Holistic AI × The Alan Turing Institute

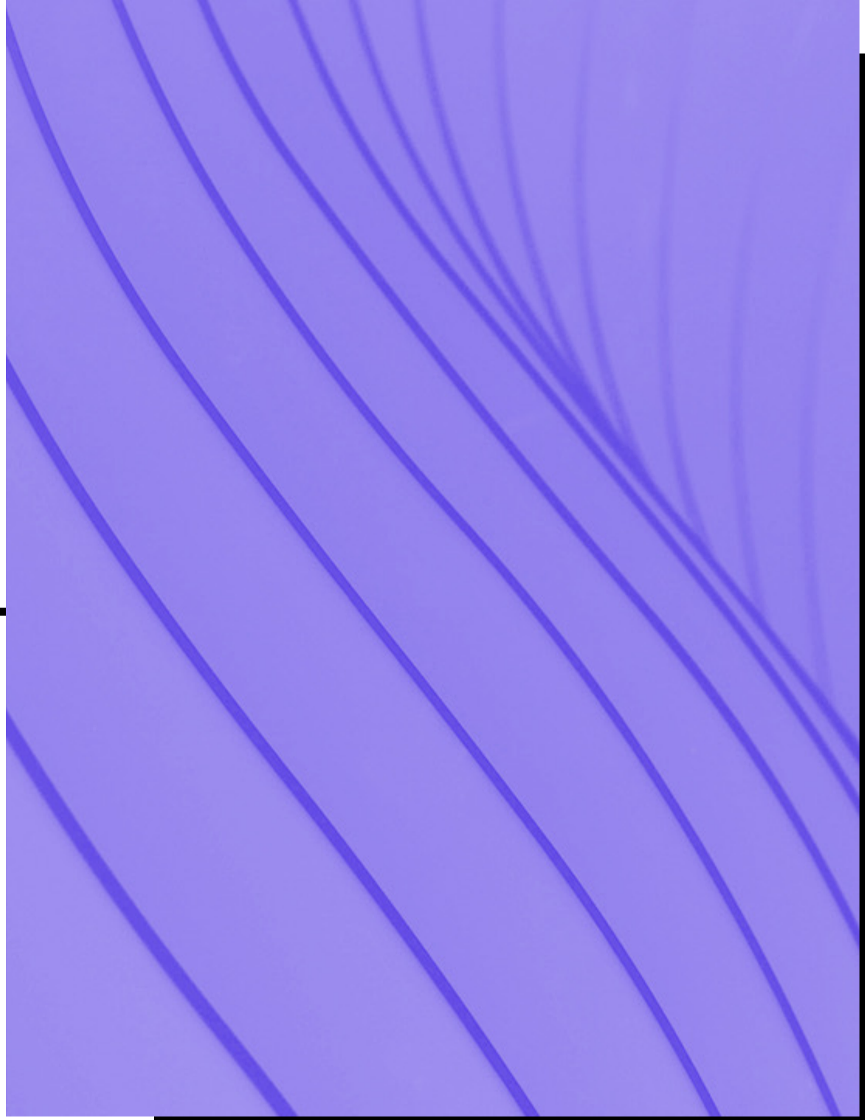
Bias in Regression Tasks – Part II

Content by: Sachin Beepath, Giulio Filippi, Nigel Kingsman, Cristian Munoz, Roseline Polle, Sara Zannone

