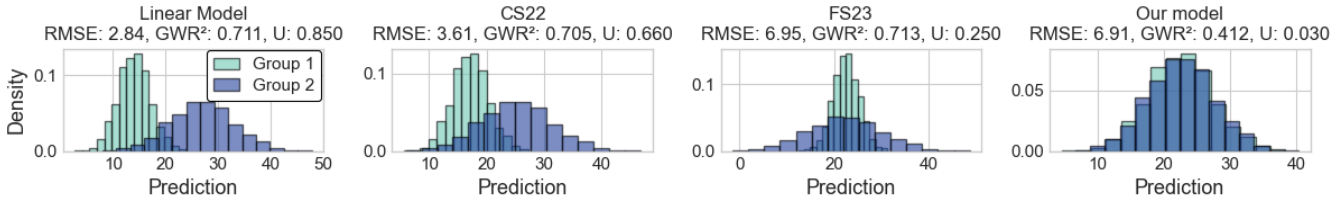

Figure 3: Comparison of group-conditioned model output distribution on synthetic data using $\mathrm{T} = (10, 2, 2, 0.7)$.

7 Numerical Experiments

We run experiments on synthetic and real-world data to: (1) validate our bias decomposition framework, (2) illustrate the transparent remediation capability of our tunable predictor under complex bias scenario.

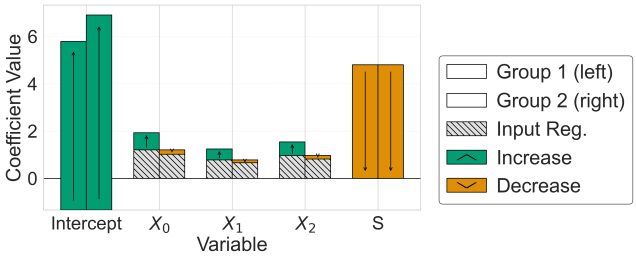

Figure 4: Coefficients adjustments for fairness, shown for a sample of features on synthetic data with $\mathrm{T} = (3, 2, 3, .7)$.

7.1 Application on Synthetic Data

We generated synthetic triplets $(\boldsymbol{X}, S, Y)$. The sensitive attribute $S \in \{1, 2\}$ is drawn from a Bernoulli distribution. Features $\boldsymbol{X} \in \mathbb{R}^d$ follow $\mathcal{N}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})$ conditional on $S = s$, introducing indirect bias through group-specific means, variances and correlations. The outcome $Y = \sum_{j=1}^{d} X_j + T_y \cdot S$ introduces direct bias via T_y. The data-generating process is governed by four control parameters $\boldsymbol{T} := (T_y, T_{\text{mean}}, T_{\text{std}}, T_{\text{corr}})$, mapping to our bias decomposition (Prop. 6). Setting a parameter to zero eliminates the corresponding bias source.

- T_y sets the **direct bias** coefficient γ^*;
- T_{mean} introduces **indirect mean bias** by shifting group 2's means: $\boldsymbol{\mu}^{(2)} = \boldsymbol{\mu}^{(1)} + T_{\text{mean}}$;
- T_{std} and T_{corr} control **indirect structural bias** via group-specific standard deviations ($\sqrt{\Sigma_{jj}^{(2)}} = \sqrt{\Sigma_{jj}^{(1)}} + \sqrt{T_{std}}$) and correlations structures within $\boldsymbol{\Sigma}^{(s)}$: $T_{\text{corr}} = 0$ yields independent features for both groups, while $T_{\text{corr}} \in (0, 1)$ yields different correlation structures between groups.

Experimentation scheme. Given $\boldsymbol{T}$, we create datasets of $d = 5$ features and $n = 20,000$ samples and split it into training (50%), testing (25%), and unlabeled (25%) subsets. As a base model, we use linear regression of Y on $(\boldsymbol{X}, S)$, using `scikit-learn` default parameters. Coefficients of this regression serve as input to build of fair linear model.

