Data Science Looks At Discrimination (R Package)

Taha Abdullah, Arjun Ashok, Shubhada Martha, Aditya Mittal, Billy Ouattara, Jonathan Tran

University of California, Davis, 95616

Introduction

The DSLD package provides statistical and graphical tools for non-statisticians and statisticians alike to detect, measure, and mitigate discrimination in real-world applications with ease.

- **Estimation**: Estimate the impact of a sensitive feature [S] on an outcome feature [Y] while accounting for potential confounders [C]
- **Prediction**: Eliminate the use of [S] in modeling while regulating the use of the proxies [O] to mitigate biased predictions

Implemented Functions

- **DsldLinear**: Comparison of conditions for sensitive groups via linear models, with and without interactions
- **DsldQeFairML**: ML algorithms such as K-Nearest Neighbors, Random Forests, Ridge Regression with explicitly deweighted features
- **DsldConfounders**: Assess possible confounding variables between a sensitive feature and the other features
- **DsldConditDisparity**: Plots [Y] against [X] with custom restrictions to extract underlying patterns with respect to different sensitive groups
- **DsldCHunting/DsldOHunting**: Confounder hunting searches for features [C] that predict both [Y] and [S], and proxy hunting searches for features [O] that predict [S]
- **FairML Wrappers**: Wrappers for FairML package including functions nclm, frrm/fgrrm, zlm
- **Python Analogs**: Python Wrappers are also available for the majority of functions
- **Installation**: Installation via https://github.com/matloff/dsld. Supplementary *Quarto Book* is also available for additional information for users.

Adjusting for Confounders

Investigating a possible gender pay gap using **svcensus** data. [Y] is wage and [S] is gender. We will treat age as a confounder [C] using a linear model

No Interactions

- Mean(W) $= \beta_0 + \beta_1 A + \beta_2 M$
- W is wage; A is age; M is an indicator feature (M = 1 for men and M = 0 for women)
- Estimate of β_2 turns out to be about 13,000, which is the (estimated) wage gap
- 95 percent Confidence interval: 13098.2091 +- 1.96 x 790.4451

Interactions

- Gender gap may be small at younger ages but much larger for older people
- Fit two linear models, one for men and one for women
- Gender pay gap is estimated to be -12753.65 at age 18, and -13459.30 at age 60. We can see that income difference by gender vary based on age

Linearity Assumptions

Graphical approach via the DSLD package may be quite informative

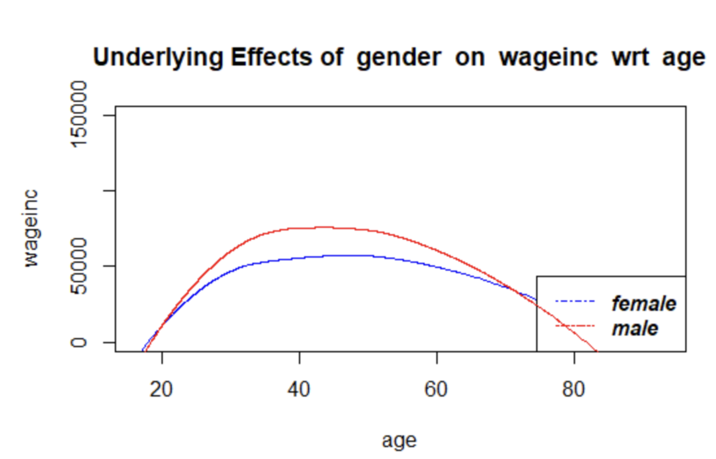

Fig. 1:Effect of Age by Race on Income

Relation looks nonlinear, possibly reflecting age discrimination against both very young and very old workers

Is the LSAT Fair?

- Concerns that the LSAT and other similar tests are biased against Black and Latino students, and might otherwise have racial issues
- Concerning **racial differences**: Two very similar people (same quality law school, undergraduate/law school grades, bar passage status) will have LSAT scores differing on average by **almost 6 points** if one person is Black and the other is white.

