# Model Mis-specification and Algorithmic Bias

**Runshan Fu, Yangfan Liang, Peter Zhang**

Carnegie Mellon University

{runshanf, yangfanl, yunz2}@andrew.cmu.edu

## Introduction

Consider a bank that uses a machine learning model to make loan approval and interest rate decisions. If individuals from one demographic group are systematically assigned higher risk scores than they deserve, and individuals from another group are systematically assigned lower risk scores than they deserve, then this may be undesirable.

Our goal in this work is to understand if and how such discrepancy is due to the algorithmic part of the machine learning process. More specifically, we want to answer these questions:

- How do we define such discrepancy?
- When would a machine learning model create discrepant outcomes for different groups of individuals?
- And when that happens, can we quantify the discrepancy?

## Problem Statement

### Main Assumptions

- The data used to train the machine learning model represent the true underlying distribution of feature and outcome variables, *e.g.*, there is no sampling bias or labeling error.
- The machine learning model produces a risk score for each individual – and in the example above, it is in turn used to make (binary) loan approval decisions and (continuous or discrete) interest rate decisions – but we focus on risk score only, not the downstream decisions such as threshold-setting in common classification tasks.

### Definitions

We denote the outcome variable that the machine learning model aims to predict as $Y$, and assume that in the true data generating process (DGP), $Y$ is a function of the feature vector $\mathbf{X} = (X_1, X_2, ..., X_n)^\intercal$:

$$Y = h(\mathbf{X}). \tag{1}$$

We use $\phi$ to denote a trained machine learning model, and the *prediction error* is defined as

$$e = \hat{Y} - Y = \phi(\mathbf{X}) - Y.$$

In this paper, we partition our results into two cases: $\phi$ is *correctly specified*, and $\phi$ is *mis-specified* (when compared with $h$). We further differentiate mis-specification into *mismatch between function classes* and *variable omission*. For example, if $h$ is a second-degree polynomial function, and $\phi$ is drawn from the family of linear functions, then $\phi$ is mis-specified in the sense of function class mismatch. If $\mathbf{X}_{\text{short}}$ is a feature vector that contains a subset of the original features, and $\phi$ only depends on $\mathbf{X}_{\text{short}}$, then $\phi$ is mis-specified in the sense of variable omission. For the latter, we also slightly abuse the notation and write $\phi(\mathbf{X}_{\text{short}})$ instead of $\phi(\mathbf{X})$.

### Our Bias Measure

Recall our first question: How do we define such discrepancy?

We denote the *population-level mean prediction error* as $b$,

$$b(\phi) = \mathbb{E}(e|\phi).$$

Prediction error $e$ is closely related to the concept of "bias" in bias-variance trade-off, as $\hat{Y}$ can be viewed as the expected prediction of $Y$ over different possible training sets given by a machine learning model $\phi$.

For exposition, we assume there are two groups of individuals: a protected group and a regular group, but many results can directly generalize to an arbitrary number of groups. We use $A$ to denote the group attribute (or sensitive attribute), where $A = 1$ represents the protected group and $A = 0$ represents the regular group. With a slight abuse of notation, the *group-level mean prediction errors* are defined as $b(\phi, A = 0) = \mathbb{E}(e|\phi, A = 0)$ and $b(\phi, A = 1) = \mathbb{E}(e|\phi, A = 1)$, respectively.

We define *outcome bias attributable to algorithm* (more simply, *algorithmic bias* or *outcome bias*) as

$$\tau(\phi) = b(\phi, A = 1) - b(\phi, A = 0).$$

In other words, bias is defined as the systematic difference in prediction errors between the two groups.

For example, if $Y$ is individual creditworthiness, and for a machine learning model, $b(\phi, A = 1) > b(\phi, A = 0)$, then this suggests that the machine learning model produces systematically higher creditworthiness scores for the individuals in the regular group, compared to those in the protected group. When $Y$ is a desirable variable (*i.e.*, higher $Y$ leads to more favorable decisions), then a positive $\tau$ suggests bias against the protected group, and a negative $\tau$ suggests bias against the regular group. If $\tau(\phi) = 0$, then $\phi$ is considered unbiased (or fair).

## Regressions

### Data generating process

For simplicity, we assume that there are two relevant random variables, $X_1$ and $X_2$, and $h$ is a linear function:

$$h(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where $\epsilon$ is random noise with $\mathbb{E}(\epsilon|\mathbf{X}) = 0$ and $\mathbb{E}(\epsilon|A = 1) = \mathbb{E}(\epsilon|A = 0) = 0$.