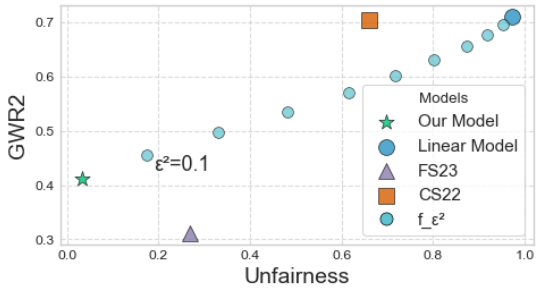

Figure 5: Analysis of Approximate fairness model on synthetic data with $\mathrm{T} = (10, 2, 3, 0.7)$.

Validating the Bias Decomposition. Fig. 2 empirically validate our bias decomposition of a base linear model: increasing (T_y) inflates the Direct Mean and Interaction terms, while increasing T_{mean} and T_{std}) primarily maps to the Indirect Mean and Indirect Structural bias components respectively. This confirms our decomposition effectively identifies the root causes of unfairness in linear models.

Fairness Mitigation and Robustness to Bias Shifts. In complex scenarios with full bias interactions $\boldsymbol{T} = (3, 2, 3, 0.7)$, our model uniquely preserves remediation capabilities (Fig. 3). The remediation operates through the fol-

Model	CRIME			LAW			GOSSIS		
	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness
Base Model Unaware	.45 ± .05	0.15 ± 0.01	0.55 ± 0.04	.15 ± .01	0.37 ± .00	.13 ± .01	.69 ± .01	10.3 ± 0.1	.14 ± .01
Base Model	.46 ± .05	0.15 ± 0.01	0.61 ± 0.04	.15 ± .01	0.37 ± .00	.43 ± .02	.69 ± .01	10.3 ± 0.1	.15 ± .01
CS22	.46 ± .05	0.15 ± 0.01	0.54 ± 0.04	.15 ± .01	0.37 ± .00	.08 ± .01	.69 ± .01	10.3 ± 0.1	.14 ± .01
FS23	.35 ± .09	0.19 ± 0.01	0.20 ± 0.05	.08 ± .05	0.39 ± .01	.15 ± .05	.51 ± .40	12.5 ± 3.3	.13 ± .07
Our model	.38 ± .07	0.19 ± 0.01	**0.12 ± 0.04**	.15 ± .01	0.37 ± .00	**.07 ± .02**	.69 ± .01	10.4 ± 0.1	**.03 ± .01**

Table 2: Comparison of model performances across all datasets. Results are presented as mean ± standard deviation over 50 runs. Bold cells indicate the lowest unfairness.

lowing mechanisms (Fig. 4): (1) **direct bias elimination** via sensitive attribute coefficient nullification and equal intercept compensation; (2) **indirect mean bias correction** through asymmetric intercept adjustment (group 1 receives larger positive shift to offset lower means); (3) **structural bias remediation** via group-specific coefficient scaling (upward for group 1, downward for group 2) and modified intercept adjustments accounting for variance differences; (4) **correlation refinement** adapting coefficients and intercepts to group-specific dependence structures. All adjustments maintain overall predictive accuracy.

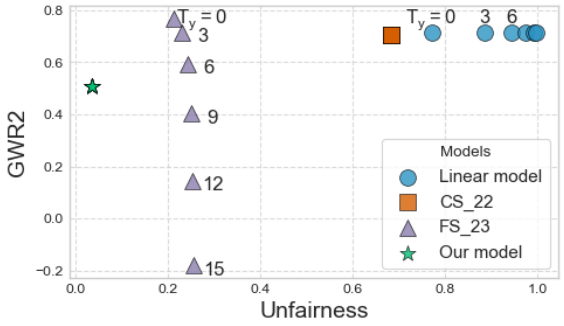

Figure 6: Analysis of Model performance *w.r.t.* direct bias shifts (T_y) on synthetic data using $\mathrm{T} = (*, 2, 2, 0.7)$.

We also test robustness under direct bias shifts by increasing T_y (Fig. 6). Our method and CS22 remain stable in both performance and fairness, while FS23 deteriorates.

Tracing the Optimal Risk–Fairness Frontier. Under ε-RI constraint, ε provides continuous control over the desired fairness level. In a full bias scenario (Fig. 5), our method either achieves higher accuracy than baselines at a given unfairness level, or ensures lower unfairness at a given accuracy.

