Quantile-Conditioned Fairness: Extending Binary Fairness Evaluation to Continuous Outcomes

Arushi Jain[1], **Praveen Thoranathula**[2],

[1]Independent
[2]Independent
arushijain1154@gmail.com, praveen.thorana@gmail.com

Abstract

Bias and fairness remain persistent challenges in the responsible deployment of machine learning systems. While most existing metrics are designed for binary classification, fairness evaluation for regression models, widely used in domains such as risk scoring, pricing, and demand forecasting, remains comparatively underexplored. We introduce a quantile-conditioned fairness framework for regression that extends conditional fairness assessment from binary to continuous outcomes. The proposed method partitions target values into quantiles, computes group-to-complement prediction ratios within each segment, and then aggregates these ratios to produce interpretable fairness scores. Through a series of controlled ablation studies on synthetic data, we analyze the effects of bias strength, protected group imbalance, and model performance. We also benchmark our solution against the open-source Dalex fairness toolkit. We further show that the same conditioning principle naturally extends to multiclass classification, treating each class as a conditioning bucket. Real-world case studies on regression and classification datasets demonstrate the practical utility of our approach. Our implementation is lightweight and easily integrable into existing model development workflows, providing a deployable framework for fairness evaluation for all domains.

Introduction

Machine learning models are increasingly deployed in high-stakes domains such as lending, healthcare, and customer retention. As a result, ensuring that such models do not introduce or amplify bias across demographic groups has become a central requirement for responsible AI deployment (Tolan 2019; Liu and Yilin Ning 2025). Existing fairness research and tooling largely centers on binary classification, with metrics such as demographic parity, equalized odds, and calibration. Yet, many deployed systems rely on regression or multiclass models (e.g., risk score, pricing, or demand forecasting), where fairness evaluation remains underexplored and difficult to interpret (Caton and Haas 2024).

While developing fairness evaluation tools for deployed models, we began with a fairness package for binary classification that relied on conditioning on the true outcome and comparing predicted probabilities across protected groups. This approach proved simple, interpretable, and effective in practice. When regression use cases emerged, however, no comparable fairness solution existed in the literature or in open-source libraries. Existing toolkits such as AIF360 (Bellamy et al. 2019) and Fairlearn (Bird et al. 2020) offer only limited regression functionality, and alternative approaches such as Dalex (Baniecki et al. 2021) approach regression fairness indirectly by testing whether predictions add information about group membership beyond targets, which is less interpretable and actionable for practitioners.

We address this gap with a quantile-conditioned fairness framework for regression. The method divides the continuous target into quantiles and, within each, computes the ratio of mean predictions for a protected group versus all others. Aggregating these ratios across quantiles yields an intuitive measure of disparity. The same conditioning principle extends seamlessly to multiclass classification by treating each class as a conditioning bucket.

We validated this framework through extensive synthetic simulations that vary injected bias, group proportions, and model accuracy, and compare its behavior with Dalex. the only prior tool for regression use cases, under identical setups. We further evaluate it on real-world regression and classification datasets, including insurance and adult income tasks (Choi 2018; Becker and Kohavi 1996), showing that it provides actionable insights for practitioners.

Our contributions are:

- A **quantile-based fairness framework** that extends binary conditional fairness testing to regression.
- A **unified treatment** of regression, binary, and multiclass prediction settings.
- **Synthetic stress tests** analyzing sensitivity to bias, class imbalance, and model quality, benchmarked against *Dalex*, the only existing regression fairness toolkit.
- **Real-world validation**, demonstrating interpretability and **practical implementation**, enabling seamless adoption in production workflows.

Related Work

Fairness evaluation has been most thoroughly developed for classification, grounded in three statistical criteria: independence, separation, and sufficiency (Barocas, Hardt, and Narayanan 2023; Steinberg, Reid, and O'Callaghan 2020).

Independence corresponds to demographic parity, requiring predictions to be independent of group membership. Separation underlies equalized odds and equal opportunity, requiring equal error rates across groups conditional on the true outcome. Sufficiency motivates predictive parity and group calibration, requiring outcomes to be independent of group given the model's prediction (McKinnon 2023). Most fairness toolkits provide extensive support for classification fairness evaluation using these metrics.

For regression, definitions and tools are far less mature and standardized. Prior work has proposed mean prediction parity (Calders et al. 2013), distributional comparisons via the Mann–Whitney U or Kolmogorov–Smirnov statistical tests (Zhao and Chen 2019; Agarwal, Dudík, and Wu 2019), and dependence-based measures such as mutual information (Steinberg et al. 2020; Steinberg, Reid, and O'Callaghan 2020). Other works introduce pairwise fairness formulations that assess disparities through instance-wise comparisons rather than group-level aggregates (Narasimhan et al. 2020). Efforts to mitigate bias during model training have also been explored, including correlation-based regularization and ridge-penalty approaches (Komiyama et al. 2018; Marco Scutari 2022). While these methods are theoretically rigorous, they primarily focus on in-training fairness enforcement rather than post-hoc fairness evaluation. Consequently, there is still no consensus on standard post-modeling regression fairness metrics, and thresholds for what constitutes "acceptable" fairness remain ambiguous.

