Fair Machine Learning

Achieving Fairness with a Simple Ridge Penalty

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

June 29, 2023

→ Introduction

Fair Linear Models

Fair Ridge Regression

An Example: Drug Consumption

Conclusions and Acknowledgements

The Problem

- Machine learning (statistical?) models are being used in applications where it is crucial to ensure the accountability and fairness of the decisions made on the basis of their outputs.

- Models are trained on historical data that contain various forms of bias, capture those biases and carry them over into current applications resulting in unfair discrimination of certain groups of people.

- The concept of fairness itself is difficult to define because it depends on the type of distortion we wish to limit and on how we characterise it mathematically.

- How can we specify fair models that capture the non-discriminating information present in the data and disregard the discriminating information?

Say that $\mathbf{y}$ is our response, $\hat{\mathbf{y}}$ are fitted values from the model, $\mathbf{S}$ are the sensitive attributes containing the discriminating information and $\mathbf{X}$ are the other predictors.

- Group fairness: predictions should be similar across the groups identified by the sensitive attributes.
  - Statistical or demographic parity ($\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{S}$).
  - Equality of opportunity ($\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{y}$).
- Individual fairness: individuals that are similar receive similar predictions

$$f(\mathbf{y}, \mathbf{S}) = \sum_{i,j} d_1(y_i, y_j) d_2(\mathbf{s}_i, \mathbf{s}_j).$$

Many, many mathematical characterisations in the literature [9, 3, 10].

Algorithmic Fairness: Different Stages

We can enforce fairness at different stages of the model selection, estimation and validation process, and for different classes of models [2]:

- Pre-processing approaches that try to transform the data to remove the underlying discrimination so that any model fitted on the transformed data is guaranteed to be fair.
- In-processing approaches that modify the model estimation process in order to remove discrimination, either by changing its objective function (typically the log-likelihood) or by imposing constraints on its parameters.
- Post-processing approaches that use a hold-out set to assess a previously-estimated model (treated as a black box) and that alter its predictions to make them fair.

In-processing approaches for black-box machine learning models such as deep neural networks fall within the realm of Explainable AI [1].

- Introduction
- Fair Linear Models
- Fair Ridge Regression
- An Example: Drug Consumption
- Conclusions and Acknowledgements

Linear Regression Models, Take I

Consider a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We can do what Zafar *et al.* [14] did:

$$\min_{\boldsymbol{\beta}} \mathrm{E}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2\right] \quad \text{such that} \quad |\,\mathrm{COV}(\mathbf{X}\boldsymbol{\beta}, S_i)| < c, c \in \mathbb{R}^+.$$

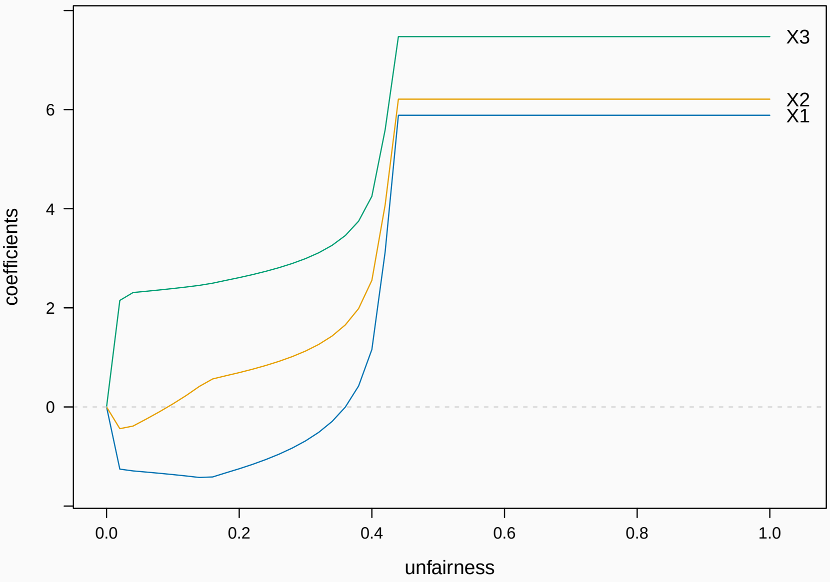
PROS:

- It's simple.
- It uses a linear measure of dependence to bound the effect of all S_i in $\mathbf{S}$ on $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$, which agrees with the loss function.

