HRP 203 Report

Introduction

There is a sizeable and growing literature on fair regression approaches. Some approaches involve modifications on input data (Feldman et al., n.d.; Gordaliza et al. 2019; Hu, Ratz, and Charpentier 2024) but many are focused on modifying the objective or loss function of predictive models in a way that values some notion of fairness. These loss functions commonly include a traditional error term, such as the mean squared error, as well as either a penalty or constraint based on some mathematical notion of unfairness, such as a requirement of statistical parity between two groups (Calders et al. 2013; Berk et al., n.d.). Some subsequent work has focused on technical methods for solving constrained or penalized prediction problems (Dwork et al., n.d.; Chzhen et al. 2020; Lohaus, Perrot, and Luxburg 2020; Mishler and Kennedy, n.d.), and some has focused on tailoring and applying these methods to important problem settings like cyberbullying (Gencoglu 2021) and health care and health policy (Zink and Rose 2020).

For this class project, I will be implementing a simplified version of these methods on simulated data in order to demonstrate the value of fair learning techniques. Specifically, I will impose penalties on the under-prediction of spending for non-female individuals, whose spending is under-predicted in the unpenalized model. I will show that this reduces under-prediction with modest but meaningful decreases in accuracy.

Methods

I will be using the HRP 203 simulated dataset for this project. This dataset has 5000 observations of five variables: `smoke`, `female`, `age`, `cardiac`, and `cost`. There is no available data dictionary on information about the data generating process. I will be focusing on predicting `cost` using the other variables except for `female`. In this exercise, we are treating `female` as a protected class that we will not include as a control but do not want to be correlated with our predictions, as we do not want to simply shift predictions up or down on the basis of this variable.

First, I will create a simple linear regression to predict spending using all other available variables except for `female`. Let there be N individuals in the dataset. For individual i, I will note cost as Y_i, an indicator for being female with F_i, and a vector of covariates Z_i (which includes a 1 for the intercept term, `smoke`, `age`, and `cardiac`. I will fit a vector of regression coefficients β according to the equation

$$Y_i = \beta Z_i + \epsilon_i$$

Next, I will identify under-prediction, as measured by the mean residuals, with respect to sex (or gender - there is no data dictionary available for the synthetic data). Let $\hat{\beta}$ be the fitted vector of all regression coefficients. For individuals whose sex is female, the mean residual will be calculated as

$$MR_F = \frac{\sum_{i=1}^{N} \left(Z_i \hat{\beta} - Y_i \right) F_i}{\sum_{i=1}^{N} F_i}$$

If this value is less than 0, this indicates that spending is under-predicted by the simple regression for individuals whose sex is female. Because the mean residual across all training observations is 0, if the mean residual for individuals whose sex is female is less than 0, then the mean residual for individuals whose sex is not female will be greater than 0 and vice versa. For the sex category that is under-predicted (with a mean residual less than 0), I will impose a penalty on under-prediction in the model fitting process. This means that in order to identify the optimal coefficient vector $\hat{\beta}$, instead of minimizing the traditional linear regression loss function of the mean squared error

$$L = \frac{1}{n} \sum_{i=1}^{n} (Y_i - Z_i \beta)^2$$

with respect to β, I will minimize a loss function that also includes a constant λ multiple of the mean residual for individuals in the under-predicted sex category (who I will note as being in group g, which has N_g individuals in it).

$$L = \frac{1}{n} \sum_{i=1}^{N} (Y_i - Z_i \beta)^2 + \frac{\lambda}{N_g} \sum_{j \in g} (Y_j - Z_j \beta)$$

The derivative of this loss function with respect to β is

$$\frac{\partial L}{\partial \beta} = -\frac{2}{N} \sum_{i=1}^{n} (Y_i Z_i' - Z_i' Z_i \beta) - \frac{\lambda}{N_g} \sum_{j \in g} Z_j'$$

I can then set this derivative equal to 0 and solve for the closed form solution for the optimal β for this penalized regression.

$$\frac{2}{n}\sum_{i=1}^{N} Z_i' Z_i \beta = \frac{2}{N}\sum_{i=1}^{n} Z_i' Y_i + \frac{\lambda}{N_g}\sum_{j\in g} Z_j'$$

$$\hat{\beta} = \left(\frac{2}{N}\sum_{i=1}^{N} Z_i' Z_i\right)^{-1} \left(\frac{2}{N}\sum_{i=1}^{N} Z_i' Y_i + \frac{\lambda}{N_g}\sum_{j\in g} Z_j'\right)$$

In this simplified example, I will be arbitrarily using a penalty value of 10. In a rigorous real-world use case, this value would be selected as the value that optimizes some pre-specified score, likely one that accounts for the desired accuracy and desired fairness.