Popular toolkits reflect this imbalance. IBM's AIF360 (Bellamy et al. 2019) implements dozens of classification metrics and mitigation algorithms, but has little direct support for regression. Microsoft's Fairlearn (Bird et al. 2020) provides a flexible framework to evaluate disparities by subgroup, but regression users must define their own fairness metrics or constraints. Other tools such as Aequitas (Saleiro et al. 2019) and Google's Fairness Indicators (Greer et al. 2019) are similarly classification-focused. Dalex (Baniecki et al. 2021) is a notable exception, offering experimental regression fairness via auxiliary prediction-of-group tests, but its indirect formulation limits interpretability.

Our framework directly addresses this gap by using conditional comparisons within outcome quantiles, avoiding arbitrary binarization of continuous outputs and yielding intuitive group-to-rest disparity ratios. This same principle unifies fairness evaluation across regression, binary, and multiclass prediction settings.

Methodology

Proposed Approach

Our approach generalizes a simple and intuitive fairness idea originally used in binary classification: **condition on the true outcome, then compare predicted scores for a group against the rest**.

Binary Classification: Let $Y \in \{0,1\}$ be the true outcome, $\hat{p} = P(Y = 1|X)$ the predicted probability, and A the protected attribute. For each target $y \in \{0,1\}$, compute:

$$r_y(g) = \frac{\mathbb{E}[\hat{p} \mid Y = y, A = g]}{\mathbb{E}[\hat{p} \mid Y = y, A \neq g]} \tag{1}$$

A ratio of 1 indicates parity: group g receives the same predicted probability as others, conditional on the true label. Ratios outside a threshold of $[0.8, 1.25]$ (Feldman et al. 2015) suggest bias.

To ensure boundedness and prevent opposite-direction disparities from canceling out, use a **direction-free version** of this ratio:

$$Fairness_y(g) = \min\left(r_y(g), 1/r_y(g)\right) \in (0,1] \tag{2}$$

Connection to Fairness Theory: Our formulation is grounded in well-established fairness principles. In binary classification, the quantity

$$\mathbb{E}[\hat{p} \mid Y = y, A = g] \tag{3}$$

is a direct empirical estimate of the conditional expectation term that appears in separation-based fairness criteria, which require predictions to be independent of group membership conditional on the true label (Barocas, Hardt, and Narayanan 2023). A perfectly fair model under separation satisfies

$$\hat{Y} \perp A \mid Y \tag{4}$$

Our ratio $r_y(g)$ is therefore a finite-sample diagnostic for violations of separation, and the direction-free transformation $Fairness_y(g)$, produces a bounded, symmetric deviation-from-parity measure.

Generalization to Regression: For continuous outcomes $Y \in \mathbb{R}$, conditioning on exact values is not feasible. Instead, partition Y into quantiles. For each quantile q, compute:

$$r_q(g) = \frac{\mathbb{E}[\hat{y} \mid Y \in q, A = g]}{\mathbb{E}[\hat{y} \mid Y \in q, A \neq g]} \tag{5}$$

Here $\hat{y}$ is the regression prediction. A fairness score for group g is obtained as a weighted average over quantiles:

$$Fairness(g) = \sum_q w_q * Fairness_q(g) \tag{6}$$

where $Fairness_q(g)$ comes from equation (2) and w_q reflects the sample proportion in quantile q.

Rationale for Quantile Conditioning: Partitioning the support of Y into quantiles provides a measurable approximation to conditioning on Y, analogous to forming a Riemann partition of the outcome space. As the number of quantile buckets increases, the aggregated quantity $Fairness_q(g)$ converges to a discretized estimate of the separation condition in equation 4.

Quantiles are preferred over fixed-width bins because they guarantee comparable sample sizes, reduce variance, and avoid pathological partitions in skewed or heavy-tailed distributions. Thus, the proposed method is a generalization of separation-style fairness to continuous outcomes using statistically stable conditional partitions.

Extension to Multiclass Classification: Multiclass is a natural extension of regression in our framework. Instead of quantiles, each class label serves as a conditioning bucket. For class c, compute:

$$r_c(g) = \frac{\mathbb{E}[\hat{p_c} \mid Y = c, A = g]}{\mathbb{E}[\hat{p_c} \mid Y = c, A \neq g]} \tag{7}$$

followed by the transformation in equation (2) and aggregation as explained in equation (6). Because we aggregate a direction-free score, the final score does not indicate over- or under-prediction. However, bias direction can be identified from bucket-level diagnostics of group vs other ratios.