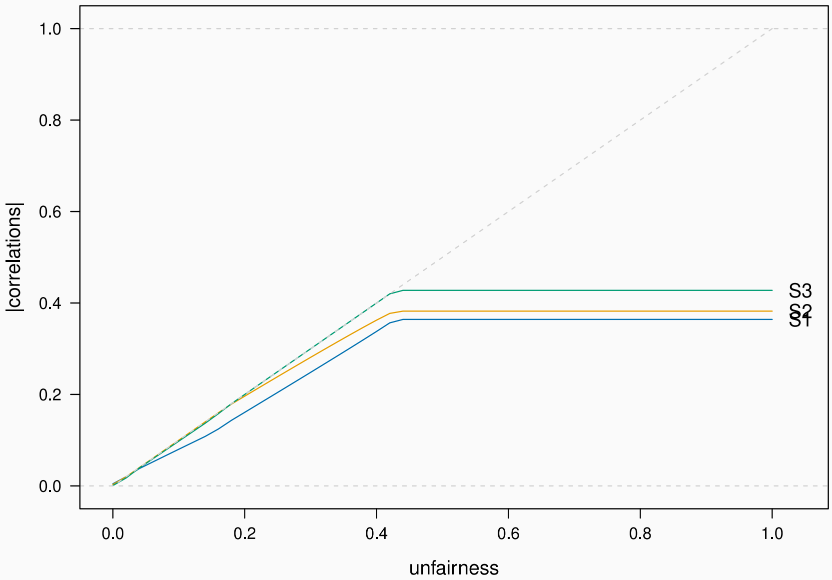
CONS:

- No distributional assumptions, so no hypothesis testing, confidence intervals, etc.
- As $c \to 0$ to enforce fairness, $\boldsymbol{\beta} \to \mathbf{0}$ and the non-discriminating information in $\mathbf{X}$ is removed along with the discriminating information.

Coefficient Profile Plots in Zafar *et al.*



Constraints in Zafar *et al.*



Komiyama *et al.* [7] did:

1. remove the association between $\mathbf{X}$ and $\mathbf{S}$ with $\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{S} + \mathbf{U}$, estimating $\widehat{\mathbf{B}}_{\mathrm{OLS}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X}$;
2. take the decorrelated predictors $\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{B}}_{\mathrm{OLS}}^{\mathrm{T}}\mathbf{S}$ which contain the component of $\mathbf{X}$ that cannot be explained by $\mathbf{S}$ ($\widehat{\mathbf{U}} \perp\!\!\!\perp \mathbf{S}$);
3. formulate the regression model $\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$;
4. formulate the fairness constraint

$$R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha})}{\mathrm{VAR}(\hat{\mathbf{y}})} = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta}};$$

5. solve the optimisation problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathrm{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] \quad \text{such that} \quad R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r, r \in [0, 1].$$

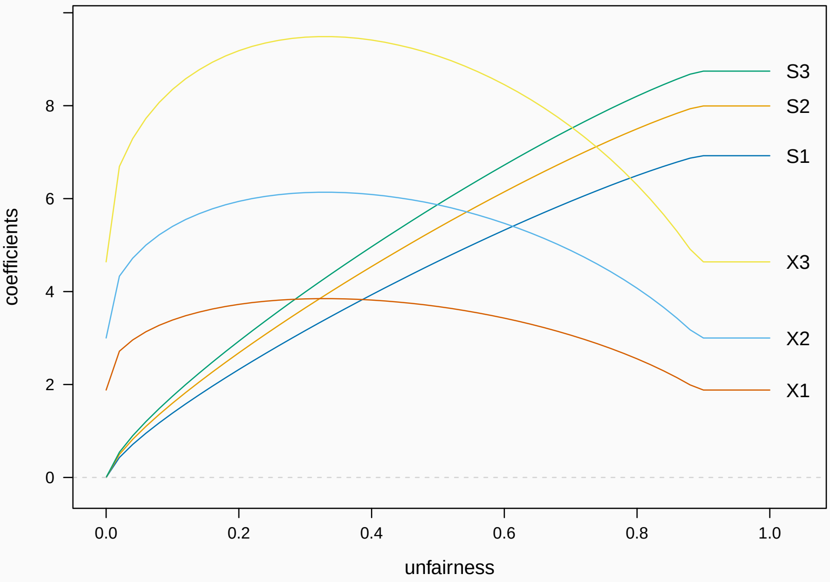
PROS:

- The formulation is simple.
- Discriminating and non-discriminating information are separated.
- The optimisation problem is quadratic-constraints quadratic programming, for which there are solvers.
- The fairness constraint is defined in terms of explained variance, the natural measure of information in a linear model.
- The bound is interpretable (0 is complete fairness, 1 is no constraint).