Speaker: Sara Zannone

Contents

II – Fairness in Regression

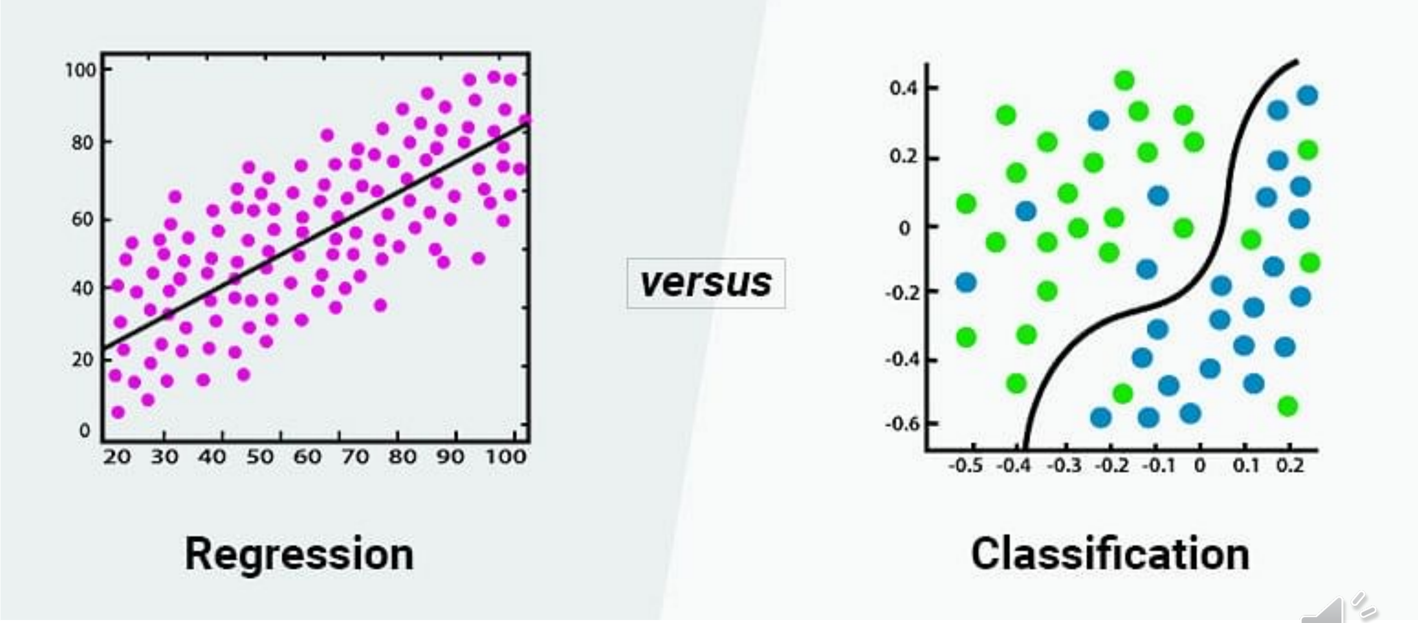


- 1) Individual vs Group Fairness
- 2) Equality of Outcome vs Equality of Opportunity

Fairness in Regression

- Fairness in regression tasks presents similar concepts to Fairness for binary classification (seen in our previous course)

- We will see in the following how we can extend these mathematical concepts from binary to continuous outputs



Individual vs Group Fairness

A first differentiation we can make is between individual and group fairness (Dwork et al. 2012)

- Individual fairness:

  Similar individuals should be treated similarly.

- Group Fairness:

  Different groups should be treated equally by the AI algorithm.

Individual vs Group Fairness

In Berk et al. 2017, we can find an example of operative definitions of individual and group fairness.

Individual vs Group Fairness

In Berk et al. 2017, we can find an example of operative definitions of individual and group fairness.

Problem setting:

In this work, fairness is achieved by minimizing the fairness cost.

Total loss function: $\ell(\boldsymbol{w}) + \lambda\, \ell_F(\boldsymbol{w}) + \gamma ||\boldsymbol{w}||_2$

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Individual fairness cost:</u>

$$\ell_{FI}(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in \boxed{S_i} \\ (\boldsymbol{x}_j, y_j) \in \boxed{S_j}}} d(y_i, y_j) \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \right)^2$$

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

Individual fairness cost:

$$\ell_{FI}(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d(y_i, y_j) \boxed{\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)^2}$$

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Individual fairness cost:</u>

$$\ell_{FI}(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} \boxed{d(y_i, y_j)} \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \right)^2$$

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Individual fairness cost:</u>

$$\ell_{FI}(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d(y_i, y_j) \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \right)^2$$

Individual Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Individual fairness cost:</u>

$$\ell_{FI}(\boldsymbol{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d(y_i, y_j) \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \right)^2$$

It measures how differently individual data points from S_i and S_j are treated (NO compensation).