Target Fairness Diagnostics

In addition to measuring fairness in model predictions, our framework also provides diagnostics of unfairness in the target variable itself. This distinction helps practitioners separate disparities originating from biased outcomes from those introduced by the predictive model.

Regression Targets: For continuous outcomes, we compute the point-biserial correlation r_{pb} between group membership ($A = g$ vs. others) and the target Y. The intuition is to test whether the average target differs systematically by group. A fairness score is then defined as:

$$Fairnesstarget(g) = 1 - |r_{pb}(Y, A = g)| \tag{8}$$

Classification Targets: For discrete outcomes, we construct a $2 \times C$ contingency table contrasting group g vs. the rest against the C target classes. We compute Cramér's V, a normalized measure of association in $[0, 1]$. The fairness score is defined as:

$$Fairnesstarget(g) = 1 - V_g \tag{9}$$

For both metrics, a value near 1 indicates little or no association between the group and the target; lower values reveal stronger associations (potentially biased targets).

Synthetic Data Generation

We generate synthetic regression datasets (Algorithm 1) with a continuous target variable, continuous model predictions and categorical protected attributes. The data generator allows explicit control over both the distributional structure of protected groups and the injection of bias into either the targets or the model predictions.

Bias injection. To simulate bias, we modify either the target variable or the predicted scores by applying a multiplicative shift factor to one or more protected groups. Formally, for a group $A = g$ with base target Y_g^0, we define the biased target as:

$$Y_g = Y_g^0 * (1 + \delta_g) \tag{10}$$

where δ_g controls the direction and magnitude of bias ($\delta_g > 0$ indicates overestimation and $\delta_g < 0$ underestimation). The same is followed for biasing predictions.

Experimental factors. We vary four key aspects of the data-generating process:

- **Bias magnitude:** Low vs. high values of δ_g.
- **Group prevalence:** Protected categories with equal, imbalanced, or rare group proportions. Bias can appear in either the prevalent or rare category.
- **Number and alignment of biased categories:** One or two groups may be biased, and their biases may be aligned (overindexed in both) or opposed (overindexed in one, underindexed in another). These settings test whether fairness metrics can detect and attribute multiple simultaneous biases.

Algorithm 1: Synthetic Data Generator for Regression

```
Input: Configuration parameters (bias factors, group proportions, biased groups, noise level)
Output: Synthetic dataset with target, prediction, and protected groups
Initialize $N$ samples and protected groups $\{\text{UTUP}, \text{UTBP}, \text{BTUP}, \text{BTBP}\}$
For each scenario $s$, sample protected groups $A_s \sim \text{Categorical}(\pi_1, \pi_2, ...\pi_n)$
Generate base targets $Y^{(0)} \sim \mathcal{N}(\mu, \sigma)$
Generate base predictions $\hat{Y}^{(0)} = \alpha Y^{(0)} + \mathcal{N}(0, \sigma_{\text{noise}})$
for each scenario $s$ in {UTUP, UTBP, BTUP, BTBP} do
    if $s$ == UTUP then
        No bias applied
    else if $s$ == UTBP then
        Apply prediction bias for group $g$:
        $\hat{Y}_g \leftarrow \hat{Y}_g^{(0)} \times$ pred_bias_factor
    else if $s$ == BTUP then
        Apply target bias for group $g$:
        $Y_g \leftarrow Y_g^{(0)} \times$ target_bias_factor
        Prediction follows target:
        $\hat{Y}_g \leftarrow \hat{Y}_g^{(0)} \times$ target_bias_factor
    else if $s$ == BTBP then
        Apply target and prediction bias for group $g$:
        $Y_g \leftarrow Y_g^{(0)} \times$ both_target_factor,
        $\hat{Y}_g \leftarrow \hat{Y}_g^{(0)} \times$ both_pred_factor
    end if
end for
return Dataset $(Y, \hat{Y}, A_s)$ and bias metadata
```

- **Model performance:** High vs. low signal-to-noise ratios, which influence prediction reliability.

Target–Prediction scenarios. For each configuration, we construct four base cases that represent how bias may appear in the real world:

- **UTUP (Unbiased Targets, Unbiased Predictions):** Neither targets nor predictions contain bias; fairness metrics should indicate parity.
- **UTBP (Unbiased Targets, Biased Predictions):** Targets are unbiased but predictions exhibit systematic bias.
- **BTUP (Biased Targets, Unbiased Predictions):** Both targets and model predictions are biased but the model does not amplify the target bias. Fairness frameworks should correctly treat the model as fair since model just reflects the bias in real-world data.
- **BTBP (Biased Targets, Biased Predictions):** Both targets and predictions are biased but the model is actually accentuating the bias further, representing the most common real-world scenario.