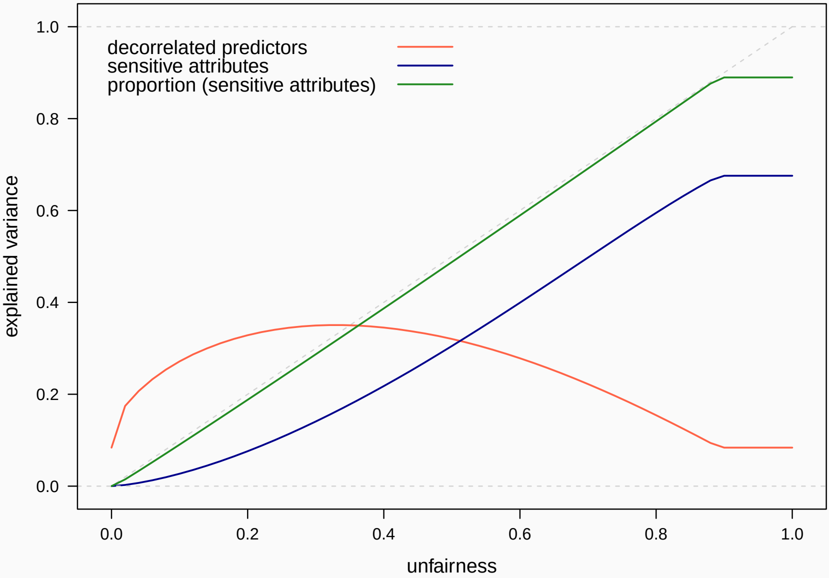
CONS:

- No distributional assumptions.
- The optimisation problem cannot be extended (or even tweaked) without losing the ability to use quadratic-constraints quadratic solvers.
- The behaviour of the estimated coefficients is weird.

Coefficient Profile Plots in Komiyama *et al.*



Constraints in Komiyama *et al.*



Introduction

Fair Linear Models

Fair Ridge Regression

An Example: Drug Consumption

Conclusions and Acknowledgements

How Can We Do Better?

Take two vintage pieces of statistics from the 1970s-1980s:

1. ridge regression [6];
2. generalised linear models [8].

We can use them (and nothing else) to fix the few CONS of the fair model from Komiyama *et al.* and keep all the PROS.

We call this approach the Fair (Generalised) Ridge Regression Model (F(G)RRM). Its selling points are:

- Modular: swappable characterisation of fairness.
- Versatile: supports all generalised linear models.
- Interpretable: both the model and the fairness constraints are interpretable and all the best practices from the literature apply.
- Statistical: model selection, model validation, hypothesis testing, confidence intervals, etc. are already available in the literature.

Fair Ridge Regression Model (FRRM)

Let's start again from $\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We want to re-create the shrinkage effects on the coefficients $\boldsymbol{\alpha}$ associated with $\mathbf{S}$ that we see in Komiyama *et al.*: we can do that with a ridge penalty,

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \widehat{\mathbf{U}}\boldsymbol{\beta}\|_2^2 + \lambda(r)\|\boldsymbol{\alpha}\|_2^2,$$

which we only apply to $\boldsymbol{\alpha}$ because by construction there is no discriminating information in $\widehat{\mathbf{U}}$. The parameter estimates are in closed form:

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} \end{bmatrix} = \begin{bmatrix} \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda(r)\mathbf{I}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y} \\ (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \end{bmatrix}.$$

But how do we control the fairness of the model?