Group Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Group fairness cost:</u>

$$\ell_{FG}(\boldsymbol{w}, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d\left(y_i, y_j\right) \left(f(\boldsymbol{x}_i) - f\left(\boldsymbol{x}_j\right)\right)\right)^{\mathbf{2}}$$

Group Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Group fairness cost:</u>

$$\ell_{FG}(\boldsymbol{w}, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d(y_i, y_j) \boxed{\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \right)} \right)^{\mathbf{2}}$$

Group Fairness

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Group fairness cost:</u>

$$\ell_{FG}(\boldsymbol{w}, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j}} d(y_i, y_j)\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^{\mathbf{2}}$$

It asks that data points from different groups have similar labels <u>on average.</u>
The model can compensate overvaluing a data point in S_i by overvaluing a data point in S_j

Example – Student’s grades

Let’s imagine that we have 3 students in total. We want to compare the grades of:

- 1 black student
- 2 white students.

For simplicity, we will assume that the model is a perfect predictor, so the true and predicted labels will be the same.

Grades for black students: $S_b = \{8\}$
Grades for white students: $S_w = \{6, 10\}$

Example

Grades for black students: $S_b = \{8\}$
Grades for white students: $S_w = \{6, 10\}$

Individual fairness: $\ell_{FI} = \frac{1}{1*2} \sum d(y_b, y_w)(y_b - y_w)^2$

where the distance: $d(y_b, y_w) = |y_b - y_w|$

$$\ell_{FI} = \frac{1}{2}(2 * 2^2 + 2 * (-2)^2) = 8$$

Example

Grades for black students: $S_b = \{8\}$
Grades for white students: $S_w = \{6, 10\}$

Group fairness: $\ell_{FG} = \left(\frac{1}{1*2}\sum d(y_b, y_w)(y_b - y_w)\right)^2$
where the distance: $d(y_b, y_w) = |y_b - y_w|$

$$\ell_{FG} = \left(\frac{1}{2}(2*2 + 2*(-2))\right)^2 = 0$$

Compensation: the model satisfies group fairness since the two groups have the same mean

Hybrid Fairness

- **Individual fairness penalty:**
  - Each cross pair $(\boldsymbol{x}_i, y_i) \in S_i, (\boldsymbol{x}_j, y_j) \in S_j$ is considered separately
- **Group fairness penalty:**
  - All cross pairs $(\boldsymbol{x}_i, y_i) \in S_i$, $(\boldsymbol{x}_j, y_j) \in S_j$ are considered together
- **Hybrid fairness penalty:**
  - Cross pairs $(\boldsymbol{x}_i, y_i) \in S_i, (\boldsymbol{x}_j, y_j) \in S_j$ with $y_i \leq \theta, y_j \leq \theta$ are considered together
  - Cross pairs $(\boldsymbol{x}_i, y_i) \in S_i$, $(\boldsymbol{x}_j, y_j) \in S_j$ with $y_i > \theta, y_j > \theta$ are considered together

Hybrid Fairness cost

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

Hybrid fairness cost:

$$\ell_{FH}(\boldsymbol{w}, S) = \left(\frac{1}{n_{1,1}n_{2,1}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i > \theta, y_j > \theta}} d(y_i, y_j)\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^{\mathbf{2}} + \left(\frac{1}{n_{1,0}n_{2,0}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i \leq \theta, y_j \leq \theta}} d(y_i, y_j)\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^{\mathbf{2}}$$

It asks that both cross pairs with labels above the threshold and cross pairs with labels below the threshold are treated similarly on average.

Compensation is possible only among datapoints whose labels are on the same side of the threshold.

Hybrid Fairness cost

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Hybrid fairness cost:</u>

$$\ell_{FH}(\boldsymbol{w}, S) = \left(\boxed{\frac{1}{n_{1,1}n_{2,1}}\sum_{\substack{(\boldsymbol{x}_i,y_i)\in S_i\\(\boldsymbol{x}_j,y_j)\in S_j\\ y_i>\theta,y_j>\theta}} d(y_i,y_j)\left(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right)}\right)^{\mathbf{2}} + \left(\boxed{\frac{1}{n_{1,0}n_{2,0}}\sum_{\substack{(\boldsymbol{x}_i,y_i)\in S_i\\(\boldsymbol{x}_j,y_j)\in S_j\\ y_i\le\theta,y_j\le\theta}} d(y_i,y_j)\left(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right)}\right)^{\mathbf{2}}$$

It asks that both cross pairs with labels above the threshold and cross pairs with labels below the threshold are treated similarly on average.