Finally, I will display plots of the predicted vs actual spending values for the penalized and unpenalized linear regressions and report the R^2 (the chosen measure of accuracy) and mean residual values for the initially under-predicted sex group under both models.

Results

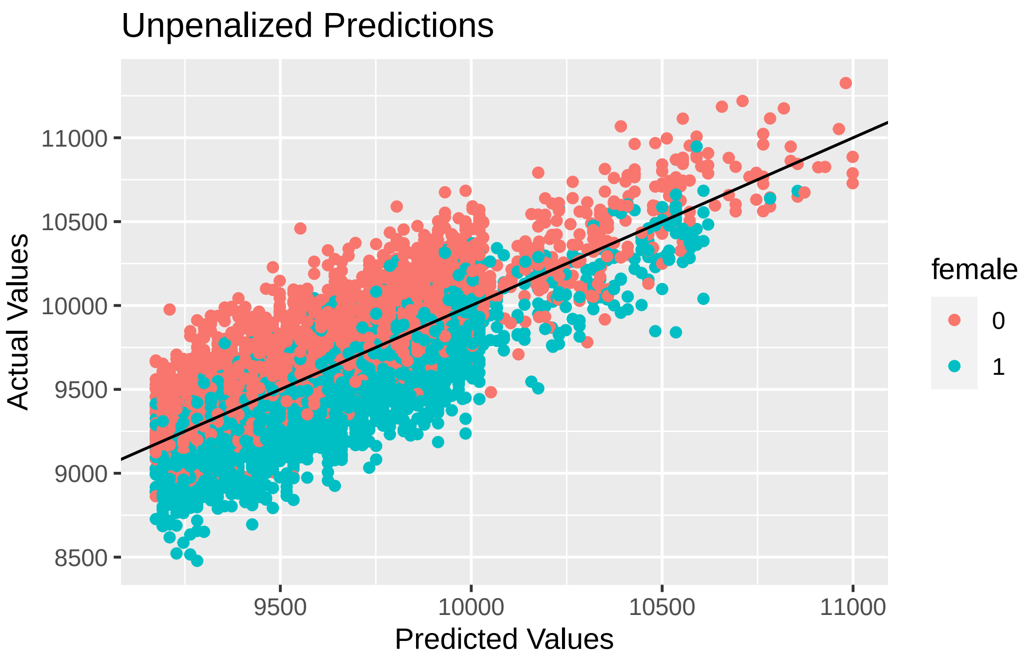
Table 1: Simulated Cohort Data by Female Sex

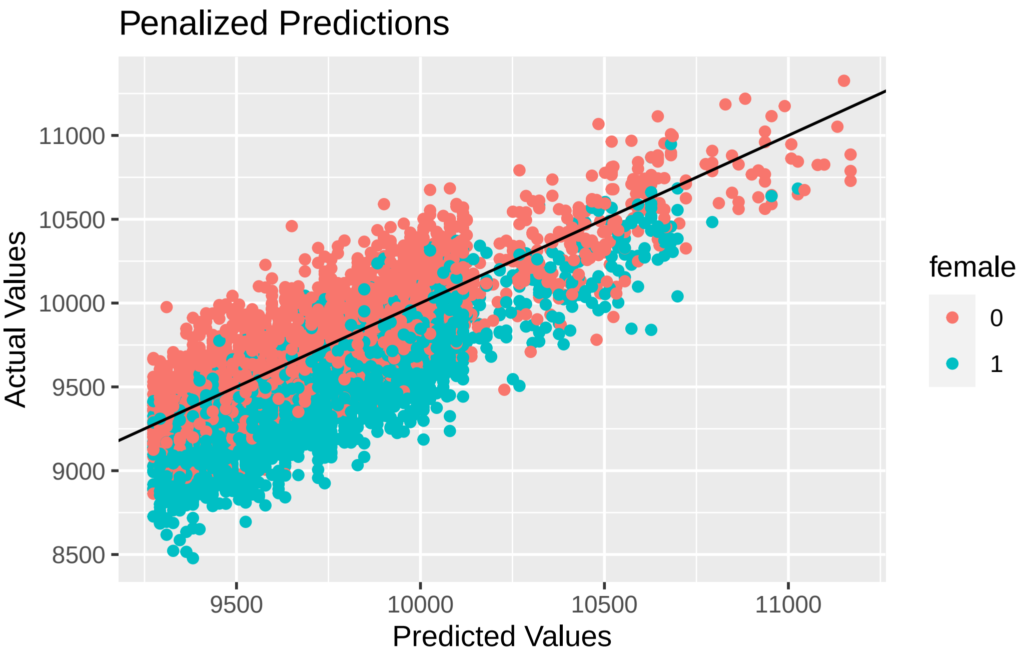
	0	1	SMD
n	2565	2435	
smoke (mean (SD))	0.11 (0.31)	0.10 (0.29)	0.038
age (mean (SD))	41.19 (13.53)	41.77 (13.55)	0.043
cardiac (mean (SD))	0.07 (0.25)	0.01 (0.09)	0.315
cost (mean (SD))	9821.72 (376.45)	9514.85 (367.88)	0.824