Classification Extension. We also generate synthetic multiclass classification datasets (Algorithm 2) with discrete class labels, class-probability predictions, and categorical protected attributes. The generator follows the same design

Algorithm 2: Synthetic Data Generator for Classification

```
Input: Configuration parameters (bias factors, group proportions, biased groups, noise level)
Output: Synthetic dataset with target, prediction, and protected groups
Initialize $N$ samples and protected groups $\{\text{UTUP}, \text{UTBP}, \text{BTUP}, \text{BTBP}\}$
For each scenario $s$, sample protected groups $A_s \sim \text{Categorical}(\pi_1, \pi_2, ...\pi_n)$
Generate base targets $Y \sim \text{Categorical}(\boldsymbol{p})$
For each class $k$, generate base predictions:
if $Y = k$, $S_k \leftarrow \text{Beta}(\alpha_{\text{correct}}, \beta_{\text{correct}})$
else $S_k \leftarrow \text{Beta}(\alpha_{\text{incorrect}}, \beta_{\text{incorrect}})$
Normalize base predictions (so $\sum_k \hat{p}_k = 1$)
for each scenario $s$ in {UTUP, UTBP, BTUP, BTBP} do
    if $s$ == UTUP then
        No bias applied
    else if $s$ == UTBP then
        Apply prediction bias for group $g$:
        $\hat{\boldsymbol{p}_g} \leftarrow \hat{\boldsymbol{p}_g} \odot \boldsymbol{pred\_bias}$; renormalize to sum to 1
    else if $s$ == BTUP then
        Apply target bias for group $g$:
        Resample $Y_g \sim \text{Categorical}\big(\boldsymbol{\tau_g}^{(\text{target\_bias})}\big)$
        Prediction follows target:
        Redraw base prediction scores and renormalize
    else if $s$ == BTBP then
        Apply target and prediction bias for group $g$:
        Resample $Y_g \sim \text{Categorical}\big(\boldsymbol{\tau_g}^{(\text{target\_bias})}\big)$;
        Redraw base prediction scores and renormalize
        $\hat{\boldsymbol{p}_g} \leftarrow \hat{\boldsymbol{p}_g} \odot \boldsymbol{pred\_bias_g}$; renormalize to sum to 1
    end if
end for
return Dataset $(Y, \hat{Y}, A_s)$ and bias metadata
```

principles as the regression setup, enabling explicit control over both the distribution of protected groups, model quality and the controlled injection of bias into either the true labels or the predicted probabilities.

For each protected group, base class prediction probabilities are drawn from Beta distributions with different parameters for correct vs incorrect class. To introduce bias in targets for a biased group, we manipulate the true class distribution in the group by resampling targets from a skewed class prior $\boldsymbol{\tau}_g$. To introduce bias in predictions, we scale predicted probabilities elementwise with a multiplicative bias pattern $\boldsymbol{pred\_bias}_g$ followed by renormalization.

Controlled synthetic datasets allow us to precisely manipulate bias magnitude, direction, and prevalence—conditions that are rarely isolatable in real data. Evaluating fairness frameworks under these controlled settings provides interpretability and diagnostic clarity: if a method cannot detect bias in well-defined synthetic scenarios, it is unlikely to perform reliably in real-world applications. To complement these controlled tests, we further evaluate the framework on real-world datasets.

Comparison with *Dalex* Approach

The ***Dalex*** library (and its R counterpart *fairmodels*) evaluates fairness in regression using independence, separation, and sufficiency diagnostics. By default, *Dalex* requires users to specify a privileged group. It then computes fairness scores for all other groups relative to this reference. However, in practice the privileged group is often not obvious and identifying one is part of the fairness question. To remove this complication, we adapt Dalex's procedure:

- For each group g, temporarily treat it as privileged.
- Compute Dalex's fairness scores for all other groups relative to g.
- Aggregate to obtain a single score for g

This modification ensures that each group's score is computed without prespecifying a "true" privileged group, yielding fairer and more symmetric comparisons across groups. *Dalex* outputs are magnitude-only ($>= 1$, anchored at $1 = parity$), so aggregating across privileged rotations does not discard directional information, only reduces granularity.

Experiments and Results

Proposed Approach vs *Dalex* on Regression

We first perform a sanity check on the 4 target-prediction scenarios on regression for both the proposed approach and *Dalex*. We make a parity plot using the target fairness scores calculated using the point-biserial correlation as discussed and prediction fairness scores from our approach and *Dalex*. The simulation dataset (generated using $\mu = 50, \sigma = 10, \alpha = 1, \sigma_{noise} = 5, n_{categories} = 3$) assumes low bias in a single category (we bias category 1 whenever bias exists, bias details described below), good model performance, and equal category distribution. We observe:

- UTUP (target and prediction bias multiplier are both 1): Both target and prediction fairness scores are almost on the $45°$ line and hence, fair.
- UTBP (target bias multiplier is 1.2): Target scores are close to 1 ($x = 1$ line) for all classes but prediction scores vary. *Dalex* is quite sensitive to the bias and shows category 1 outside the bias thresholds.
- BTUP (target and prediction bias multiplier are both 1.2): Prediction scores are close to 1 ($y = 1$ line) for all classes but target scores vary.
- BTBP (target bias multiplier is 0.85 and prediction bias multiplier is 0.75): Both target and prediction fairness scores are away from the $45°$ line and hence, show bias. *Dalex* is quite sensitive to the bias and shows category 1 outside the bias thresholds.