Let $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \beta_2)$, and $\mathbf{X} = (1,\ X_1,\ X_2)$ with a slight abuse of notation, then the true data generating process can be written as

$$Y = \mathbf{X}^\intercal \boldsymbol{\beta} + \epsilon.$$

### Correctly Specified Model

**Proposition 1.** *When the model is correctly specified, the mean prediction error in the population equals zero. That is, if $\xi$ is a random error with $\mathbb{E}(\xi|\mathbf{X}) = 0$ and $\phi(X) = \mathbf{X}^\intercal\boldsymbol{\gamma} + \xi$, then $\mathbb{E}(e) = 0$.*

*Proof.* $\gamma$ is estimated through the following criteria

$$\boldsymbol{\gamma} = \underset{\boldsymbol{\gamma}'}{\operatorname{argmin}}\, \mathbb{E}\left[(Y - \mathbf{X}^\intercal\boldsymbol{\gamma}')^2\right]. \tag{2}$$

We can easily find that the solution to Eq.2 satisfies the following:

$$\boldsymbol{\gamma} = (\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal))^{-1}\,\mathbb{E}(\mathbf{X}Y) = (\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal))^{-1}\,\mathbb{E}(\mathbf{X}(\mathbf{X}^\intercal\boldsymbol{\beta} + \epsilon)) = \boldsymbol{\beta}. \tag{3}$$

Thus, the best linear projection of $Y$ given $\mathbf{X}$ is

$$\hat{Y} = \mathbf{X}^\intercal\boldsymbol{\gamma} = \mathbf{X}^\intercal\boldsymbol{\beta}. \tag{4}$$

By Eq.4, the prediction error $e = \hat{Y} - Y$ would be equal to the negative random noise $\epsilon$

$$e = \hat{Y} - Y = \mathbf{X}^\intercal\boldsymbol{\beta} - Y = -\epsilon.$$

Therefore, we have

$$\mathbb{E}(e) = \mathbb{E}(-\epsilon) = 0.$$

□

**Corollary 2.** *When the model is correctly specified, the algorithmic bias equals zero. That is, $\tau = \mathbb{E}(e|A = 1) - \mathbb{E}(e|A = 0) = \mathbb{E}(\epsilon|A = 1) - \mathbb{E}(\epsilon|A = 0) = 0$.*

This result follows directly from assumptions.

Thus, we can see that as long as the model is correctly specified, there needs not be any bias, even if feature values are correlated with the sensitive attribute $A$.

### Model Mis-specification and Bias

Consider the case when the model is mis-specified with omitted variables ($X_2$ is omitted from the model). Here the estimation model is

$$Y = \mathbf{X}_{\text{short}}^\intercal \boldsymbol{\gamma}_{\text{short}} + u,$$

where $\mathbf{X}_{\text{short}} = (1, X_1)$, $\boldsymbol{\gamma}_{\text{short}} = (\gamma_0, \gamma_1)$, and $u$ is a random error with $\mathbb{E}(u|\mathbf{X}_{\text{short}}) = 0$. We denote the prediction error as $e_{\text{short}}$:

$$e_{\text{short}} = \hat{Y}_{\text{short}} - Y.$$

**Proposition 3.** *When the model is mis-specified with omitted variables, the mean prediction error in the population equals zero,* i.e., $\mathbb{E}(e_{short}) = 0$.

*Proof.* Similar to Eq.3, here $\boldsymbol{\gamma}_{\text{short}}$ is derived as follows:

$$\begin{aligned}\boldsymbol{\gamma}_{\text{short}} &= (\mathbb{E}(\mathbf{X}_{\text{short}}\mathbf{X}_{\text{short}}^\intercal))^{-1}\,\mathbb{E}(\mathbf{X}_{\text{short}}Y)\\ &= \frac{1}{\mathbb{E}(X_1^2) - (\mathbb{E}\,X_1)^2}\begin{pmatrix}\beta_0(\mathbb{E}(X_1^2) - (\mathbb{E}\,X_1)^2) + \beta_2(\mathbb{E}\,X_1^2\,\mathbb{E}\,X_2 - \mathbb{E}\,X_1\,\mathbb{E}\,X_1X_2)\\ \beta_1(\mathbb{E}(X_1^2) - (\mathbb{E}\,X_1)^2) + \beta_2(\mathbb{E}\,X_1X_2 - \mathbb{E}\,X_1\,\mathbb{E}\,X_2)\end{pmatrix}.\end{aligned}$$

Thus, we know that:

$$\begin{aligned}\gamma_0 &= \beta_0 + \frac{\mathbb{E}\,X_1^2\,\mathbb{E}\,X_2 - \mathbb{E}\,X_1\,\mathbb{E}\,X_1X_2}{Var(X_1)}\beta_2,\\ \gamma_1 &= \beta_1 + \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2.\end{aligned}$$