The simulated dataset has 5000 individuals in it, 2435 of whom have the sex female. There are heavily overlapping distributions of all variables across the sex categories, but the mean values of `cost` and `cardiac` are lower among individuals whose sex is female, with high standardized mean differences.

I find that the mean residual for individuals whose sex is female is 147.012621 and it is -139.561689 for other individuals. This indicates that non-females are under-predicted by the baseline model.

```
# add intercept to cohort data
cohort <- cbind(int = 1, cohort)
# prepare variables for matrix calculation
X <- data.matrix(cohort[c('int','smoke','age','cardiac')])
y <- data.matrix(cohort['cost'])
n <- dim(cohort)[1]
lam <- 100
# calculate penalty term
pen_term <- colMeans(cohort[cohort$female==0,
                            c('int','smoke','age','cardiac')])
# calculated Beta closed form solution
Beta <- solve(t(X) %*% X / n) %*% (t(X) %*% y / n + lam * pen_term)
```


Unpenalized Predictions



The penalized predictions are higher for all individuals, but they are differentially higher for non-female individuals, shrinking the residual gaps between sex groups.

For the penalized regression, I find that the mean residual for individuals whose sex is female is 244.6014513 and it is -37.2727228 for other individuals. This indicates that under-prediction for non-females is much lower than it is in the baseline model because the predictions are shifting up. The R^2 value for the unpenalized model is 0.6258169, while it is 0.5627066 for the penalized model. There are some decreases in accuracy that come with the substantial decrease in under-prediction.

Citations

Berk, Richard, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. n.d. "A Convex Framework for Fair Regression." https://doi.org/10.48550/arXiv.1706.02409.

Calders, Toon, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. "Proceedings - IEEE International Conference on Data Mining, ICDM." In. https://doi.org/10.1109/ICDM.2013.114.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2020. "Fair Regression via Plug-in Estimator and Recalibration with Statistical Guarantees." In, 33:1913719148. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/hash/ddd808772c035aed516d42ad3559be5f-Abstract.html.

Dwork, Cynthia, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. n.d. "Decoupled Classifiers for Fair and Efficient Machine Learning." https://doi.org/10.48550/arXiv.1707.06613.

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. n.d. "Certifying and Removing Disparate Impact." https://doi.org/10.48550/arXiv.1412.3756.

Gencoglu, Oguzhan. 2021. "Cyberbullying Detection with Fairness Constraints." *IEEE Internet Computing* 25 (1): 20–29. https://doi.org/10.1109/MIC.2020.3032461.

Gordaliza, Paula, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. 2019. "International Conference on Machine Learning." In, 2357–65. PMLR. https://proceedings.mlr.press/v97/gordaliza19a.html.

Hu, François, Philipp Ratz, and Arthur Charpentier. 2024. "A Sequentially Fair Mechanism for Multiple Sensitive Attributes." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (11): 12502–10. https://doi.org/10.1609/aaai.v38i11.29143.

Lohaus, Michael, Michael Perrot, and Ulrike Von Luxburg. 2020. "International Conference on Machine Learning." In, 6360–69. PMLR. https://proceedings.mlr.press/v119/lohaus20a.html.

Mishler, Alan, and Edward Kennedy. n.d. "FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes."

Zink, Anna, and Sherri Rose. 2020. "Fair Regression for Health Care Spending." *Biometrics* 76 (3): 973–82. https://doi.org/10.1111/biom.13206.