We next try different experimental factors such as bias strength, category distribution, number of biased categories and model performance to compare our approach and *Dalex*. We compare across 3 synthetic setups (detailed in appendix) and discuss results for the BTBP case since this is the most common real-world scenario (All configs generate targets from $\sim \mathcal{N}(50, 10)$).

Based on Figures 2, 3, and 4, our framework produces direct group-to-rest ratios that are immediately interpretable

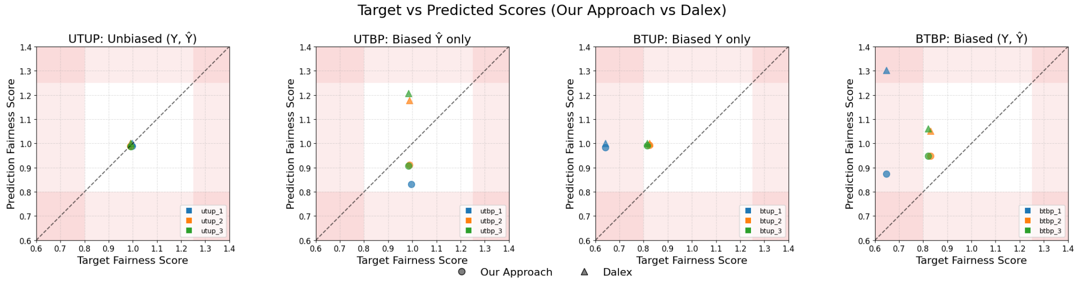

Figure 1: Parity check of regression fairness under four synthetic bias scenarios (UTUP, UTBP, BTUP, BTBP). Each point represents a protected group; circles = our method, triangles = Dalex. The white region denotes the acceptable fairness band [0.8,1.25]. Our method tracks prediction bias without inflating disparities, whereas Dalex shows much larger deviations. Dalex values exceeding the axis range (e.g., 6.65 in UTBP Category 1) are truncated for readability.

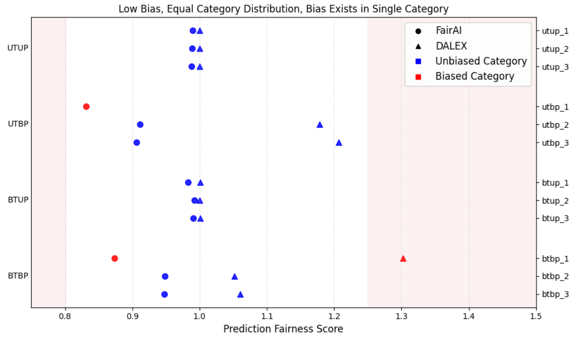

Figure 2: Low Bias, Equal Category, Single Class Biased Simulation. Points outside the fairness band [0.8,1.25] (in red zone) indicate bias. Despite only small injected bias, Dalex flags UTBP and BTBP as unfair; in UTBP, its prediction fairness reaches 6.65 (truncated here for readability).

and map cleanly onto quantitative fairness deviations. In contrast, *Dalex* relies on inference-based adjustments whose practical meaning is less transparent. This difference becomes especially apparent in low-bias settings, where *Dalex* often reports extremely large fairness scores even when only minimal bias is injected. Our method also provides clearer attribution of bias to the correct protected category, whereas *Dalex* can misidentify unbiased categories as biased. In multi-category settings with poor model performance, our approach continues to detect bias when it exists, while *Dalex*—being tightly coupled to the model's raw prediction scores—can fail to do so reliably.

Overall, these experiments show that our metric produces stable, interpretable fairness ratios that scale with bias strength, remain robust under imbalance and model quality, and accurately attribute disparities to specific groups. In contrast, *Dalex* often detects only aggregate effects and lacks the resolution to identify which groups and outcome regions are driving the bias.

Ablation Study on Regression

Because regression is the main focus of our contribution, we perform ablation studies on the synthetic regression datasets.

