Decomposing Direct and Indirect Biases in Linear Models Under Demographic Parity Constraint (Student Abstract)

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

[1]Milliman France, R&D Department, Paris AI Lab, 75017 Paris, France
[2]ENSAE Paris - IP Paris, 91120 Palaiseau, France
[3]Université du Québec à Montréal (UQAM), 66023 Montréal, Canada
bertille.tierny@ensae.fr, charpentier.arthur@uqam.ca, francois.hu@milliman.com

Abstract

Linear models are widely used in high-stakes decision-making due to their interpretability, but fairness constraints like Demographic Parity (DP) create opaque effects on model coefficients and predictive bias distribution. We propose a post-processing framework that can be applied on top of any linear model to decompose bias into direct (sensitive-attribute) and indirect (correlated-features) components. Our method analytically characterizes how DP reshapes each coefficient, enabling transparent feature-level interpretation.

1 Introduction

Linear models are extensively used in high-stakes domains such as credit scoring, hiring, and healthcare, where algorithmic decisions must satisfy fairness requirements (Obermeyer et al. 2019; Barocas, Hardt, and Narayanan 2023). However, these models can encode biases both *directly* through sensitive attributes and *indirectly* through correlated features (Hajian and Domingo-Ferrer 2012; Nabi and Shpitser 2018; Tang, Zhang, and Zhang 2023). While Demographic Parity (DP) is a widely adopted fairness criterion requiring predictions to be independent of sensitive attributes, existing approaches for linear models (Chzhen and Schreuder 2022; Fukuchi and Sakuma 2023) rely on overly restrictive assumptions and lack systematic tools for decomposing and interpreting bias sources, leaving practitioners without actionable insights for model auditing and bias mitigation. We address this gap by proposing a post-processing framework that decomposes bias into direct and indirect components while providing closed-form solutions for fair linear regressors. Our contributions include:

- A closed-form solution for optimal fair linear regressors that can be applied to any linear model.
- A decomposition of direct (sensitive-attribute) and indirect (correlated-features) bias contributions.

2 The Proposed Framework

We consider a general setting where the outcome $Y \in \mathbb{R}$ is generated by:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta^*} \rangle + \gamma^* S + \beta_0^* + \zeta, \tag{1}$$

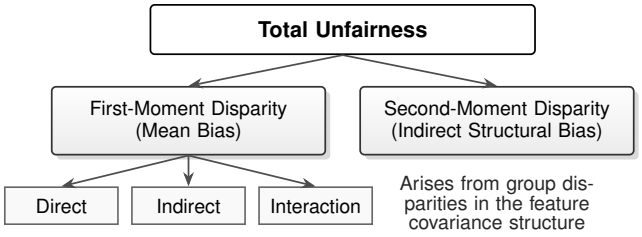

Figure 1: Conceptual decomposition of the total unfairness measure.

where $\boldsymbol{X} \in \mathbb{R}^d$ is a non-sensitive feature vector and $S \in \{1, \dots, M\}$ is a discrete sensitive attribute. We suppose that 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 $\boldsymbol{X}$.

Our key innovation lies in decomposing the unfairness measure of any linear predictor f. The unfairness 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.

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

We decompose the unfairness measure of any linear predictor into *First-Moment Disparity* (differences in prediction means across groups) and *Second-Moment Disparity* (differences in prediction variances and correlations):

$$\mathcal{U}(f) = \underbrace{\mathrm{Var}(\mathbb{E}[f|S])}_{\text{FMD}} + \underbrace{\mathrm{Var}(\sqrt{\mathrm{Var}(f|S)})}_{\text{SMD}}. \tag{3}$$

The FMD component further decomposes into three sources of bias as illustrated in (Fig. 1). This comprehensive decomposition enables practitioners to perform complete bias diagnosis, by identifying and quantifying each source of unfairness.

Beyond bias diagnosis, we develop an interpretable post-processing framework that transforms any linear model into a fair predictor. Our approach enables to navigate the fundamental trade-off between predictive accuracy and fairness by adopting the ε-Relative Fairness Improvement framework from (Chzhen and Schreuder 2022). The optimal predictor

Metric	Base Model	CS22	FS23	Our Model
RMSE	10.3 ± 0.1	10.3 ± 0.1	12.5 ± 3.3	10.4 ± 0.1
Unfairness	$.15 \pm .01$	$.14 \pm .01$	$.13 \pm .07$	$.03 \pm .01$

Table 1: Model performance comparison on GOSSIS dataset. Results are presented as mean $\pm$ standard deviation over 50 runs.

