Anti-discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models

Fei Huang Xi Xin
UNSW Sydney

June 21, 2022
2022 Virtual ASTIN AFIR/ERM Colloquia

Introduction: Indirect Discrimination and Pricing Fairness

- A grey area in regulation: direct discrimination is prohibited, but indirect discrimination using proxies or more complex and opaque algorithms can be tolerated without restrictions.
  - What are the existing regulations related to indirect discrimination?
  - Which fairness criteria to use?
  - Which insurance pricing model to use?
  - How are they linked to each other?

Direct Discrimination

- **Direct discrimination** refers to the direct use of a protected attribute that is determined by law and prohibited from being used as a risk-rating factor.
- Common protected attributes include race, national or ethnic origin, religion or belief, gender, sexual orientation, age and disability, which usually vary by jurisdiction, line of business and even different insurance stages.

Indirect Discrimination

- **Indirect Discrimination.** After avoiding direct discrimination, indirect discrimination occurs when
  - a person is still treated unfairly than another person
  - by virtue of implicit inference from their protected characteristics,
  - based on an apparently neutral practice such as using one or a group of proxy variables from the non-protected characteristics of policyholders (i.e. identifiable proxy), or an opaque algorithm (i.e. unidentifiable proxy).
- **Disparate impact discrimination** originated in the United States and we believe it is a subset of (very similar to) indirect discrimination and intends to cover unintentional discrimination. A related concept is **Disparate Treatment**.
- **Proxy discrimination**

Algorithmic Discrimination and Responses to Big Data

- **Algorithmic discrimination** refers to the biased outcomes or decisions produced by algorithms and is usually considered as a subset of indirect discrimination.
- Insurers can offer personalized pricing for policyholders with the help of big data analytics, which makes insurance products unaffordable or unobtainable for high-risk individuals.
- Insurance regulators are publicly seeking advice on algorithmic discrimination issues, such as NAIC (US), EIOPA (EU), and AHRC (AUS).
- US price optimisation ban and UK price walking ban.

Current Laws and Regulations on indirect discrimination

- For insurance, the current regulation on indirect discrimination is mainly through *prohibiting or restricting the use of certain proxies for protected features*.
- Some traditionally or recently recognized proxy variables, such as zip code, credit information, education level and occupation, are regulated mainly because of their negative impact on (racial) minorities and low-income individuals.
- In the United States, insurers are prohibited or severely restricted to use drivers' education and occupation in automobile insurance rating in at least four states (Consumer Reports 2021).
- To the extent of our knowledge, there is no existing legal framework in any jurisdiction to assess indirect discrimination in the insurance sector.

Insurance Pricing Regulation Examples

- Actuarial fairness: similar risks should be treated similarly
- "Unfair discrimination" prohibition + protected variables + insurance exemptions
- EU unisex rule: no pricing difference due to gender

$$\text{price}(X_{NP}, X_P = \text{Female}) = \text{price}(X_N P, X_P = \text{Male}), \forall X_{NP}$$

  - on the use of gender variable
  - indirect discrimination
- Restrictions/prohibitions on proxy and protected variables
- "Disparate impact standard" (HUD) on Fair Housing Act (FHA) covering home insurance
- Colorado Senate Bill 21-169 (Conditional demographic parity)
- Community rating (AUS Private Health Insurance, Compulsory Third Party Auto Insurance)

Fairness Criteria for Insurance Pricing

- Let X_P denote the protected attribute, which is a categorical variable and has only two groups $X_P = \{a, b\}$.
- Let X_{NP} denote other available (non-protected) attributes, and hence the feature space is $X = \{X_P, X_{NP}\}$.
- Let $\hat{Y}$ denote the predictor or the decision outcome of interest, $\hat{Y} \in \mathbb{R}$. In our context, $\hat{Y}$ is the premium charged by the insurer, and in this paper, we assume that $\hat{Y}$ is approximately equal to the pure premium and ignore any expenses or profit loadings.
- Let Y denote the observed outcome of interest, $Y \in \mathbb{R}$. Note that Y is not known when the policy is issued, Y is a measure of real claim experience observed by the insurer over a given period after policy issuance.

Individual Fairness and Group Fairness

- **Individual Fairness** is analogous to the idea of treating similar people similarly (see Dwork et al. (2012), Kusner et al. (2017) and Zemel et al. (2013)).
- **Group Fairness** is sometimes used interchangeably with demographic parity or statistical parity, but here we adopt the broader meaning of group fairness, as opposed to individual fairness.
  - achieve parity across groups based on a protected feature (e.g., race or gender) in order to protect minority groups in insurance practices
- Conflict between individual and group fairness

Individual Fairness

- **Definition 1 – Fairness Through Unawareness:** fairness is achieved if the protected attribute X_P is not explicitly used in calculating the insurance premium $\hat{Y}$.
- **Definition 2 – Fairness Through Awareness:** a predictor $\hat{Y}$ satisfies fairness through awareness if it gives similar predictions to similar individuals (Dwork et al. 2012; Kusner et al. 2017).
- **Definition 3 – Counterfactual Fairness:** a predictor $\hat{Y}$ is counterfactually fair for an individual if "its prediction in the real world is the same as that in the counterfactual world where the individual had belonged to a different demographic group." (Kusner et al. 2017; Wu, Zhang, and Wu 2019)

$$\begin{aligned}\mathbb{P}(\hat{Y}_{X_P \leftarrow b}(U) = y \mid X_{NP} = x, X_P = b) \\ = \mathbb{P}(\hat{Y}_{X_P \leftarrow a}(U) = y \mid X_{NP} = x, X_P = b)\end{aligned}$$

Group Fairness Criteria

- **Definition 4 – Demographic Parity (or Statistical Parity):** a predictor $\hat{Y}$ satisfies demographic parity if

$$\mathbb{P}(\hat{Y}|X_P = a) = \mathbb{P}(\hat{Y}|X_P = b)$$

- **Definition 5 – Disparate Impact (the Four-Fifths Rule):** a predictor $\hat{Y}$ has no disparate impact if the following ratio is above than a certain threshold τ (Feldman et al. 2015):

$$\frac{\mathbb{P}(\hat{Y} = \hat{y}|X_P = b)}{\mathbb{P}(\hat{Y} = \hat{y}|X_P = a)} > \tau$$

- **Definition 6 – Conditional Demographic Parity (or Conditional Statistical Parity):** a predictor $\hat{Y}$ satisfies conditional demographic parity if

$$\mathbb{P}(\hat{Y}|L = l, X_P = a) = \mathbb{P}(\hat{Y}|L = l, X_P = b)$$

where L denotes a subset of "legitimate" attributes within unprotected attributes in the feature space ($L \subseteq X_{NP} \subset X$) (Corbett-Davies et al. 2017; Verma and Rubin 2018).

  - Unawareness $\leftrightarrow$ Conditional Demographic Parity $\leftrightarrow$ Demographic Parity

Other Fairness Criteria

- Hybrid notions between individual and group fairness
- Equalized odds/equal opportunities (separation) – unsuitable for insurance
- Actuarial group fairness (Dolman and Semenovich 2019)
- Calibration (sufficiency)

Note:

- Incompatibility of metrics

Bias Mitigation Techniques

- Preprocessing (e.g. disparate impact remover, resampling, reweighting)
- Inprocessing (e.g. regularisation, adversarial learning)
- Postprocessing (e.g. controlling for the variable, discrimination-aware pruning and relabeling of tree leaves)

Anti-Discrimination Insurance Pricing Models

In the following formulas, we show each model in its simplest form as a linear model for illustration purpose.