Fair Ridge Regression Model (FRRM)

For a given level of fairness $r \in [0, 1]$:

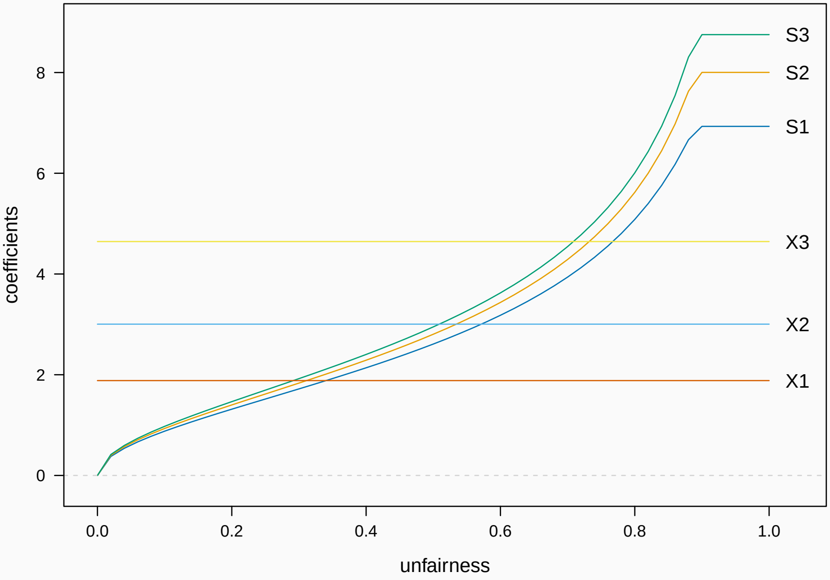
1. Compute $\widehat{\mathbf{U}}$ from $\mathbf{X}, \mathbf{S}$.
2. Estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} = (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y}$.
3. Estimate $\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y}$. Then:
   3.1 If $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \leqslant r$, set $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = \widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$.
   3.2 Otherwise, find the value of $\lambda(r)$ that satisfies

$$\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} = \frac{r}{1-r}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$$

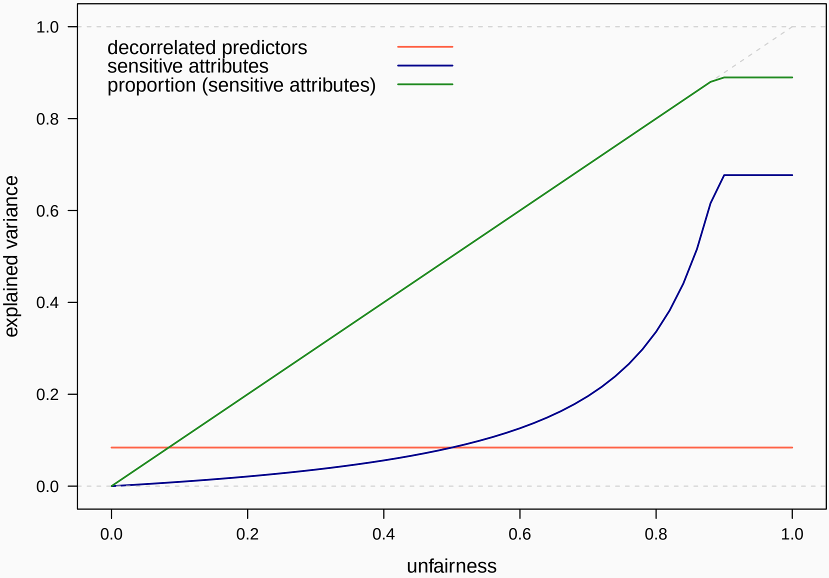
and estimate the associated $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ in the process.

This approach is guaranteed to have a single solution which can be found with a simple univariate root finding algorithm regardless of the number of variables involved.

Coefficients in FRRM



Constraints in FRRM



Different Characterisations of Fairness in FRRM

Furthermore, $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ depend on the fairness constraint only through $\lambda(r)$ so we can easily replace $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (which enforces statistical parity) with other constraints.

1. Equality of opportunity:

$$R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\phi})}{\mathrm{VAR}(\mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi})}$$

   where $\boldsymbol{\phi}, \psi$ are the coefficients of $\hat{\mathbf{y}} = \mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi} + \boldsymbol{\varepsilon}^*$.
2. Individual fairness:

$$D_{\mathrm{IF}} = \frac{f(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \mathbf{y}, \mathbf{S})}{f(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \mathbf{y}, \mathbf{S})}, f(\boldsymbol{\alpha}, \mathbf{y}, \mathbf{S}) = \sum_{i,j} d(y_i, y_j)(\mathbf{s}_i\boldsymbol{\alpha} - \mathbf{s}_j\boldsymbol{\alpha})^2$$