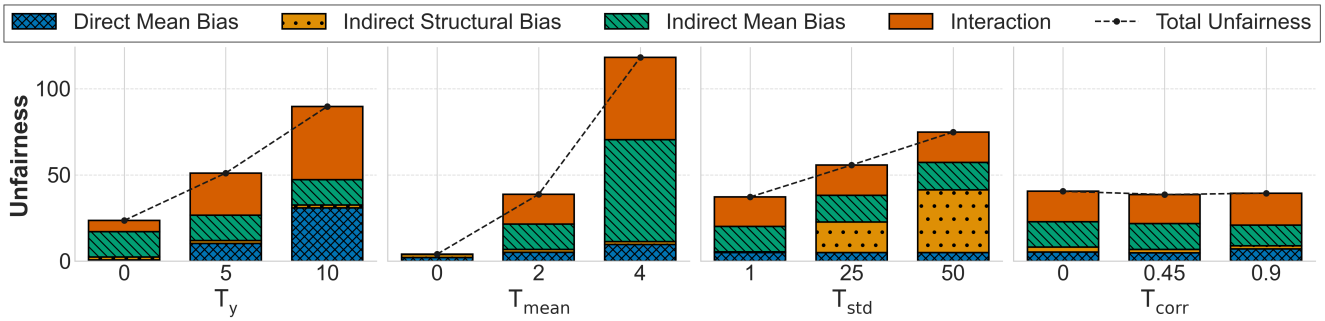

Figure 2: Bias decomposition of the linear model on synthetic data using by default $\mathrm{T} = (3, 2, 3, 0.7)$.

under this constraint, f^*_ε, is a linear interpolation of the fair f^*_{DP} and the Bayes f^* predictors :

$$f^*_\varepsilon = (1 - \sqrt{\varepsilon}) f^*_{DP} + \sqrt{\varepsilon} f^* \quad . \tag{4}$$

Our contribution is deriving the explicit closed-form expression under our general setting (Eq. 1):

$$f^*_\varepsilon(\boldsymbol{x}, s) = \sigma^{(s)}_\varepsilon \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta^*} \rangle}{\sigma^{(s)}_{f^*}} \right) + \mu^{(s)}_\varepsilon \quad , \tag{5}$$

where the mean and std $(\mu^{(s)}_\varepsilon, \sigma^{(s)}_\varepsilon)$ are convex combinations of the group-specific $(\mu^{(s)}_{f^*}, \sigma^{(s)}_{f^*})$ and population-averaged statistics.

3 Experiments

3.1 Synthetic Data

Experimentation scheme We generated synthetic triplets[1] $(\boldsymbol{X}, S, Y)$ where we can precisely control each source of bias through four control parameters $T := (T_y, T_{\text{mean}}, T_{\text{std}}, T_{\text{corr}})$. T_y sets the **direct bias**; T_{mean} introduces **indirect mean bias**; and T_{std} and T_{corr} introduce **indirect structural bias**. When a parameter is set to zero, the corresponding source of bias is absent. The sensitive attribute S is drawn from a Bernoulli distribution.

Validating the Bias Decomposition Fig. 2 applies our decomposition to a linear model trained on synthetic data. The results empirically validate our theory: increasing the direct bias parameter (T_y) primarily inflates the Direct Mean and Interaction terms, while increasing the indirect parameters ($T_{\text{mean}}, T_{\text{std}}$) maps clearly to the Indirect Mean and Indirect Structural bias components, respectively. This confirms our decomposition as a reliable and practical diagnosis tool.

Fairness Mitigation Our framework provides transparency into the mitigation mechanism through explicit coefficient mappings from biased to fair models (Figure 3).

[1]https://github.com/bias-mitigator/interpretable.git

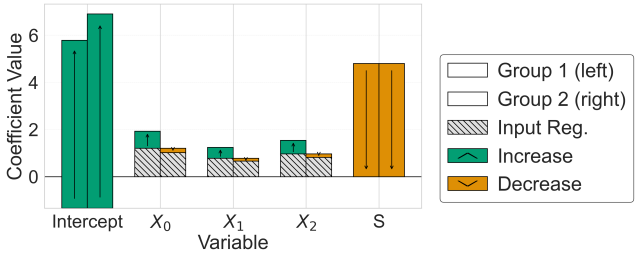

Figure 3: Model coefficient adjustments to achieve fairness in a direct and indirect bias scenario, $\mathrm{T} = (3, 2, 3, 0.7)$.

3.2 Real Data

Comparison w.r.t state-of-the-art. The application of our method on Gossis dataset (Raffa et al. 2022) demonstrate that our model achieves substantial unfairness reductions while maintaining predictive performance (Table 1).

Conclusion & Future Work

This work addresses an important gap in fairness literature by providing the first systematic decomposition of direct and indirect bias sources in linear models under DP. Beyond diagnosis, our interpretable remediation framework enables to achieve optimal performance-fairness trade-offs.

Future work will aim to generalize the bias decomposition methodology to broader model classes, particularly generalized linear models (GLM), and investigate the theoretical question of whether GLM retain their structural properties after fairness interventions.

References

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

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.

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

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.

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.

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.

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.