- Model 1: Baseline (Full) Model $\hat{Y}_{full} = f(X_{NP}, X_P)$

$$\hat{Y}_{full} = \mathbf{1}\ b_{0,1} + X_P\ b_{1,1} + X_{NP}\ b_{2,1}$$

- Model 2: Excluding Protected Variables – **Definition 1: Fairness though unawareness** $\hat{Y}_{full} = f(X_{NP})$

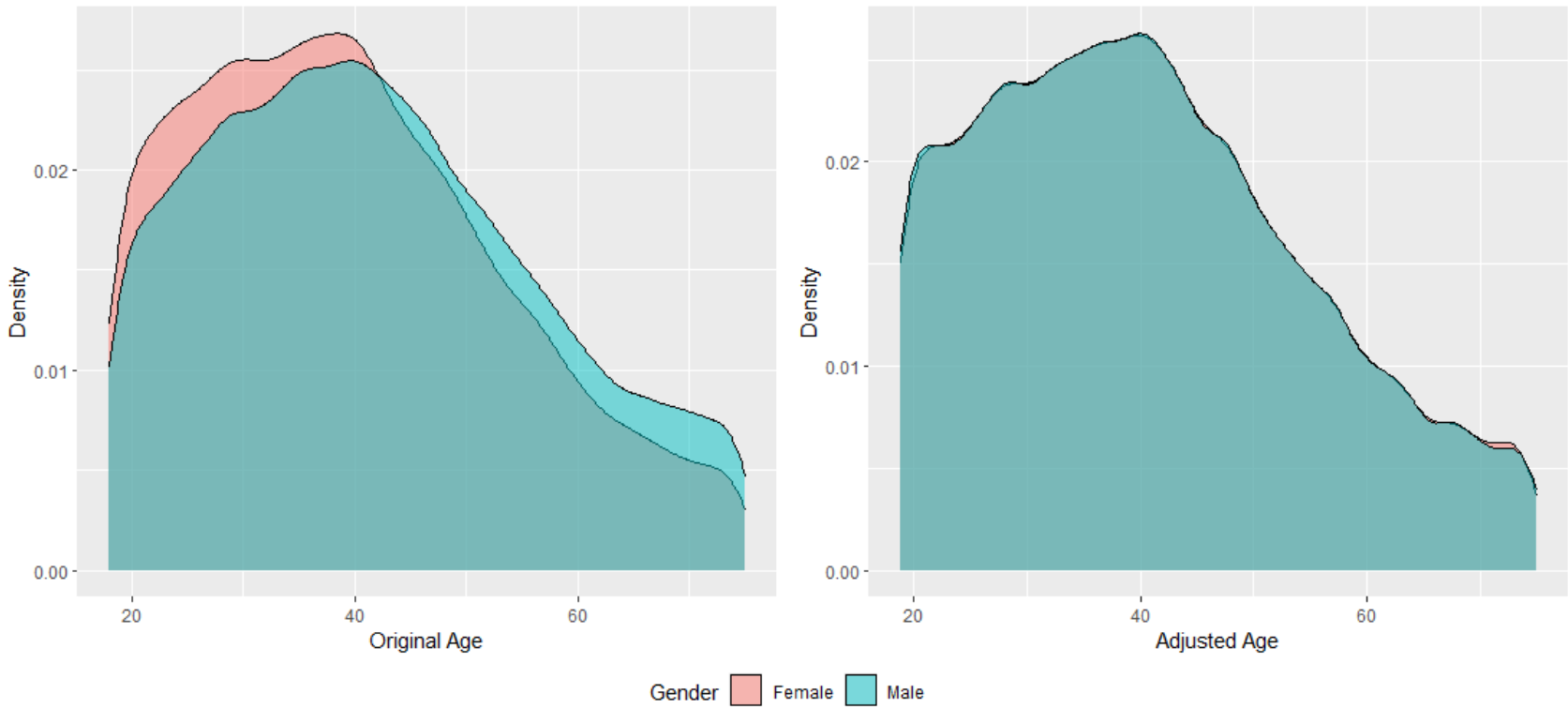
$$\hat{Y}_{restricted} = \mathbf{1}\ b_{0,2} + X_{NP}\ b_{2,2}$$

- Model 3: Fitting with Unbiased Data – **Definition 4: Demographic Parity** $\hat{Y}_{full} = f(X^{\star}_{NP})$

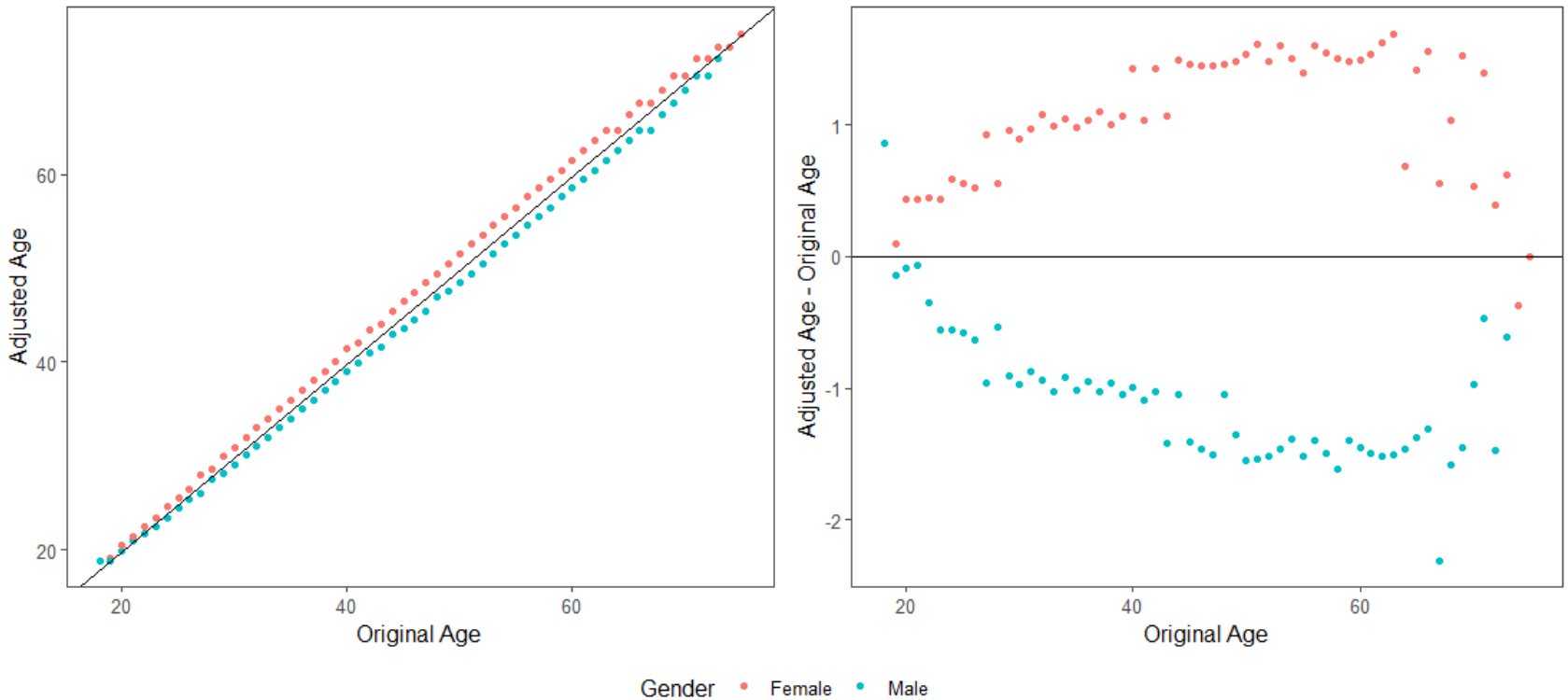
$$\hat{Y}_{FH} = \mathbf{1}\ b_{0,3} + X^{\star}_{NP}\ b_{2,3}$$

  - Method 1: Using **Disparate Impact (DI) Remover** (Feldman et al. 2015)
  - Method 2: Using Orthogonal Predictors (Frees and Huang 2021)
  - Preprocessing

How Does DI Remover Work on Age?