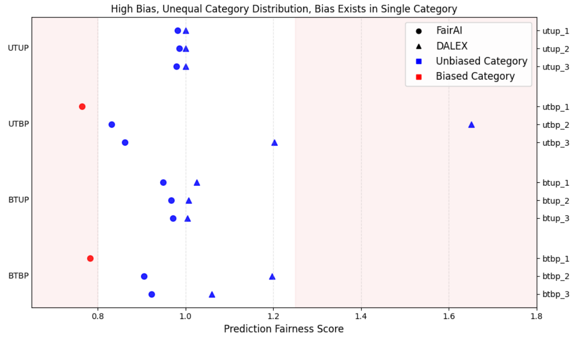

Figure 3: Results on High Bias, Unequal Category, Single Class Biased Simulation. While both our approach and *Dalex* detect bias when it exists (red dots go in red zones), *Dalex* produces extremely large fairness scores (e.g., 51.38 in UTBP-Category 1 and 1.82 in BTBP-Category 1, truncated in graph for readability) which may not directly reflect the amount of bias, making interpretation difficult. It also flags UTBP-Category 2 as unfair despite no injected bias.

All ablations generate targets from $\sim \mathcal{N}(50, 10)$, use a three-category setting with one biased group (category 1) and repeat evaluation across 30 random seeds, reporting the mean fairness score and 95% confidence intervals. We observe the following from Figure 5:

- Varying Bias magnitude: Here, we assume equal category distribution and $\sigma_{noise} = 5$. As prediction bias strengthens relative to target bias, fairness for the biased group falls below the 0.8 threshold. The trend is generally monotonic within confidence intervals, confirming that our metric is sensitive to bias strength.
- Group distribution: Here, we assume $target\_bias = 0.75$, $prediction\_bias = 0.6$ and $\sigma_{noise} = 5$. Under equal distributions, fairness scores for the biased group diverge cleanly from those of unbiased groups. With skewed distributions, the biased group still registers as less fair, but the error bars widen, particularly for smaller categories.
- Model performance: Here, we assume equal category

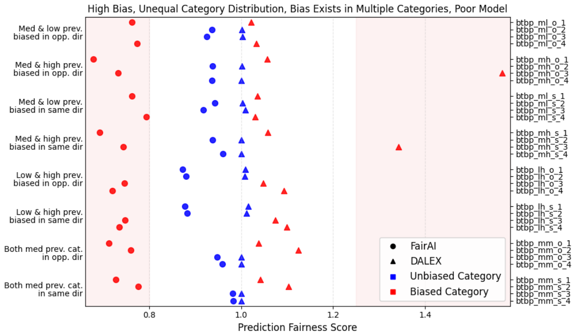

Figure 4: Results on High Bias, Unequal Category, Multiple Class Biased, Poor Model Simulation. Our approach detects bias in most cases (red circles in red biased region) whereas Dalex fails to detect bias for a lot of biased categories (red triangle in white regions).

distribution, $target\_bias = 0.75$ and $prediction\_bias = 0.6$. With stronger predictive models, fairness scores display clearer separation between biased and unbiased groups. As model quality degrades, the fairness scores become noisier. Nonetheless, in most cases the biased group still deviates from parity, showing that the metric is informative even when predictions are weak, though error bars widen slightly.

Overall, the ablation results show that the proposed metric responds appropriately to injected bias, class imbalance, and model noise. Fairness estimates are stable across 30 random seeds, with tight confidence intervals in Figure 5. While imbalance and high noise naturally increase uncertainty, the metric consistently identifies biased categories.

Results on Multiclass Extension

We evaluate the multiclass extension of our framework using synthetic datasets with 3 outcome classes. We consider the case of 3 protected categories where exactly one (category 1) is biased on the BTBP case. To introduce target bias, we change class 0 target distribution in biased category to $\tau_g^{target\_bias} = (0.2, 0.5, 0.3)$ and to introduce prediction bias, we multiply prediction scores with $pred\_bias = (0.5, 3, 1.5)$. We consider three scenarios:

- Equal Distribution: Category distribution is equal and correct class probabilities are sampled from $\beta(7, 2)$ and incorrect class probabilities are sampled from $\beta(1.5, 6)$.
- Unequal Distribution: Category distribution is (0.4, 0.4, 0.2) and correct class probabilities are sampled from $\beta(7, 2)$ and incorrect class probabilities are sampled from $\beta(1.5, 6)$.
- Equal Distrbution with poor model: Category distribution is equal and correct class probabilities are sampled from $\beta(4, 3)$ and incorrect class probabilities are sampled from $\beta(2, 4)$.

Based on Figure 6, our appraoch is able to detect bias in all 3 cases. We can further see the individual class fairness scores to decode unfair category-class combinations.

Feature	Class	Target	Prediction
sex	female	0.943	0.963
sex	male	0.943	0.964
smoker	no	0.213	**0.115**
smoker	yes	0.213	**0.766**
region	northeast	0.994	0.924
region	northwest	0.960	0.955
region	southeast	0.926	0.894
region	southwest	0.957	0.958

Table 1: Results on Insurance dataset for Regression Task

Evaluation on Real-world Data

Regression Task on Insurance Charges: We train a simple linear regression model on the publicly available Medical Cost Insurance dataset (Choi 2018) to predict insurance charges. We consider categorical variables region, sex, and smoker as protected attributes and do not include them in model inputs. We observe from table 1:

- For demographic variables (sex/ region), targets are relatively balanced and predictions are close to parity.
- For domain variables (smoker), strong disparities appear in both labels and predictions, with the model accentuating the bias. These differences represent causal effects (smoking → higher charges).