3. Any Convex combination of $R^2_{\mathbf{S}}(\cdot)$, $R^2_{\mathrm{EO}}(\cdot)$, $D_{\mathrm{IF}}(\cdot)$ and others.

Different Families of Models: FGRRM

Starting from the general formulation of a generalised linear model

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}, \qquad \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}), \qquad \boldsymbol{\eta} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta},$$

where $g(\cdot)$ is the link function, we can draw on [5, 12, 13] to estimate

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} D(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda(r)\|\boldsymbol{\alpha}\|_2^2.$$

where $D(\cdot)$ is the deviance of the model.
The ridge penalty $\lambda(r)$ can then be estimated to give

$$\frac{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \boldsymbol{\beta})}{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \mathbf{0})} \leqslant r.$$

For Gaussian GLMs we obtain FRRM again, but we can also work with Binomial GLMs, Poisson GLMs, Multinomial GLMs and Cox proportional hazards models.

- ✓ Introduction
- ✓ Fair Linear Models
- ✓ Fair Ridge Regression
- → An Example: Drug Consumption
- Conclusions and Acknowledgements

Drug Consumption Survey from Fehrman *et al.* [4]

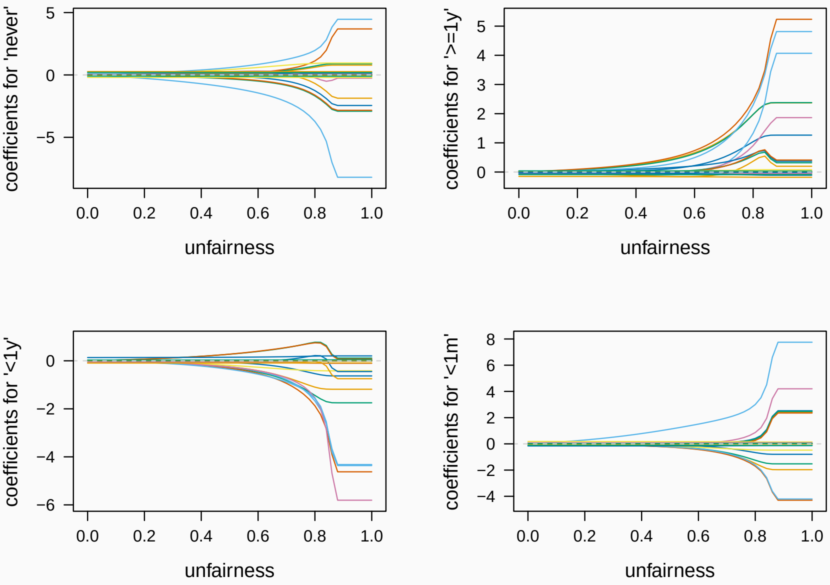
The data:

- 18 different drugs measured as `"Never Used"`, `"Used over a Decade Ago"`, `"Used in Last Decade"`, `"Used in Last Year"`, `"Used in Last Month"`, `"Used in Last Week"`, `"Used in Last Day"`.
- Impulsivity (`Impulsivity`), sensation seeking (`SS`).
- Personality traits: neuroticism (`Nscore`), extroversion (`Escore`), openness to experience (`Oscore`), agreeableness (`Ascore`) and conscientiousness (`Cscore`).
- Age, gender, race, education level.

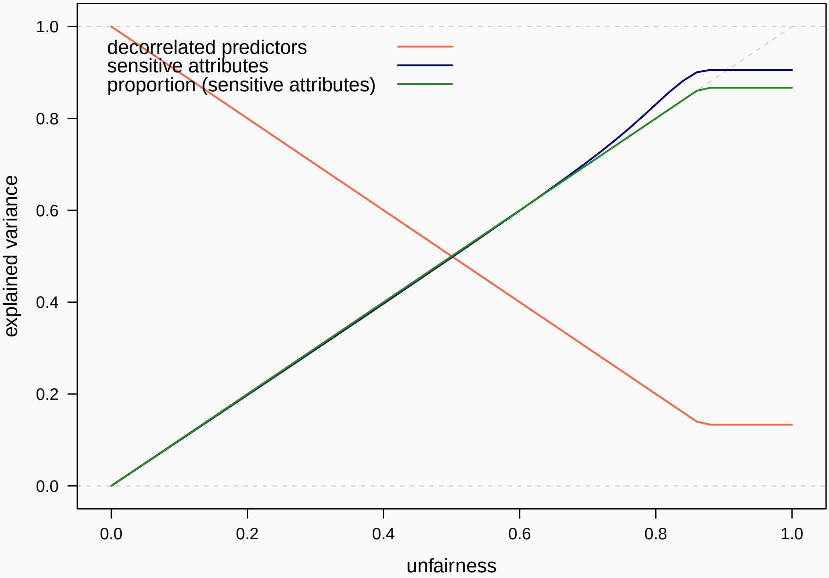
The model, a multinomial FGRRM:

- Response: LSD use.
- Sensitive attributes: age, gender, race.
- Predictors: education level, personality traits, impulsivity, sensation seeking.

Drugs Consumption: Coefficients



Drugs Consumption: Coefficients



- ✓ Introduction
- ✓ Fair Linear Models
- ✓ Fair Ridge Regression
- ✓ An Example: Drug Consumption
- → Conclusions and Acknowledgements

Conclusions

- Fairness is increasingly a concern as machine learning models become an integral part of automated decision support systems.
- Explainable AI investigates the explainability and fairness of black-box models such as deep neural networks, but simpler models are also in common use and should be made to be fair.
- The literature, by and large, studies fairness as an optimisation problem and produces models whose statistical properties and best practices are unknown.
- Classical statistics provides all the tools to formulate versatile fair models that we know how to interpret and to use.

Francesca Panero
Assistant Professor, Department of Statistics
London School of Economics

Manuel Proissl
Quantum Industry Applications Lead
IBM

This material has been published in [11]:

M. Scutari, F. Panero and M. Proissl (2022). "Achieving Fairness with a Simple Ridge Penalty." *Statistics and Computing*, 32, 77.
https://doi.org/10.1007/s11222-022-10143-w

Software: https://cran.r-project.org/web/packages/fairml/

Thanks!

Any questions?

References I

- A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al.
  Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI.
  Information Fusion, 58:82–115, 2020.
- B. D'Alessandro, C. O'Neil, and T. LaGatta.
  Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification.
  Big Data, 5(2):120–134, 2017.
- E. Del Barrio, P. Gordaliza, and J. M. Loubes.
  Review of Mathematical Frameworks for Fairness in Machine Learning, 2020.
- E. Fehrman, A. K. Muhammad, E. M. Mirkes, et al.
  The Five Factor Model of Personality and Evaluation of Drug Consumption Risk.
  In *Data Science*, pages 231–242, 2017.
- J. Friedman, T. Hastie, and R. Tibshirani.
  Regularization Paths for Generalized Linear Models via Coordinate Descent.
  Journal of Statistical Software, 33(1):1–22, 2010.
- A. E. Hoerl and R. W. Kennard.
  Ridge Regression: Biased Estimation for Nonorthogonal Problems.
  Technometrics, 12(1):55–67, 1970.

References II

- J. Komiyama, A. Takeda, J. Honda, and H. Shimao.
  Nonconvex Optimization for Regression with Fairness Constraints.
  Proceedings of Machine Learning Research, 80:2737–2746, 2018.
  35th International Conference on Machine Learning.
- P. McCullagh and J. A. Nelder.
  Generalized Linear Models.
  CRC press, 2nd edition, 1989.
- N. Mehrabi, F. Morstatter, N. Saxena, et al.
  A Survey on Bias and Fairness in Machine Learning.
  ACM Computing Surveys, 54(6):115, 2021.
- D. Pessach and E. Shmueli.
  A Review on Fairness in Machine Learning.
  ACM Computing Surveys, 55(3):51, 2022.
- M. Scutari, F. Panero, and M. Proissl.
  Achieving Fairness with a Simple Ridge Penalty.
  Statistics and Computing, 32:77, 2022.
- N. Simon, J. Friedman, T. Hastie, and R. Tibshirani.
  Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.
  Journal of Statistical Software, 39(5):1–13, 2011.

References III

J. K. Tay, B. Narasimhan, and T. Hastie.
Elastic Net Regularization Paths for All Generalized Linear Models.
Journal of Statistical Software, 106(1):1–31, 2023.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi.
Fairness Constraints: a Flexible Approach for Fair Classification.
Journal of Machine Learning Research, 20:1–42, 2019.