The Effect of DI Remover by Age and Gender



- The DI remover strongly preserves rank within groups.

Anti-Discrimination Insurance Pricing Models (Con't)

- Model 4: Fitting with Unbiased Data After Controlling for Legitimate Predictors – **Definition 6: Conditional Demographic Parity** $\hat{Y}_{full} = f(X_{NP_{legit}}, X^{\star}_{NP_{not}})$

$$\hat{Y}_{XH} = \mathbf{1}\ b_{0,4} + X^{\star}_{NP_{not}}\ b_{2,4} + X_{NP_{legit}}\ b_{3,4}$$

- Model 5: Controlling for the Protected Variable

$$\hat{Y}_{PS} = \mathbf{1}\ b_{0,1} + \bar{X}_P\ b_{1,1} + X_{NP}\ b_{2,1}$$

where the coefficients $b_{0,1}$, $b_{1,1}$, and $b_{2,1}$ are from the full model (`Model 1`) and $\bar{X}_P$ is the average over the protected variables.

  - Pope and Sydnor (2011) and Lindholm et al. (2020)
  - Postprocessing Model 1

Empirical Data

- We analyze a dataset from a French private motor insurance drawn from the `R` package `CASdatasets` (Dutang, Charpentier, and Dutang (2015)) – `pg15training`, which was used for the first pricing game organized by the French institute of Actuaries in 2015.
- We focus on the material damage (e.g., damage to a building or another vehicle) coverage.
- Data: contains 100,000 third-party liability (TPL) policies observed from 2009 to 2010.

Pure Premium Modelling

- Explanatory variables: Age, Gender, Bonus, Group1 (car group), Density (the density of inhabitants), Value (car value), Insurance Score[1].
- Protected variable: Gender
- Response variable: Pure premium (frequency x severity)
- Methods:
  - Standard frequency-severity **GLM** approach using Poisson regression and gamma regression.
  - Frequency-severity approach built on **Extreme Gradient Boosting (XGBoost)** models using Poisson deviance loss for claims frequency and gamma deviance loss for claims severity. We perform a grid search for tuning hyperparameters by steps using five-fold cross-validation.
- Performance (Accuracy) measures: Normalized Gini Index and Root Mean Square Error
- Portfolio bias adjustment (Lindholm et al. 2020)

[1]We create an insurance score for each policyholder using Type (car type), Category (car category), Occupation, Group2 (region of the driver home) and Age.

Gender Proxy

- We develop an artificial gender proxy for the probability of being female for each driver, which takes into account ten moderately efficient gender proxy variables.
- We simulate five male binary proxy variables and five female binary proxy variables. For example, in order to simulate the male proxy variable, given the gender of a person, each male has a 60% chance of being in the positive class, while each female only has a 40% chance.
- Although it may constitute indirect discrimination in the EU under the Gender Directive, or intentional discrimination in the United States, this artificial proxy predictor is added to `Model 2`, `Model 3` and `Model 4`, leading to `Model 2'`, `Model 3'` and `Model 4'` respectively.

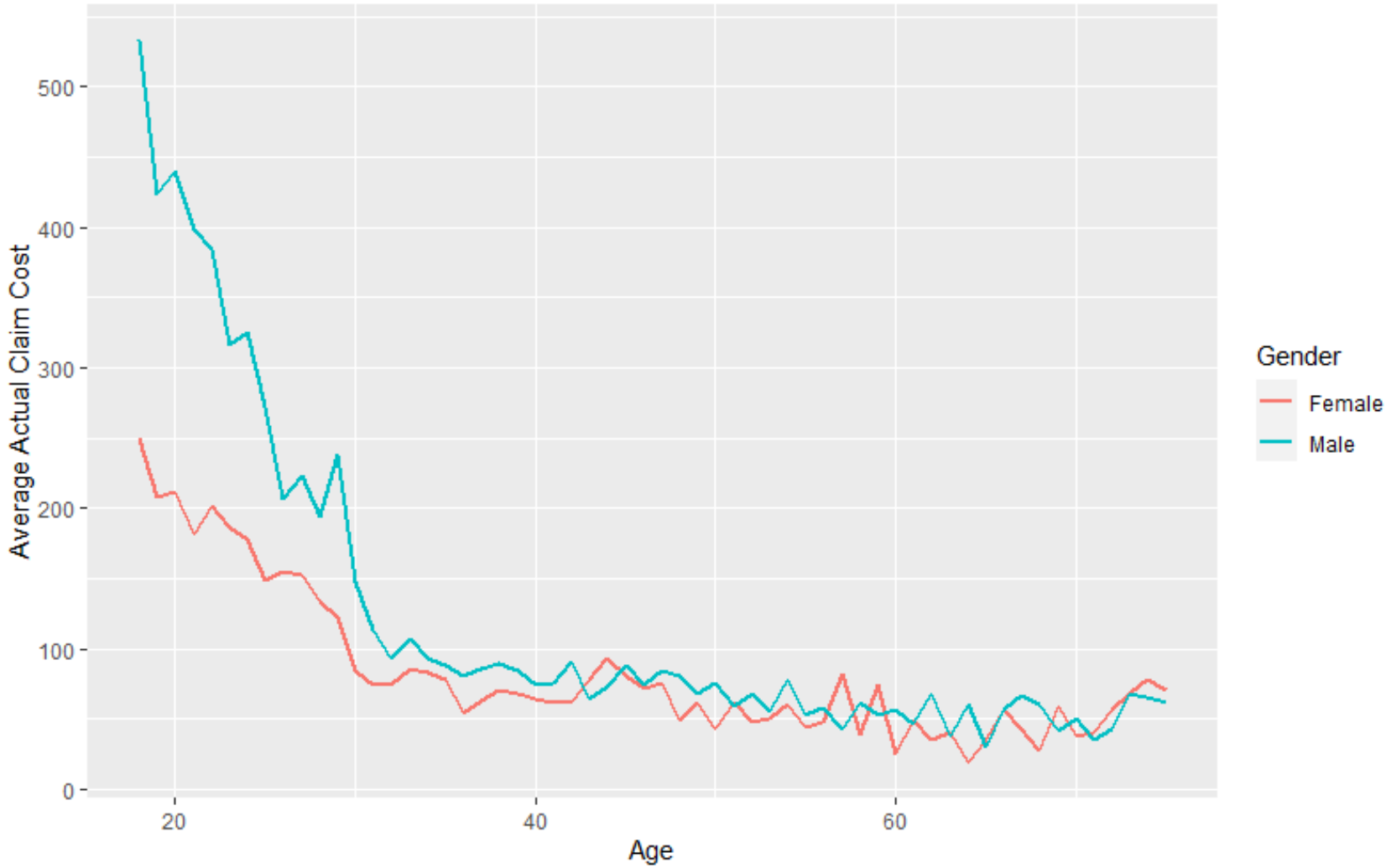
Model Fitting

- Comparison of Means of Predicted Pure Premiums by Model, Method and Gender after portfolio level adjustment.
- Insurance Score is the only non-legitimate predictor in Model 4 (baseline scenario).

