FATE P03, P04

Algorithmic Fairness I

Ashwin S they/them

ashwin.singh01@estudiant.upf.edu

Universitat Pompeu Fabra Barcelona

Fairness, Accountability, Transparency and Ethics of Data Processing (FATE)

Programming Sessions

upf.

1. **6 Labs =** 3 Modules x 2 Sessions
2. **Grading Policy**

~~M1: Data Anonymization~~	~~**(35%)**~~	~~Submit by 11:59 PM, Feb 03, 2025~~
M2: Algorithmic Fairness I	**(30%)**	Submit by 11:59 PM, Feb 20, 2025
M3: Algorithmic Fairness II	**(35%)**	Submit by 11:59 PM, Mar 11, 2025

You must attend at least one session of each module to be eligible for grading.

3. **Queries / Discussion**

   Please post on the Aula Global Forum for everyone's benefit.

4. **Solved Practices:** To be available after each deadline.

Outline

In **Module I**, we will focus on **Regression**

1. “Fairness via Unawareness”
2. Omitted Variable Bias
3. Fairness Criteria for Regression
    a. Counterfactual Fairness
    b. Statistical (or Demographic) Parity

Datasets

1. Discrimination in Salaries
2. Law School Success

Discrimination in Salaries Data

upf.

Sex	Rank	degree	Years of Experience	Salary
male	full	doctorate	35	36350
male	associate	doctorate	19	24750
male	assistant	doctorate	23	19175
female	full	doctorate	27	26775
female	full	masters	32	24900
…	…	…	…	…

From a small midwestern college from United States in the early 1980s.

N = 52 Faculty

- 38 Male (73%)
- 14 Female (27%)

Underlying Biases

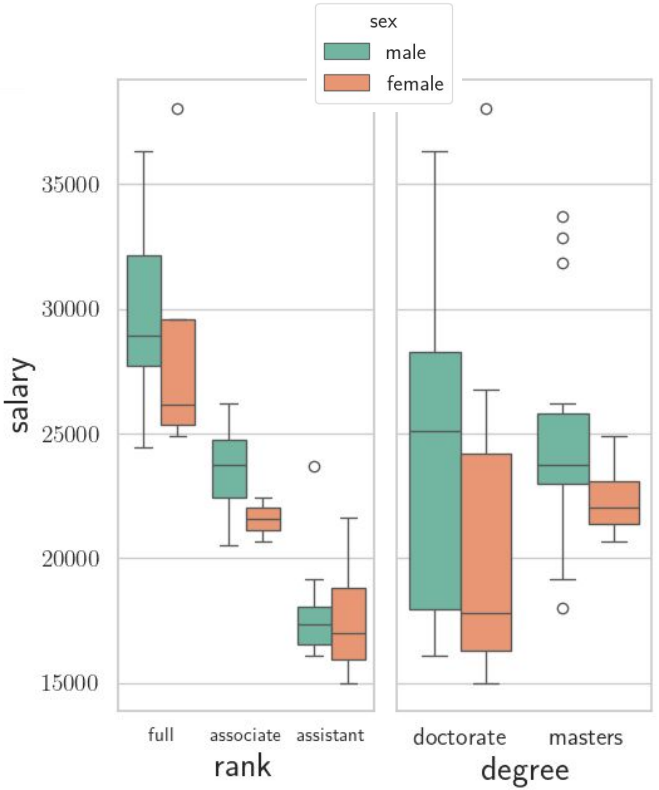
Faculty with Sex = Male

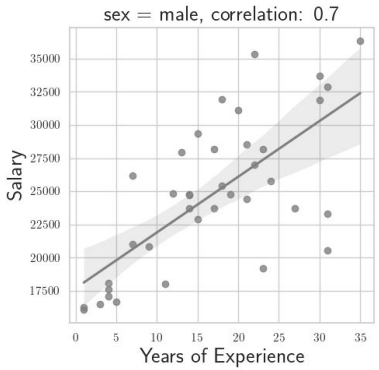
- 26% assistant, 32% associate, 42% full
- 37% masters, 63% doctorate

