Fair learning and testing for unfairness given protected features

Insurance Data Science 2024

Tessa Steensgaard, University of Copenhagen (presenter)

Niklas Pfister, University of Copenhagen

Munir Eberhardt Hiabu, University of Copenhagen

Stockholm, June 2024

Set-up

Let $P \in \mathcal{P}$ be the observed distribution of (X, A, Y) from the set of distributions $\mathcal{P}$ where

$X \in \mathcal{X}$ is a set of risk factors,

$A \in \mathcal{A}$ is a set of protected features,

$Y \in \mathbb{R}$ is the response variable.

Set-up

Let $P \in \mathcal{P}$ be the observed distribution of (X, A, Y) from the set of distributions $\mathcal{P}$ where

$X \in \mathcal{X}$ is a set of risk factors,

$A \in \mathcal{A}$ is a set of protected features,

$Y \in \mathbb{R}$ is the response variable.

Consider the mapping $m : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and the stochastic error term $\varepsilon \in \mathbb{R}$. Assume that the response depends on the covariates via the structural assignment

$$Y = m(X, A) + \varepsilon \quad \text{and} \quad \mathbb{E}_P[\varepsilon \mid X, A] = 0.$$

Set-up

Let $P \in \mathcal{P}$ be the observed distribution of (X, A, Y) from the set of distributions $\mathcal{P}$ where

$X \in \mathcal{X}$ is a set of risk factors,

$A \in \mathcal{A}$ is a set of protected features,

$Y \in \mathbb{R}$ is the response variable.

Consider the mapping $m : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and the stochastic error term $\varepsilon \in \mathbb{R}$. Assume that the response depends on the covariates via the structural assignment

$$Y = m(X, A) + \varepsilon \quad \text{and} \quad \mathbb{E}_P[\varepsilon \mid X, A] = 0.$$

We identify m as

$$m(x, a) = m_0 + m_X(x) + m_A(a) + m_{X,A}(x, a),$$

with

$$m_0 = \mathbb{E}_P[Y],$$
$$m_X : x \mapsto \int m(x,a) p_A(a)\, \mathrm{d}a - m_0,$$
$$m_A : a \mapsto \int m(x,a) p_X(x)\, \mathrm{d}x - m_0,$$
$$m_{X,A} : (x,a) \mapsto m(x,a) - m_X(x) - m_A(a) - m_0.$$

Dependence removing shift

We can write the density of P as

$$p(x,a,y) = p_{Y|X,A}(y|x,a)p_{A|X}(a|x)p_X(x).$$

Dependence removing shift

We can write the density of P as

$$p(x,a,y) = p_{Y|X,A}(y|x,a)p_{A|X}(a|x)p_X(x).$$

Definition (Dependence removing shift)

Let the map $\tau : \mathcal{P} \to \mathcal{P}$ be the distributional shift such that the density of $\tau(P)$ satisfies

$$\tau(p)(x,a,y) = p_{Y|X,A}(y|x,a)p_A(a)p_X(x),$$

where $\tau(p)$ denotes the density of $\tau(P)$.

Unfairness

Definition (Unfair estimator)

We say that an estimator $g : \mathcal{X} \to \mathbb{R}$ is unfair if there exist a $\tilde{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that

$$\mathbb{E}_P\left[(Y - g(X))^2\right] \leq \mathbb{E}_P\left[(Y - \tilde{g}(X,A))^2\right], \text{ and} \qquad (P\text{: Observed})$$

$$\mathbb{E}_{\tau(P)}\left[(Y - g(X))^2\right] > \mathbb{E}_{\tau(P)}\left[(Y - \tilde{g}(X,A))^2\right]. \qquad (\tau(P)\text{: Dependence removing})$$

Unfairness

Definition (Unfair estimator)

We say that an estimator $g : \mathcal{X} \to \mathbb{R}$ is unfair if there exist a $\tilde{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that

$$\mathbb{E}_P\left[(Y - g(X))^2\right] \leq \mathbb{E}_P\left[(Y - \tilde{g}(X, A))^2\right], \text{ and} \qquad (P\text{: Observed})$$

$$\mathbb{E}_{\tau(P)}\left[(Y - g(X))^2\right] > \mathbb{E}_{\tau(P)}\left[(Y - \tilde{g}(X, A))^2\right]. \qquad (\tau(P)\text{: Dependence removing})$$

Proposition

We state the following results:

(i) The estimator $\mathbb{E}_P[Y \mid X = x]$ in the case $\operatorname{cov}(X, A) \neq 0$ is unfair.

(ii) The best not unfair estimator is

$$m_0 + m_X(X) = \int m(x, a) p_A(a)\, \mathrm{d}a.$$

In particular, if $m_X(X) = 0$ then the best not unfair estimator is a constant.

Proposing an estimator for $m_0 + m_X(X) = \int m(x,a) p_A(a)\,\mathrm{d}a$

Assumption

No interactions exist between X and A, i.e. $m_{X,A}(X,A) = 0$.