	Model 1	Model 2	Model 2'	Model 3	Model 3'	Model 4	Model 4'	Model 5
GLM Male	130.47	114.03	118.53	117.38	120.78	115.26	119.40	113.95
GLM Female	95.66	124.05	116.25	118.23	112.34	121.90	114.73	124.18
XGBoost Male	130.98	114.46	118.63	117.76	120.63	116.46	119.74	114.24
XGBoost Female	94.64	123.29	116.06	117.58	112.60	119.83	114.14	123.68

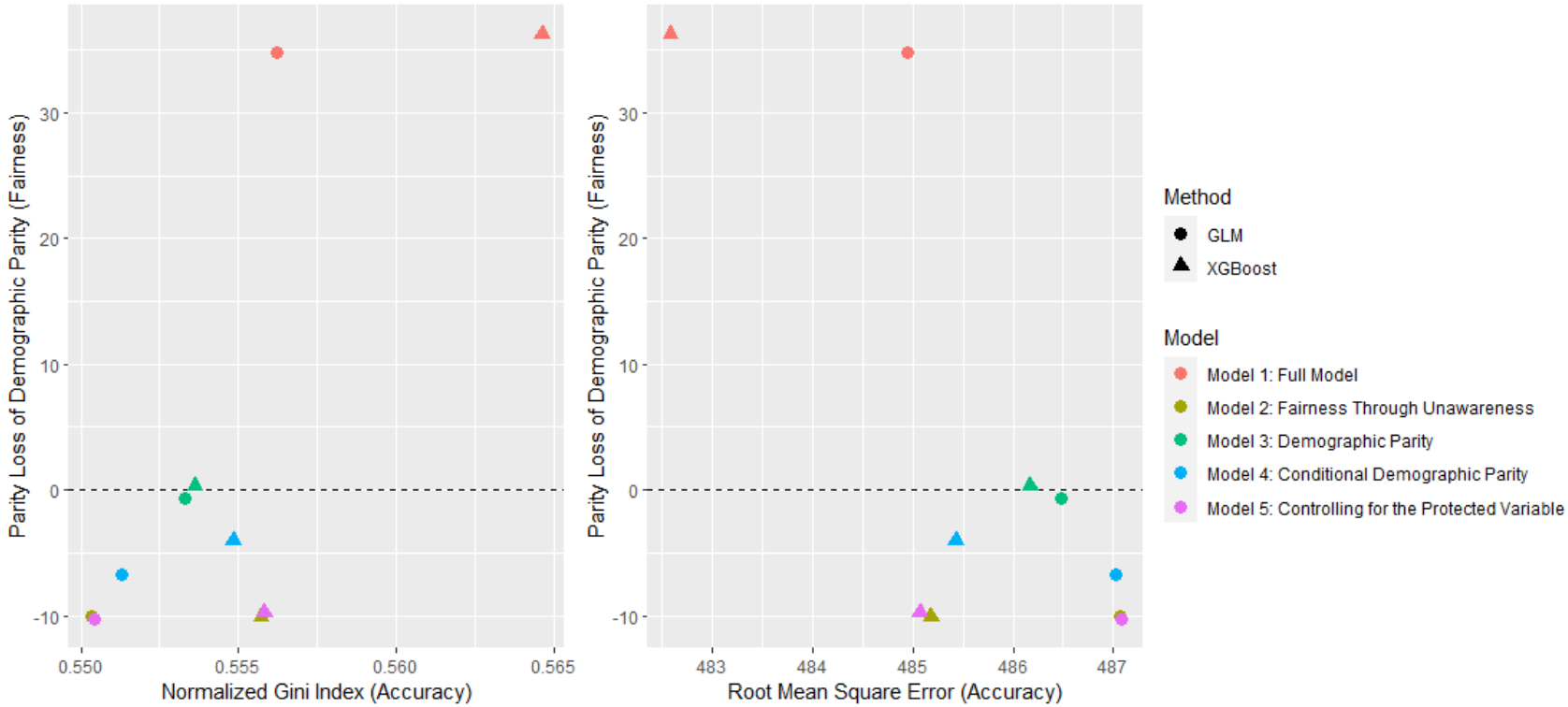
- The gender difference of Model 2 moves in the opposite direction compared to Model 1.
- Gender proxies can induce proxy discrimination, but can also improve group fairness in certain circumstances.

Average Actual Claim Cost by Age and Gender



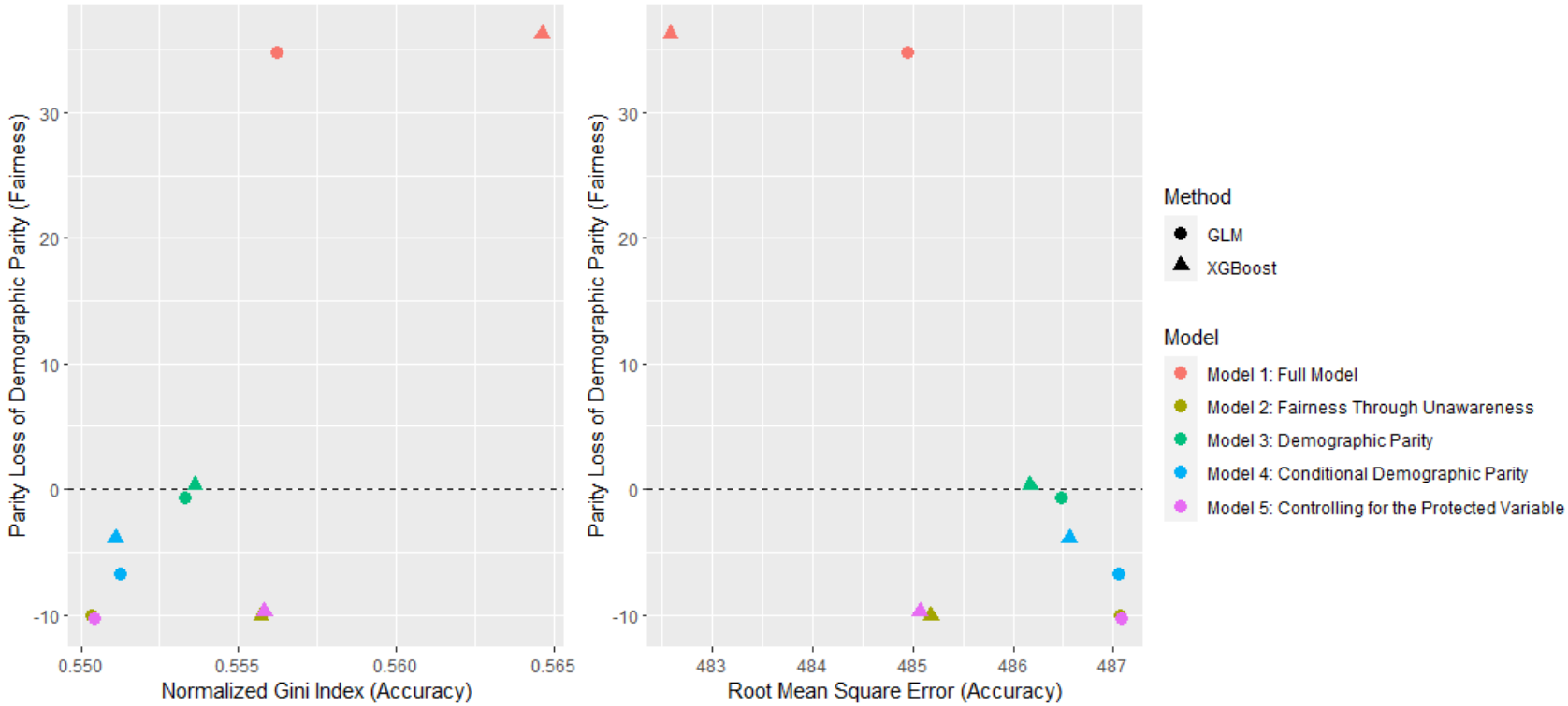
Fairness-Accuracy Plot (Scenario 1)