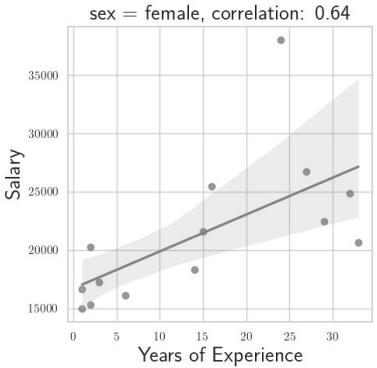
Faculty with Sex = Female

- 57% assistant, 14% associate, 29% full
- 29% masters, 71% doctorate

What can we observe?







A Test of Fairness via "Unawareness"

Linear Regression

$$\text{Salary } y = w^T x = w_0 + w_1 * \text{Rank} + w_2 * \text{Degree} + w_3 * \text{Experience} + w_4 * \text{Sex}$$

Rank: assistant = 1, associate = 2, full = 3

Degree: masters = 1, doctorate = 2

Sex: male = 0, female = 1

- **Aware Model** ▸ Trained on all Attributes
- **Blind Model** ▸ Trained on all Attributes except Sex
- **M Model** ▸ Trained on Data where Sex = Male
- **F Model** ▸ Trained on Data where Sex = Female

Žliobaitė, I., Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence & Law 24, 183–201 (2016). https://doi.org/10.1007/s10506-016-9182-5

A Test of Fairness via "Unawareness"

upf.

	Intercept	Rank	Degree	Experience	Sex
Aware Model	11558	4993	397	102	-950
Blind Model	11424	5230	179	87	
M Model	11736	5031	-30	129	
F Model	7718	4567	2398	115	

	RMSE (M)	RMSE (F)
Aware Model	2558	3454
Blind Model	2566	3525

Discuss

What can we infer from these results?

Omitted Variable Bias

upf.

True Model ▸ $y = a_0 + a_1x_1 + a_2x_2 + \beta s$

Fitted Model ▸ $y' = b_0 + b_1x_1 + b_2x_2$

where **s** : sensitive attribute and **$\beta \neq 0$**

b_0, b_1, b_2 will be **biased** due to the **omission** of s

when s is **correlated** with x_1 or x_2 or both…

Žliobaitė, I., Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence & Law 24, 183–201 (2016). https://doi.org/10.1007/s10506-016-9182-5

Law School Success Data

Race	Sex	UGPA	LSAT	ZFYA
White	Male	3.4	39	-0.5
Black	Female	3.2	27	-1.1
Asian	…	2.8	…	0.28
Hispanic	…	3.0	…	0.58
Other	…	2.4	…	…
…	…	…	…	…

N = 21,790 Students

From **163** Law Schools in the United States (1998).

Sex

56% Male,44% Female

Race

84% White, 6% Black, 4% Asian …

Fairness Criteria for Regression

upf.

Counterfactual Fairness $[\, Y \perp A \mid X \,]$

A predictor is **counterfactually fair** if it's decision towards an individual is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group.

Formally, Y is counterfactually fair if under any context $X{=}x, A{=}a$, Y satisfies:

$$\mathbb{E}\,[\, Y \mid X = x\,, A = a\,] = \mathbb{E}\,[\, Y \mid X = x\,, A = a'\,]$$

Discuss: **When is counterfactual fairness useful?**

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17).

Fairness Criteria for Regression

upf.

Statistical (or Demographic) Parity $[\ Y \perp A\ ]$

A predictor satisfies **statistical parity** if its predictions are independent of the sensitive attribute.

Formally, Y satisfies **statistical parity** if

$$\forall\ a \in A\ \ P[\ Y \geq z \,|\, A = a\ ] = P[\ Y \geq z\ ]$$

Discuss: When is statistical parity useful?

Agarwal, A., Dudik, M. & Wu, Z.S.. (2019). Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research.