Classification Task on Adult Income: We train a simple logistic regression model on the publicly available Adult dataset (Becker and Kohavi 1996) to predict high vs low income. We consider categorical variables sex, race and occupation as protected attributes and do not include them in model inputs. We observe from table 2:

- For demographic variables (sex, race), the targets show some disparity but predictions are close to parity.
- For domain variables (occupation), mild disparities appear in both labels and predictions, with the model slightly accentuating the bias.

Across both datasets, our framework highlights meaningful contrasts in terms of target and prediction fairness. We also demonstrate the importance of jointly reporting target fairness and prediction fairness to properly interpret whether disparities originate from data labels, model outputs, or domain realities.

Practical Considerations

Complexity and deployability. The proposed metric is computationally lightweight because it requires only groupwise conditional averages over quantile buckets, making it a simple post-hoc transformation of model outputs rather than a retraining step. Its computation scales linearly with dataset size and number of quantiles, dominated by a single grouping operation, and therefore remains efficient even for large datasets and fine-grained partitions. Since the method depends only on predictions and protected-group labels, it can be seamlessly integrated into existing pipelines without reaccessing the underlying model, supporting its deployability in practical settings.

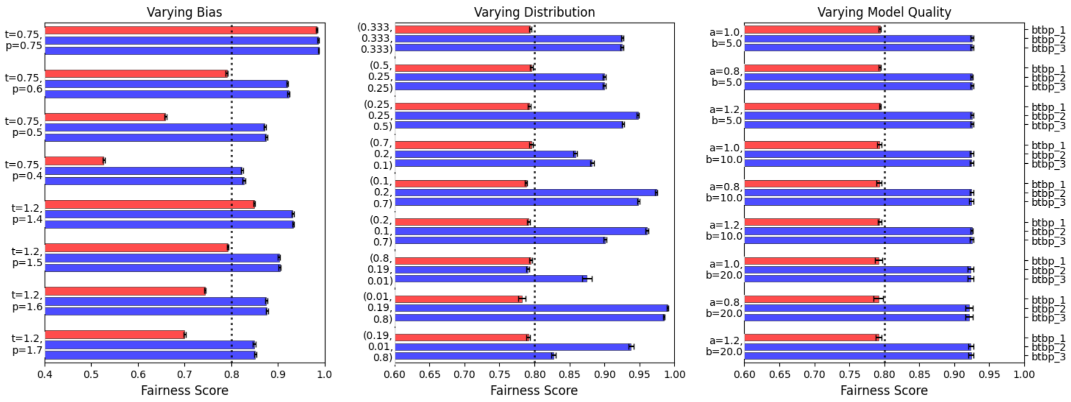

Figure 5: Ablation study of our regression fairness metric under 3 synthetic settings. Each panel shows mean fairness across 30 seeds with 95% confidence intervals; Category 1 is the only biased group, and values near 1 indicate parity. (a) Varying Bias: Each group corresponds to a different (target bias, prediction bias) scenario. (b) Varying Distribution: Groups correspond to different category proportions. (c) Varying Model Quality: Groups correspond to models of the form $\hat{Y} = aY + \mathcal{N}(0, b)$.

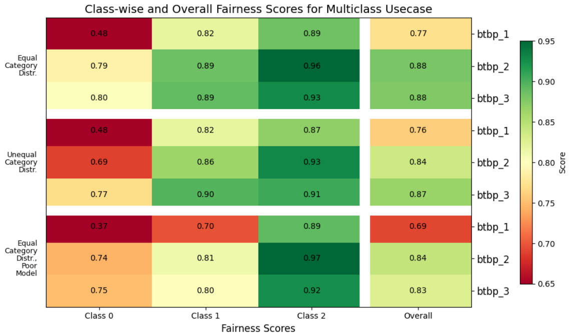

Figure 6: Results on Multiclass Scenarios. Each row group represents a scenario, each row within a group represents protected category with category 1 biased. First 3 columns indicate class-level fairness and last column indicates overall fairness. Our approach is able to detect bias in all 3 cases.