7.2 Results on Real-World Data

We use three benchmarks. (1) **GOSSIS** (Raffa et al. 2022) contains medical data from over 130,000 patients admitted to intensive care units. The task consists in predicting the vital variable *h1_diaspb_max* with ethnicity as protected attribute. (2) **CRIME** (Redmond and Baveja 2002) includes US communities' demographic and crime data with 1994 samples. We predict the number of violent crimes per 10^5 population with a sensitive attribute based on Black population percentage (Calders et al. 2013). (3) **LAW** covers law school admissions. We predict normalized GPA using race as protected attribute.

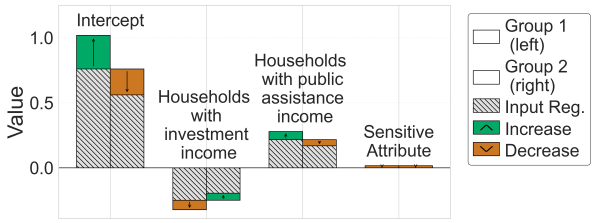

Figure 7: Analysis of coefficient shifts from the linear model to our fair model on the CRIME dataset.

Comparison w.r.t state-of-the-art. Experimental results (Table 2) shows our model effectively reduces unfairness while maintaining competitive predictive performance. The Unaware baseline confirms that omitting the sensitive attribute fails to eliminate discrimination. On LAW, where direct bias dominates, CS22 performs well by mitigating this bias component; nevertheless, our model achieves lower unfairness. Compared to the best-performing baselines, we achieve substantial unfairness reduction for each dataset, while preserving competitive accuracy.

Feature-level interpretation on CRIME Dataset. While the direct bias is nullified (Fig. 7), the model mitigates indirect biases through group-specific coefficient adjustments.

Conclusion

We propose a closed-form solution for fair linear regression that enables exact control over the risk-fairness trade-off via the optimal predictor f_ε^*. Building upon this Gaussian framework, we introduce a novel decomposition of unfairness into direct and indirect components, highlighting four distinct sources, including the previously overlooked **Indirect Structural Bias** arising from disparities in predictive variance.

Our results demonstrate that mean-based fairness alone is insufficient. By explicitly accounting for structural disparities, our method ensures fairness in both average predictions and predictive certainty across groups. The decomposition, along with the Group-Weighted R^2, provides actionable tools for diagnosing unfairness in linear models. While grounded in Gaussian assumptions, our approach shows strong empirical robustness on real-world data. Future work may extend these insights to non-linear models and broader fairness notions.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Calders, T.; Karim, A.; Kamiran, F.; Ali, W.; and Zhang, X. 2013. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, 71–80. IEEE.

Chzhen, E.; and Schreuder, N. 2022. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4): 2416–2442.

Del Barrio, E.; Gordaliza, P.; and Loubes, J.-M. 2020. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*.

Denis, C.; Elie, R.; Hebiri, M.; and Hu, F. 2024. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130): 1–46.

Fukuchi, K.; and Sakuma, J. 2023. Demographic parity constrained minimax optimal regression under linear model. *Advances in Neural Information Processing Systems*, 36: 8653–8689.

Gaucher, S.; Schreuder, N.; and Chzhen, E. 2023. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, 2436–2459. PMLR.

Hajian, S.; and 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.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hu, F.; Ratz, P.; and Charpentier, A. 2024. A sequentially fair mechanism for multiple sensitive attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12502–12510.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Raffa, J. D.; Johnson, A. E. W.; O'Brien, Z.; Pollard, T. J.; Mark, R. G.; Celi, L. A.; Pilcher, D.; and Badawi, O. 2022. The Global Open Source Severity of Illness Score (GOSSIS). *Critical Care Medicine*, 50(7): 1040–1050.

Redmond, M.; and Baveja, A. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3): 660–678.

Santambrogio, F. 2015. *Optimal transport for applied mathematicians*. Springer.

Tang, Z.; Zhang, J.; and Zhang, K. 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s): 1–37.