Proposing an estimator for $m_0 + m_X(X) = \int m(x,a) p_A(a)\,\mathrm{d}a$

Assumption

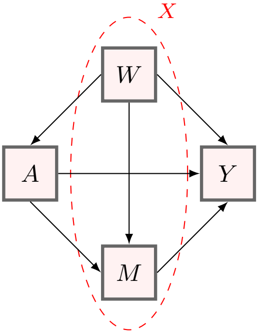
No interactions exist between X and A, i.e. $m_{X,A}(X,A) = 0$.

Assume that A is a one-dimensional binary random variable, i.e. $\mathcal{A} = \{0,1\}$. We will consider a partially linear model of the form

$$\begin{aligned} Y &= \theta A + m_X(X) + \varepsilon, \\ A &= w(X) + \varepsilon_A, \end{aligned} \quad \text{with} \quad \begin{aligned} \mathbb{E}[\varepsilon \mid A, X] &= 0, \\ \mathbb{E}[\varepsilon_A \mid X] &= 0. \end{aligned}$$

In this model, it holds that

$$\begin{aligned} \frac{\operatorname{cov}(Y - m_X(X), A - w(X))}{\operatorname{cov}(A, A - w(X))} &= \frac{\operatorname{cov}(\theta A + \varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta + \frac{\operatorname{cov}(\varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta. \end{aligned}$$



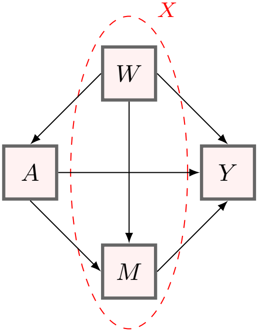
Proposing an estimator for $m_0 + m_X(X) = \int m(x,a) p_A(a)\,\mathrm{d}a$

Assumption

No interactions exist between X and A, i.e. $m_{X,A}(X,A) = 0$.

Assume that A is a one-dimensional binary random variable, i.e. $\mathcal{A} = \{0,1\}$. We will consider a partially linear model of the form

$$\begin{aligned} Y &= \theta A + m_X(X) + \varepsilon, \\ A &= w(X) + \varepsilon_A, \end{aligned} \quad \text{with} \quad \begin{aligned} \mathbb{E}[\varepsilon \mid A, X] &= 0, \\ \mathbb{E}[\varepsilon_A \mid X] &= 0. \end{aligned}$$



In this model, it holds that

$$\begin{aligned} \frac{\operatorname{cov}(Y - m_X(X), A - w(X))}{\operatorname{cov}(A, A - w(X))} &= \frac{\operatorname{cov}(\theta A + \varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta + \frac{\operatorname{cov}(\varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta. \end{aligned}$$

We obtain an estimator of θ by using the plug-in estimators of m_X and w.

Proposing an estimator for $m_0 + m_X(X) = \int m(x,a) p_A(a)\,\mathrm{d}a$

Assumption

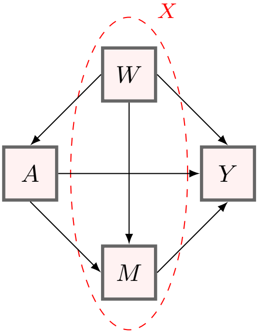
No interactions exist between X and A, i.e. $m_{X,A}(X,A) = 0$.

Assume that A is a one-dimensional binary random variable, i.e. $\mathcal{A} = \{0,1\}$. We will consider a partially linear model of the form

$$\begin{aligned} Y &= \theta A + m_X(X) + \varepsilon, \\ A &= w(X) + \varepsilon_A, \end{aligned} \quad \text{with} \quad \begin{aligned} \mathbb{E}[\varepsilon \mid A, X] &= 0, \\ \mathbb{E}[\varepsilon_A \mid X] &= 0. \end{aligned}$$

In this model, it holds that

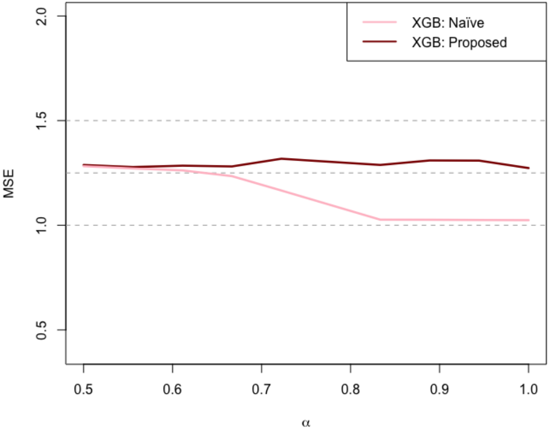
$$\begin{aligned} \frac{\operatorname{cov}(Y - m_X(X), A - w(X))}{\operatorname{cov}(A, A - w(X))} &= \frac{\operatorname{cov}(\theta A + \varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta + \frac{\operatorname{cov}(\varepsilon, \varepsilon_A)}{\operatorname{cov}(A, \varepsilon_A)} \\ &= \theta. \end{aligned}$$



We obtain an estimator of θ by using the plug-in estimators of m_X and w. Finally, we determine the estimator of m_X by training a machine learning model to $Y - \hat{\theta} A$ with features X.