Hybrid Fairness cost

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Hybrid fairness cost:</u>

$$\ell_{FH}(\boldsymbol{w}, S) = \left(\frac{1}{n_{1,1}n_{2,1}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i > \theta, y_j > \theta}} d(y_i, y_j)\left(\boxed{f(\boldsymbol{x}_i)} - \boxed{f(\boldsymbol{x}_j)}\right)\right)^2 + \boxed{\left(\frac{1}{n_{1,0}n_{2,0}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i \le \theta, y_j \le \theta}} d(y_i, y_j)\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^2}$$

Compensation is possible only among datapoints whose labels are on the same side of the threshold.

Hybrid Fairness cost

The fairness cost $\ell_F(\boldsymbol{w})$ will take on different forms depending on the type of fairness.

<u>Hybrid fairness cost:</u>

$$\ell_{FH}(\boldsymbol{w}, S) = \left(\frac{1}{n_{1,1}n_{2,1}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i = y_j = 1}} d(y_i, y_j) \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^{\mathbf{2}} + \left(\frac{1}{n_{1,0}n_{2,0}} \sum_{\substack{(\boldsymbol{x}_i, y_i) \in S_i \\ (\boldsymbol{x}_j, y_j) \in S_j \\ y_i = y_j = 0}} d(y_i, y_j) \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)\right)^{\mathbf{2}}$$

In the following, we will focus on group fairness.

Equality of Opportunity vs Equality of Outcome

As in the case for binary classification, we split fairness notions into two main categories:

- Equality of Opportunity
- Equality of Outcome

Equality of Opportunity

Equality of Opportunity:

The accuracy of an AI system should be the same across all groups

Gender Shades: facial recognition algorithm shows disparity accuracy when it comes to gender and skin colour

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Equality of Opportunity vs Equality of Outcome

As in the case for binary classification, we split fairness notions into two main categories:

- Equality of Opportunity:
  The accuracy of an AI system should be the same across all groups
- Equality of Outcome :
  The distribution of the model's output should be similar across groups.

Equality of Opportunity

Bounded Group Loss (Agarwal et al. 2019):

- A predictor f satisfies bounded group loss at level η if the average loss is smaller than η for all protected attributes:

- In equations:

$$\min_{f \in F} \mathbb{E}[\ell(y, f(\boldsymbol{x}))] \quad \text{such that } \forall a \in A$$
$$\mathbb{E}[\ell(y, f(\boldsymbol{x}))|A = a] \leq \eta$$

Equality of Outcome

Statistical Parity (Agarwal et al. 2019):

- A predictor f satisfies statistical parity if *f(X)* is independent of the protected attribute *A*

- In equations:

$$\min_{f \in F} \mathbb{E}[\ell(y, f(\boldsymbol{x}))] \quad \text{such that } \forall a \in A, z \in [0,1]$$
$$|P[f(\boldsymbol{x}) \geq z | A = a] - P[f(\boldsymbol{x}) \geq z]| \leq \epsilon$$

Binarization

- Statistical Parity:

$$\min_{f \in F} \mathbb{E}[\ell(y, f(\boldsymbol{x}))] \quad \text{such that } \forall a \in A, z \in [0,1]$$
$$|P[f(\boldsymbol{x}) \geq z | A = a] - P[f(\boldsymbol{x}) \geq z]| \leq \epsilon$$

- Binarization is needed to extend SP from classification to regression

- $P[f(\boldsymbol{x}) \geq z | A = a]$ can be seen as the Success Rate for a specific group when the value z is taken as threshold
- Other forms of binarization are possible, as we will see in the next lecture

Conclusion

- Individual and Group
- Hybrid Fairness

- Equality of Opportunity: Bounded Group Loss
- Equality of Outcome: Statistical Parity

Contents

- Part I – Introduction to Regression
- Part II – Fairness in Regression
- **Part III – Measuring Bias in Regression**
- Part IV – Mitigating Bias in Regression

References

[1] Berk et al, 2017, A Convex Framework for Fair Regression (https://arxiv.org/pdf/1706.02409.pdf)

[2] Dwork et al. 2012, Fairness Through Awareness (https://arxiv.org/abs/1104.3913)