- Note: group fairness (demographic parity) is used in the plots.



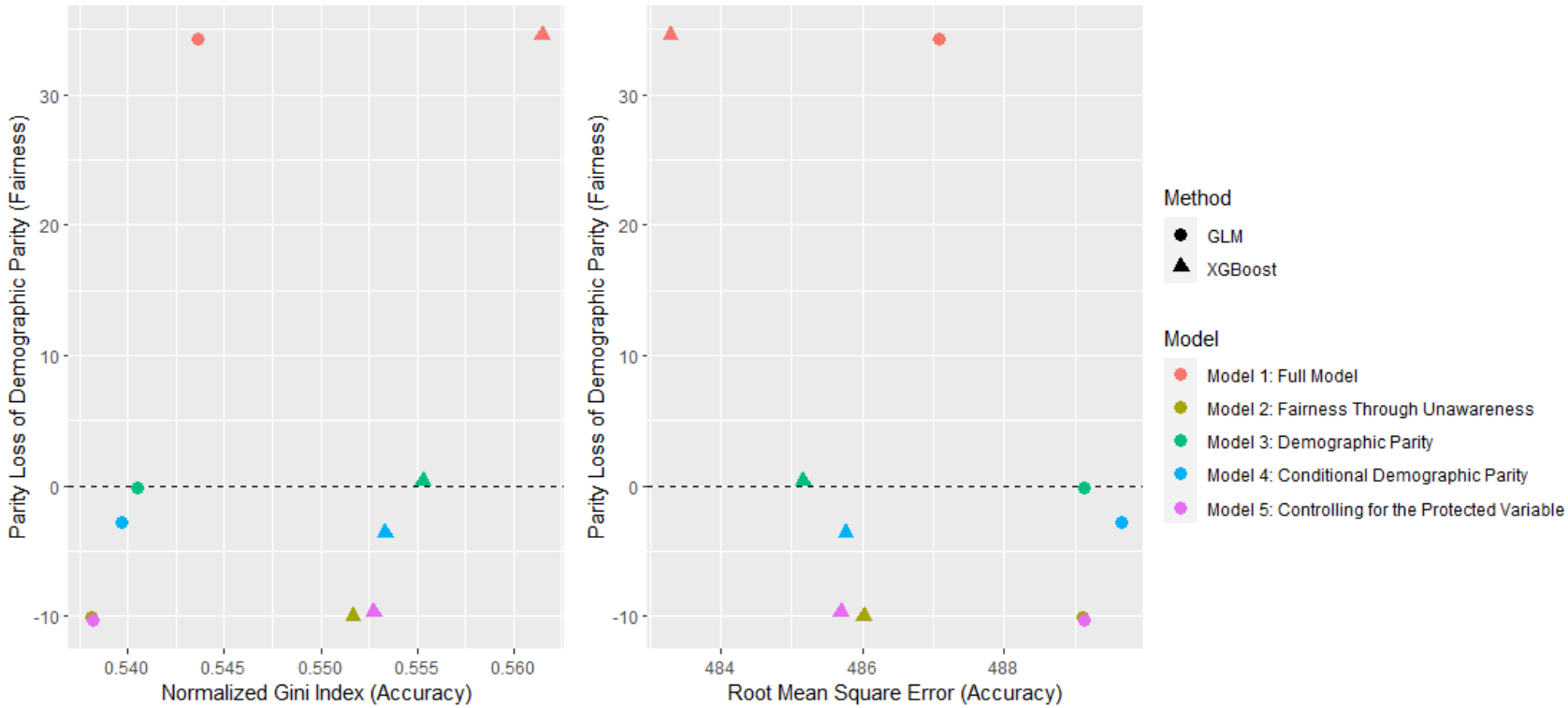
- This is the baseline scenario with Insurance Score being the only non-legitimate predictor in Model 4.
- XGBoost performs better than GLM, with different order of model accuracy performance between GLM and XGBoost.
- For GLM, model 3 is the best among models 2-5 in terms of both fairness and accuracy.
- Models 2 and 5 are similar.

Fairness-Accuracy Plot (Scenario 2)



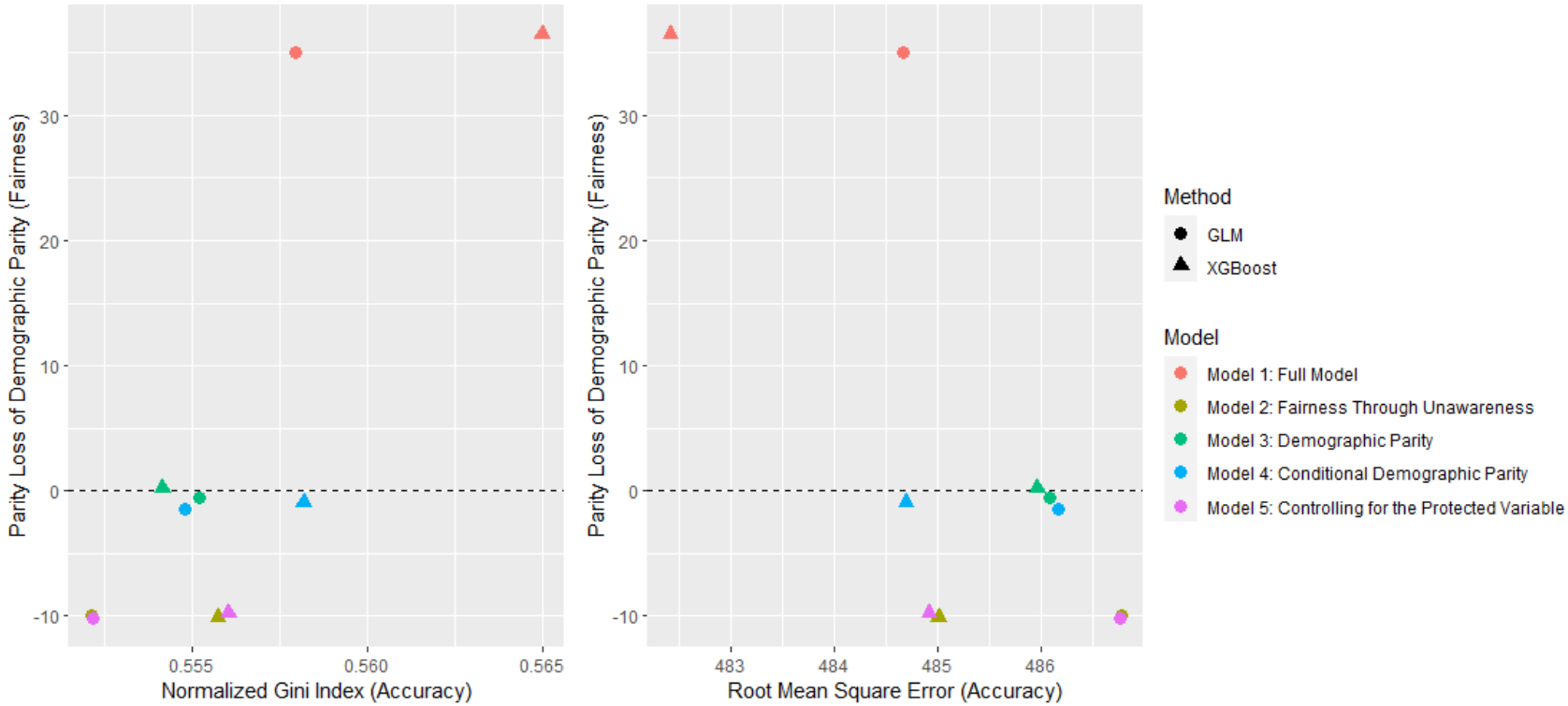
- We consider both Insurance Score and Density in Model 4 to be non-legitimate.
- The performance of XGBoost Model 4 drops a lot in accuracy.
- Compared with Scenario 1, adjusting Density with the DI remover does not improve the fairness of either method and reduces the accuracy of the XGBoost method.