Exploratory Data Analysis

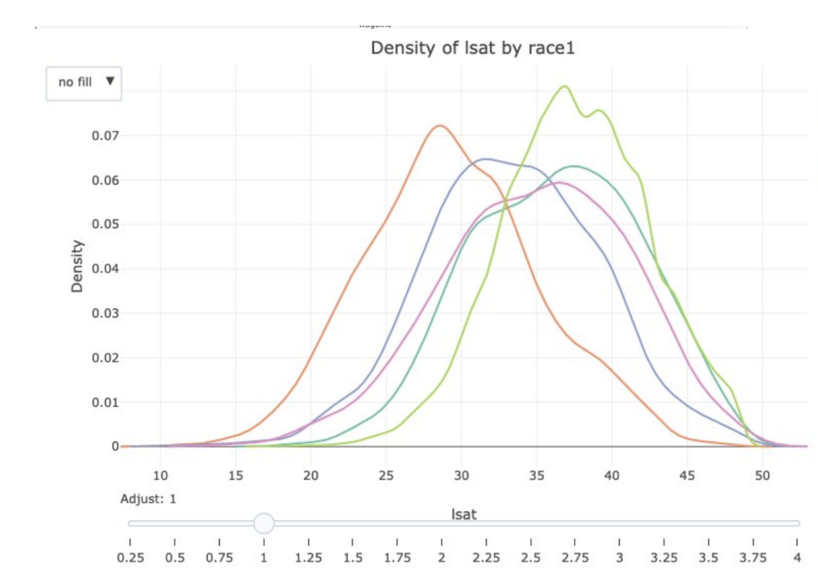

Fig. 2:Distribution of LSAT Scores by Race

- Distribution of LSAT scores for white students appears to be higher than others, particularly compared to black students

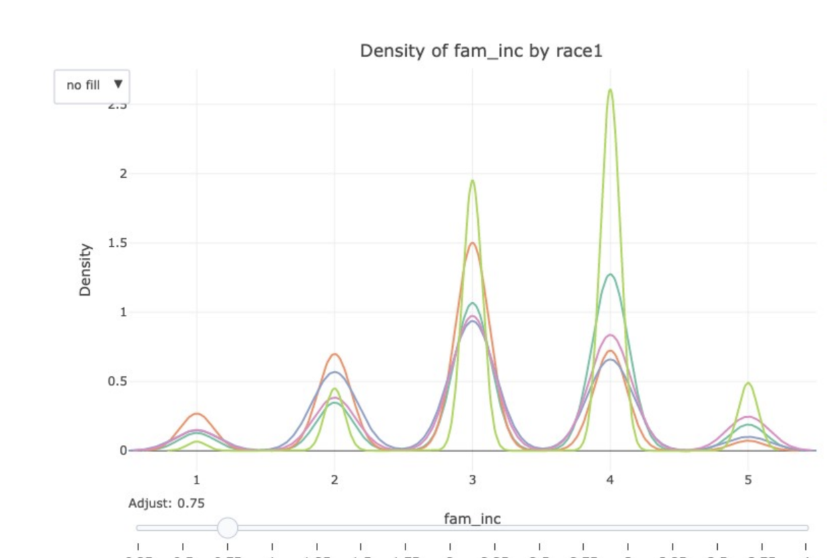

Fig. 3:Distribution of Family Income by Race

- White students tend to fall under higher family income group as opposed to other races

Mitigating Bias for FairML

- **Goal**: Predict [Y] from [X] and [O], omitting [S]
- Concern that we may be indirectly using [S] via [O]. We want to limit the usage of proxies.
- [O] is related to [S]; the stronger the relation, the less weight we will put on that feature in predicting [Y]
- The inherent tradeoff of **increasing fairness** is **reduced utility** (reduced predictive power)

Measuring Utility

- Measuring effectiveness or value of a model in making accurate predictions or decisions
- Mean Squared Error for continuous [Y]
  Misclassification rate for binary [Y]

Measuring Fairness

- Measuring algorithmic discrimination empirically
- Correlation between predicted [Y], to be denoted $[\hat{Y}]$, and [S]

Comparing Empirical Results

- Compare base K-Nearest Neighbors (qeKNN) with dsldQeFairKNN
- Proxy [O] "occupation" will be deweighted to 0.2 to limit its effect

Fairness/Utility Tradeoff	Fairness	Utility
qeKNN	0.1943313	25452.08
dsldQeFairKNN	0.0814919	26291.38

Table 1:Fairness/Utility Results across KNN Models

- $\rho(\hat{Y}, \text{S})$ decreased significantly. Test Accuracy increased by about 700 dollars
- We see an increase in fairness at the cost of utility