The prediction error in this model would be as follows:

$$\begin{aligned}e_{\text{short}} &= \hat{Y}_{\text{short}} - Y = (\gamma_0 + \gamma_1 X_1) - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon)\\ &= (\gamma_0 - \beta_0) + \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2 X_1 - \beta_2 X_2 - \epsilon.\end{aligned} \tag{5}$$

Therefore we have:

$$\mathbb{E}(e_{\text{short}}) = \frac{\mathbb{E}\,X_1^2\,\mathbb{E}\,X_2 - \mathbb{E}\,X_1\,\mathbb{E}\,X_1X_2}{Var(X_1)}\beta_2 + \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2\,\mathbb{E}\,X_1 - \beta_2\,\mathbb{E}\,X_2 = 0.$$

□

**Proposition 4.** *When the model is mis-specified with omitted variables, the group-level mean prediction errors would not equal zero, unless for each feature, its conditional expectation (on $A$) is the same for $A = 0$ and $A = 1$. That is $\mathbb{E}(e_{short}|A = a) \neq 0$ where $a = 0$ or $1$, unless $\mathbb{E}(X_1|A = a) = \mathbb{E}X_1$ and $\mathbb{E}(X_2|A = a) = \mathbb{E}X_2\ \forall a$.*

*Proof.* From Eq. 5 we have the following:

$$\begin{aligned}&\mathbb{E}(e_{\text{short}}|A = a)\\ &= \frac{\mathbb{E}\,X_1^2\,\mathbb{E}\,X_2 - \mathbb{E}\,X_1\,\mathbb{E}\,X_1X_2}{Var(X_1)}\beta_2 + \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2\,\mathbb{E}(X_1|A = a) - \beta_2\,\mathbb{E}(X_2|A = a)\\ &\neq 0,\end{aligned}$$

unless $\mathbb{E}(X_1|A = a) = \mathbb{E}X_1$ and $\mathbb{E}(X_2|A = a) = \mathbb{E}X_2, \forall a = 0, 1$.

□

**Proposition 5.** *When the model is mis-specified with omitted variables, the difference between the group-level mean prediction errors would not equal zero and the model is biased under our fairness notion, unless the group-specific mean of features satisfies a specific condition. That is, $\tau = \mathbb{E}(e_{short}|A = 1) - \mathbb{E}(e_{short}|A = 0) \neq 0$, unless $\frac{Cov(X_1,X_2)}{Var(X_1)}\left[\mathbb{E}(X_1|A = 1) - \mathbb{E}(X_1|A = 0)\right] = \left[\mathbb{E}(X_2|A = 1) - \mathbb{E}(X_2|A = 0)\right]$.*

*Proof.*

$$\begin{aligned}&\mathbb{E}(e_{\text{short}}|A = 1) - \mathbb{E}(e_{\text{short}}|A = 0)\\ &= \frac{Cov(X_1, X_2)}{Var(X_1)}\beta_2\left[\mathbb{E}(X_1|A = 1) - \mathbb{E}(X_1|A = 0)\right] - \beta_2\left[\mathbb{E}(X_2|A = 1) - \mathbb{E}(X_2|A = 0)\right]\\ &\neq 0,\end{aligned}$$

and the inequality would hold unless a specific condition holds. □

**Proposition 6.** $\mathbb{E}(e_{short}|A = 1)\Pr(A = 1) + \mathbb{E}(e_{short}|A = 0)\Pr(A = 0) = 0.$

*Proof.*

$$0 = \mathbb{E}(e_{\text{short}}) = \mathbb{E}(e_{\text{short}}|A = 1)\Pr(A = 1) + \mathbb{E}(e_{\text{short}}|A = 0)\Pr(A = 0) \tag{6}$$

□

**Corollary 7.** *When two groups have the same size, this would lead to the worst-case bias of $|\,\mathbb{E}(e_{short}|A = 1) - \mathbb{E}(e_{short}|A = 0)| = 2|\,\mathbb{E}(e_{short}|A = 1)|$.*

*Proof.* Since two groups have equal size, $\Pr(A = 1) = \Pr(A = 0)$. From Eq.6 we know that $\mathbb{E}(e_{\text{short}}|A = 1) + \mathbb{E}(e_{\text{short}}|A = 0) = 0$. Thus the statement is true. □

## Summary of results

- First, we prove that when a model is correctly specified, the mean prediction errors and bias are zero.
- We then quantify errors and biases when a model is mis-specified with omitted variables. In particular, we show that group-level mean prediction errors can be large even if population-level mean prediction error is small. Such errors are different across groups, leading to bias. The magnitude of such bias increases as the model misspecification worsens.