Fairness-Accuracy Plot (Scenario 3)



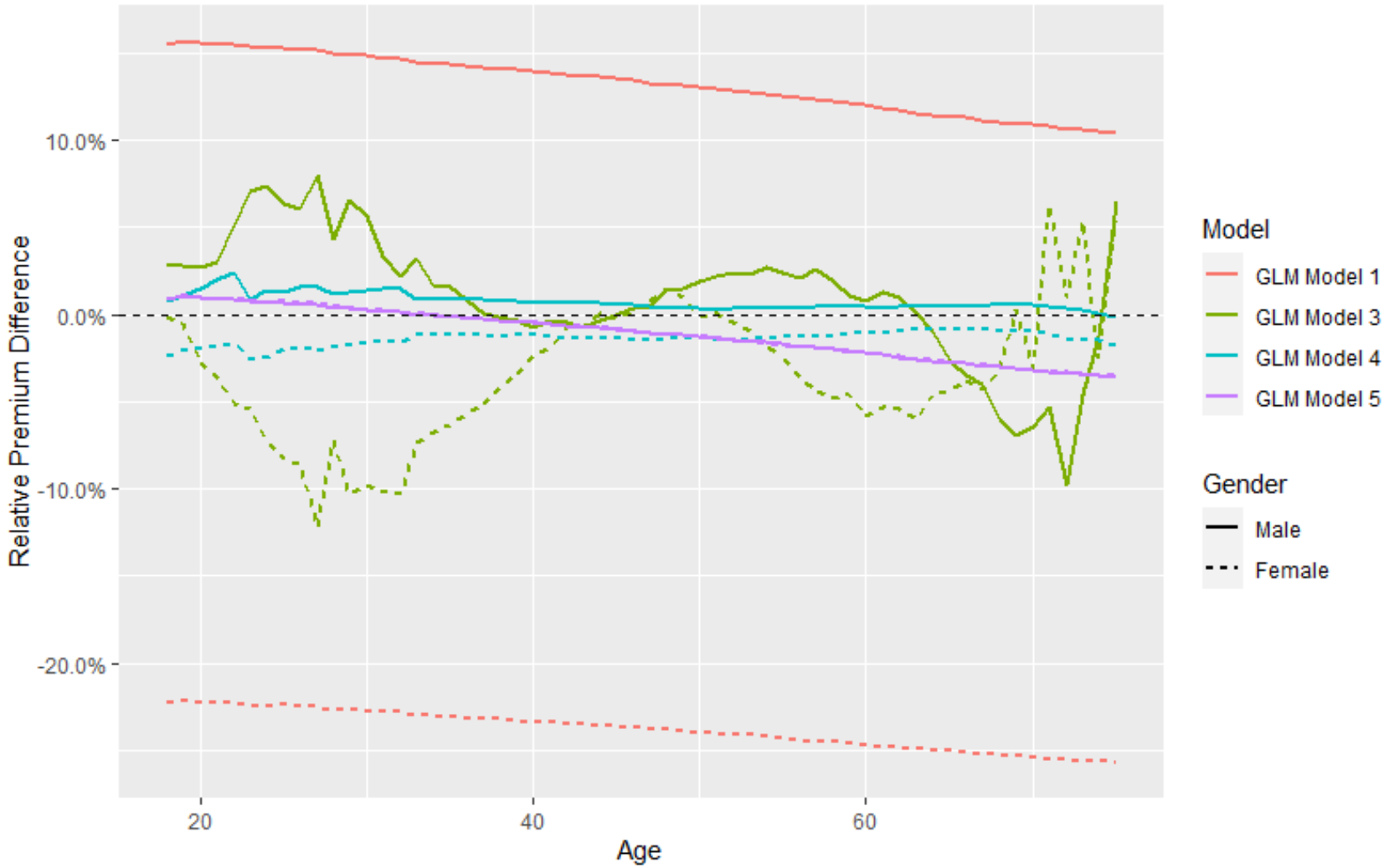
- We exclude Density in modelling compared to Scenario 1.
- Model 3 outperforms Models 2, 4, and 5 using both methods.

Fairness-Accuracy Plot (Scenario 4)