Numerical Example

Naïve: $$\int \hat{m}(x,a)\hat{p}_A(a)\,\mathrm{d}a.$$



$$A = \text{Bern}(0.5),$$
$$X_1 = \alpha A + (1-\alpha)N(0,1),$$
$$X_2 = N(0,1),$$
$$Y = \sin(X_2) + A + N(0,1).$$

Unfairness Test

Theorem

Given an estimator $g : \mathcal{X} \to \mathbb{R}$ *if we can find a benchmark estimator* $\tilde{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *such that*

$$\mathbb{E}_P[(Y - \tilde{g}(X,A))^2] \leq \mathbb{E}_P[(Y - g(X))^2], \text{ and}$$

$$\operatorname{cov}_P(m_A(A), \tilde{g}(X,A)) < \operatorname{cov}_P(m_A(A), g(X)),$$

then g *is unfair according to Definition 2 under Assumption 1.*

Unfairness Test

Theorem

Given an estimator $g : \mathcal{X} \to \mathbb{R}$ *if we can find a benchmark estimator* $\tilde{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *such that*

$$\mathbb{E}_P[(Y - \tilde{g}(X,A))^2] \leq \mathbb{E}_P[(Y - g(X))^2], \text{ and}$$

$$\operatorname{cov}_P(m_A(A), \tilde{g}(X,A)) < \operatorname{cov}_P(m_A(A), g(X)),$$

then g *is unfair according to Definition 2 under Assumption 1.*

Theorem

Assume that $\hat{\theta}$ *is estimated on data different from* (X,A,Y)*. Then*

$$\sqrt{n}\left(\widehat{\operatorname{cov}}_P\left(\hat{\theta}A, \tilde{g}(X,A) - g(X)\right) - \operatorname{cov}_P(\theta A, \tilde{g}(X,A) - g(X))\right) \to \mathcal{N}\left(0, \sigma^2\right),$$

as $n \to \infty$, *where* $\sigma^2 = \theta^2 \operatorname{var}_P((A - \mathbb{E}_P[A])(g(X) - \mathbb{E}_P[g(X)]))$.

Unfairness Test

Theorem

Given an estimator $g : \mathcal{X} \to \mathbb{R}$ *if we can find a benchmark estimator* $\tilde{g} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *such that*

$$\mathbb{E}_P[(Y - \tilde{g}(X,A))^2] \leq \mathbb{E}_P[(Y - g(X))^2], \text{ and}$$

$$\operatorname{cov}_P(m_A(A), \tilde{g}(X,A)) < \operatorname{cov}_P(m_A(A), g(X)),$$

then g *is unfair according to Definition 2 under Assumption 1.*

Theorem

Assume that $\hat{\theta}$ *is estimated on data different from* (X, A, Y)*. Then*

$$\sqrt{n}\left(\widehat{\operatorname{cov}}_P\left(\hat{\theta}A, \tilde{g}(X,A) - g(X)\right) - \operatorname{cov}_P(\theta A, \tilde{g}(X,A) - g(X))\right) \to \mathcal{N}\left(0, \sigma^2\right),$$

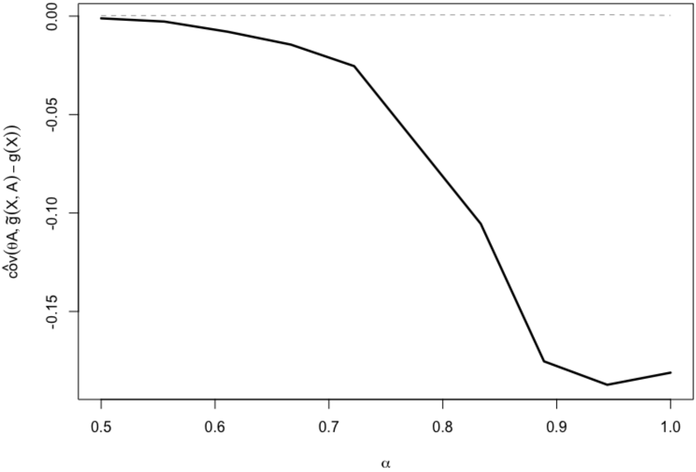
as $n \to \infty$*, where* $\sigma^2 = \theta^2 \operatorname{var}_P((A - \mathbb{E}_P[A])(g(X) - \mathbb{E}_P[g(X)]))$.

We can now construct a test for

$$H_0 : \quad \operatorname{cov}_P(\theta A, \tilde{g}(X,A) - g(X)) < 0.$$

Numerical Example

We choose $\gamma \in \mathbb{R}$ such that when we add $\gamma\hat{\theta}A$ to our proposed estimator it has the same performance under P as the plug-in estimator. Now, we consider the difference in the estimators' covariances with $\hat{\theta}A$.



Thank you for your attention.