Feature	Class	Target	Prediction
sex	female	0.784	0.975
sex	male	0.784	0.966
race	amer-ind-esk	0.972	0.965
race	asian-pac-isl	0.990	0.958
race	black	0.911	0.962
race	other	0.969	0.940
race	white	0.915	0.980
occup.	adm-clerical	0.910	0.953
occup.	armed-force	0.997	0.895
occup.	craft-repair	0.988	0.904
occup.	exec-manag	0.785	0.868
occup.	farm-fish	0.948	0.968
occup.	hand-clean	0.913	0.893
occup.	mach-op-inspct	0.931	0.894
occup.	other-service	0.844	0.909
occup.	priv-house-serv	0.963	0.925
occup.	prof-specialty	0.814	**0.770**
occup.	protective-serv	0.972	0.906
occup.	sales	0.964	0.992
occup.	tech-support	0.975	0.919
occup.	transport-moving	0.979	0.926

Table 2: Results on Adult dataset for Classification Task

Choice of quantiles. In all regression experiments we fix the number of quantile buckets to a single value (10) for simplicity. The choice of quantile granularity induces a standard bias-variance trade-off: using fewer buckets yields more stable but coarser estimates of conditional fairness, whereas using many buckets increases resolution at the cost of higher variance, especially for small protected groups. In practice, we recommend checking that each protected group has a minimum level of support in the majority of buckets (e.g., via simple count diagnostics), and re-running the analysis with a coarser partition as a robustness check when group sizes are very small.

Conclusion and Future Work

We presented a unified framework for fairness evaluation in regression and classification by conditioning on outcome partitions and comparing group-to-rest prediction ratios, extending outcome-conditioned fairness testing beyond the binary setting. Synthetic experiments show that the metric sensitively tracks injected bias, provides interpretable per-group diagnostics, and avoids the instability of auxiliary model inference-based baselines such as *Dalex*, while remaining lightweight and easy to integrate as a post-hoc analysis. Results on real-world datasets further demonstrate its practical utility in identifying domain-sensitive fairness disparities. Future work includes a more systematic study of the multiclass extension, broader benchmarking across diverse real-world datasets, and a deeper examination of quantile granularity selection.

References

Agarwal, A.; Dudík, M.; and Wu, Z. S. 2019. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. arXiv:1905.12843.

Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.; and Biecek, P. 2021. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *Journal of Machine Learning Research*, 22(214): 1–7.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4:1–4:15.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

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.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7): 1–38.

Choi, M. 2018. Medical Cost Personal Datasets. Kaggle.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. arXiv:1412.3756.

Greer, C.; Joshi, M.; Fang, H.; Jindal, S.; Shukla, K.; Aka, O.; Kleinfeld, S.; Chang, A.; Hanna, A.; and Nanas, D. 2019. Fairness Indicators: Scalable Infrastructure for Fair ML Systems.

Komiyama, J.; Takeda, A.; Honda, J.; and Shimao, H. 2018. Nonconvex Optimization for Regression with Fairness Constraints. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2737–2746. PMLR.

Liu, M.; and Yilin Ning, X. L. M. M. Y. S. X. L. D. M. J. L. J. X. D. S. W. T. L. T.-E. C. J. C. L. O. Z. L. T. T. F. T. N. R. F. W. L. A. C. M. E. H. O. N. L., Salinelat Teixayavong. 2025. A scoping review and evidence gap analysis of clinical AI fairness. *NPJ digital medicine*, 8(360).

Marco Scutari, M. P., Francesca Panero. 2022. Achieving fairness with a simple ridge penalty. *Statistics and Computing*, 32(77).

McKinnon, A. D. 2023. Assessing Bias in ML Models.

Narasimhan, H.; Cotter, A.; Gupta, M.; and Wang, S. 2020. Pairwise Fairness for Ranking and Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5248–5255.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2019. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577.

Steinberg, D.; Reid, A.; and O'Callaghan, S. 2020. Fairness Measures for Regression via Probabilistic Classification. arXiv:2001.06089.

Steinberg, D.; Reid, A.; O'Callaghan, S.; Lattimore, F.; McCalman, L.; and Caetano, T. 2020. Fast Fair Regression via Efficient Approximations of Mutual Information. arXiv:2002.06200.

Tolan, S. 2019. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. arXiv:1901.04730.

Zhao, C.; and Chen, F. 2019. Rank-Based Multi-task Learning for Fair Regression. In *2019 IEEE International Conference on Data Mining (ICDM)*, 916–925.

APPENDIX

Configurations for Regression Synthetic Experiments

Simulation	Config
Low bias, equal category distribution with single-biased category	α=1, σ_{noise}=5, 3 categories with category 1 biased, equal category distribution, target_bias = 0.85, pred_bias = 0.75
High bias, unequal category distribution with single-biased category	α=1, σ_{noise}=5, 3 categories with category 1 biased, category distribution = (0.4, 0.4, 0.2), target_bias = 0.75, pred_bias = 0.60.
High bias, unequal category distribution with two biased categories, poor model performance	α=1, σ_{noise}=20, 4 categories with 2 categories biased (mentioned in figure), category distribution = (0.25, 0.25, 0.45, 0.05), target_bias for cat1 = 1.2, pred_bias for cat2 = 1.4, target_bias for cat2 = 1.4 (same) or 0.85 (opp), pred_bias for cat2 = 1.1 (same) or 0.65 (opp)

Table 3: Comparison of Our Approach with Dalex: Simulation configs