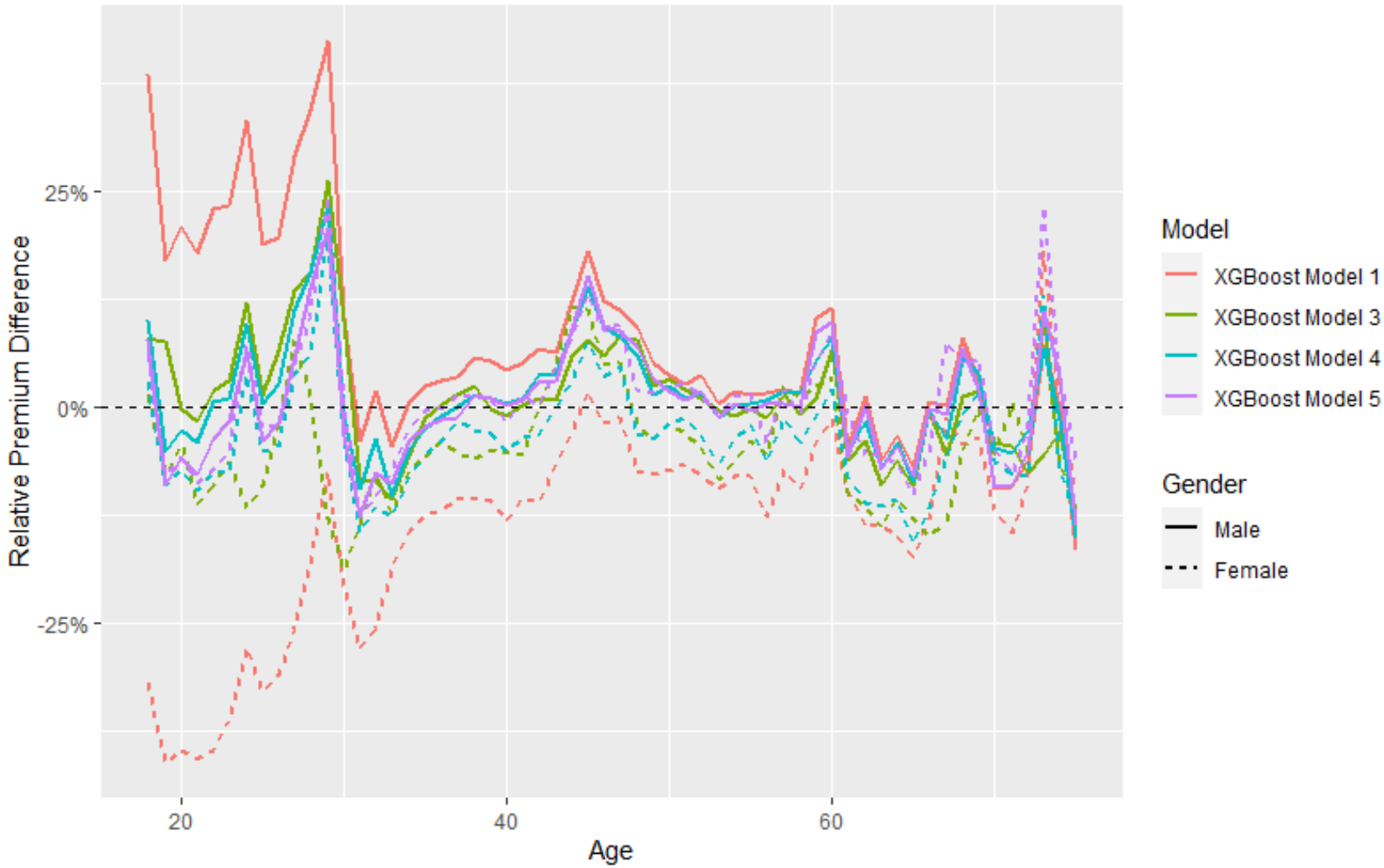


- We consider Age as the only non-legitimate variable in Model 4 compared to Scenario 1.
- Comparing with Models 2, adjusting age using DI remover improves both the accuracy and fairness of both method.
- Model 4 outperforms Models 2, 3, and 5 using XGBoost.

Relative Premium Difference (GLM Models vs. GLM Model 2)



Relative Premium Difference (XGBoost Models vs. GLM Model 2)



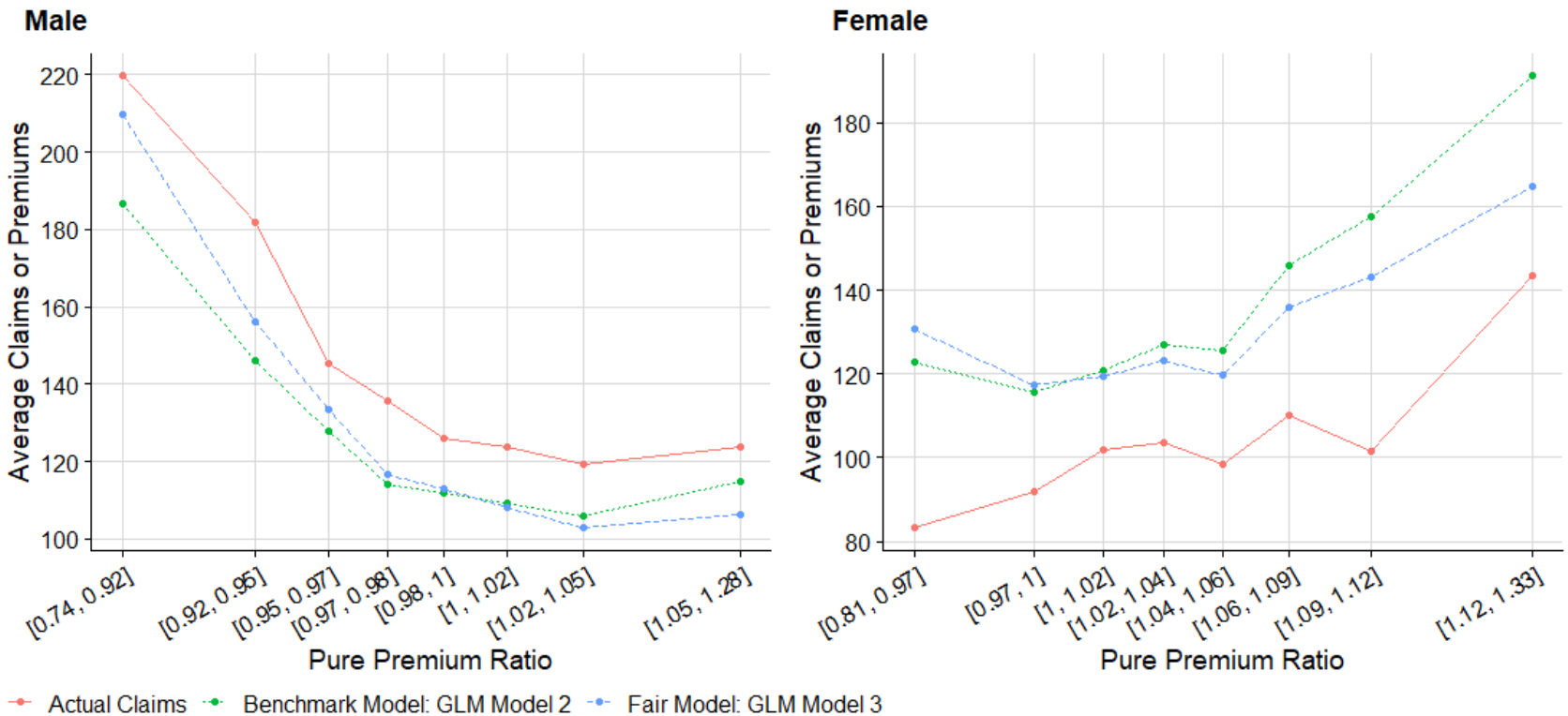
Double Lift Charts

- Analyse **adverse selection and consumer behavior**, assuming the competitor (benchmark) model is GLM Model 2.
- First, we find the pure premium ratio for each individual based on a pair of benchmark and fair models, and sort the ratios from lowest to highest.

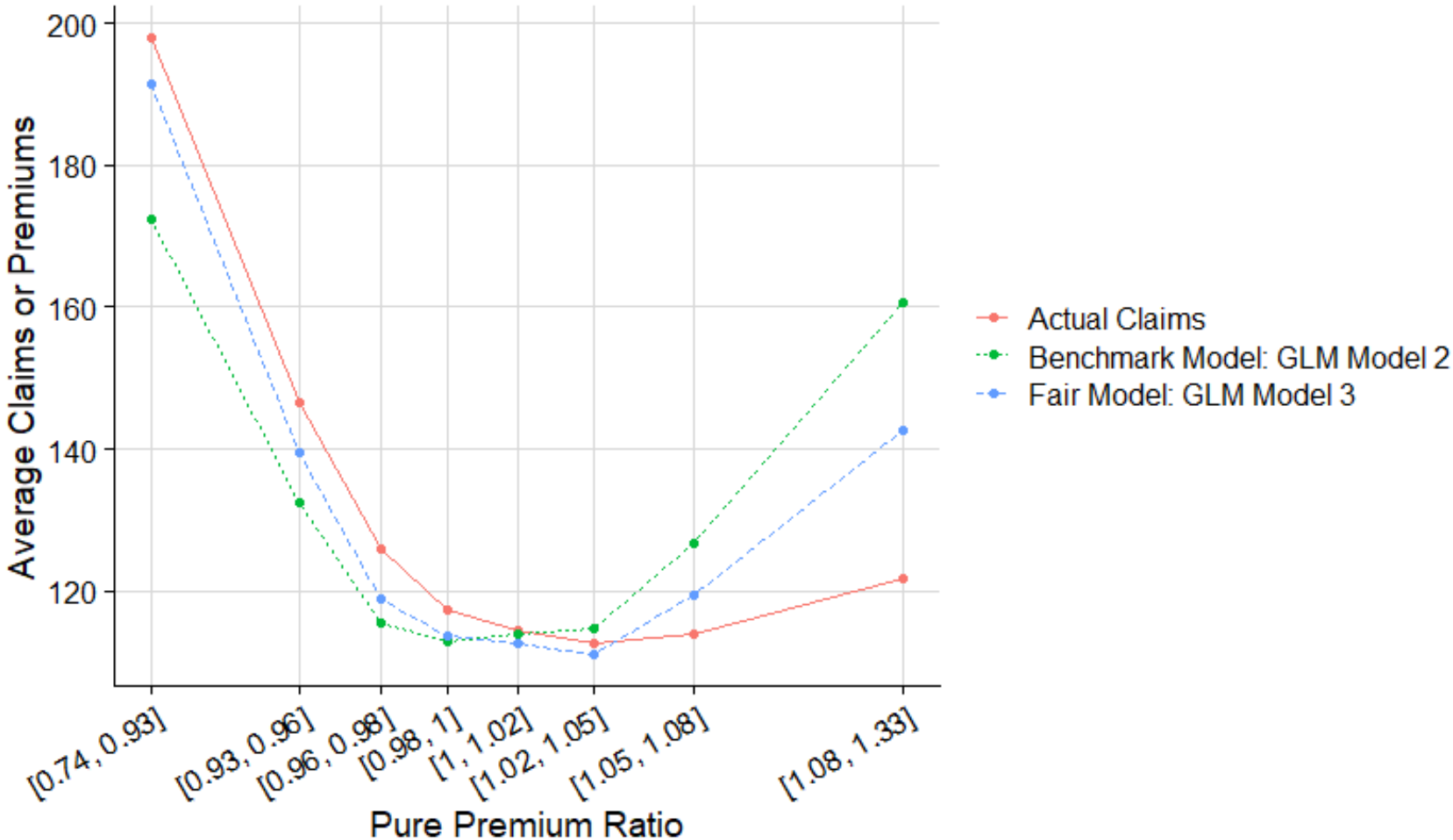
$$\text{Pure premium ratio} = \frac{\text{Predicted Premium of Benchmark Model}}{\text{Predicted Premium of Fair Model}}$$

- Then, we create bins of equal volume exposure based on the pure premium ratios calculated.
- For each bin, we calculate the average predicted premium for each model and the average actual experience based on actual claims.

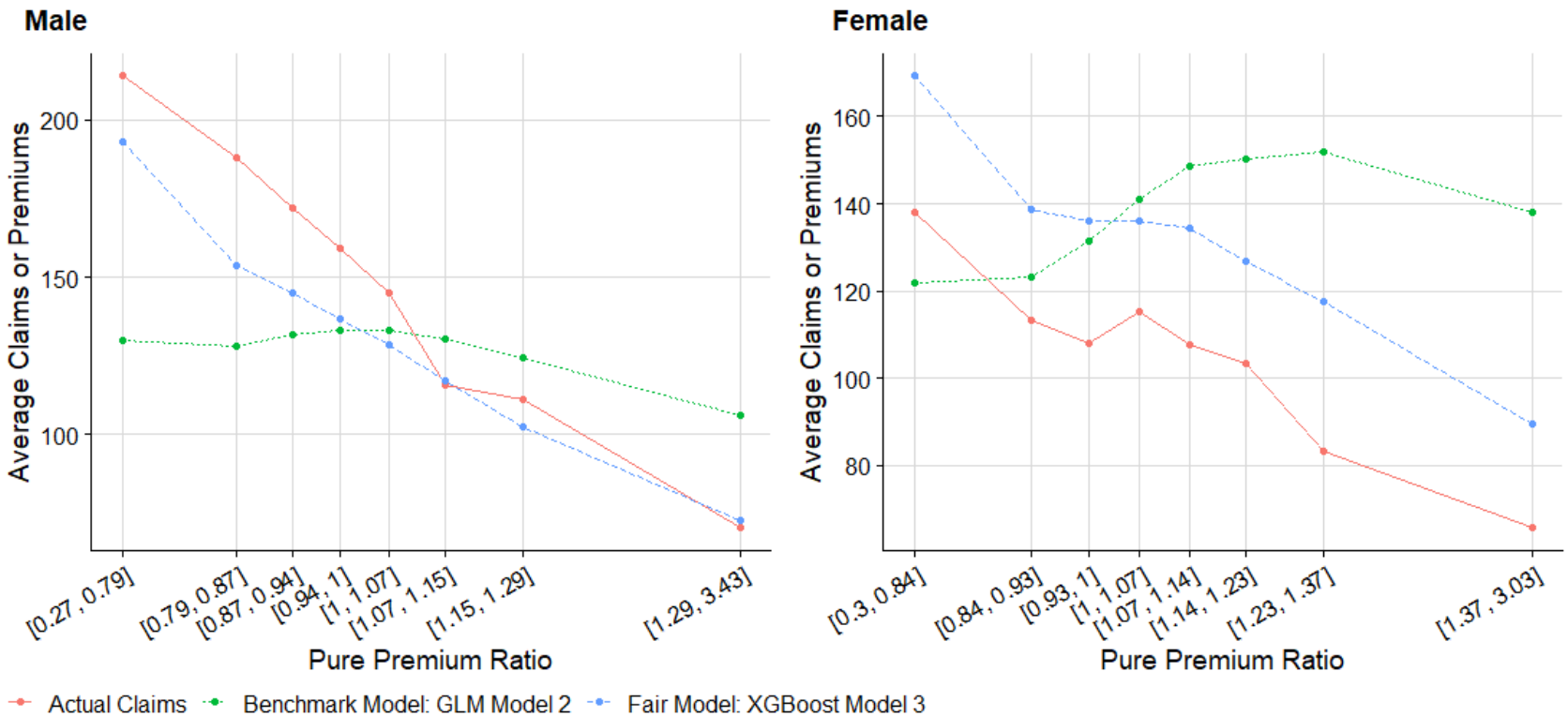
Double Lift Charts by Gender (GLM Model 3 vs. GLM Model 2)



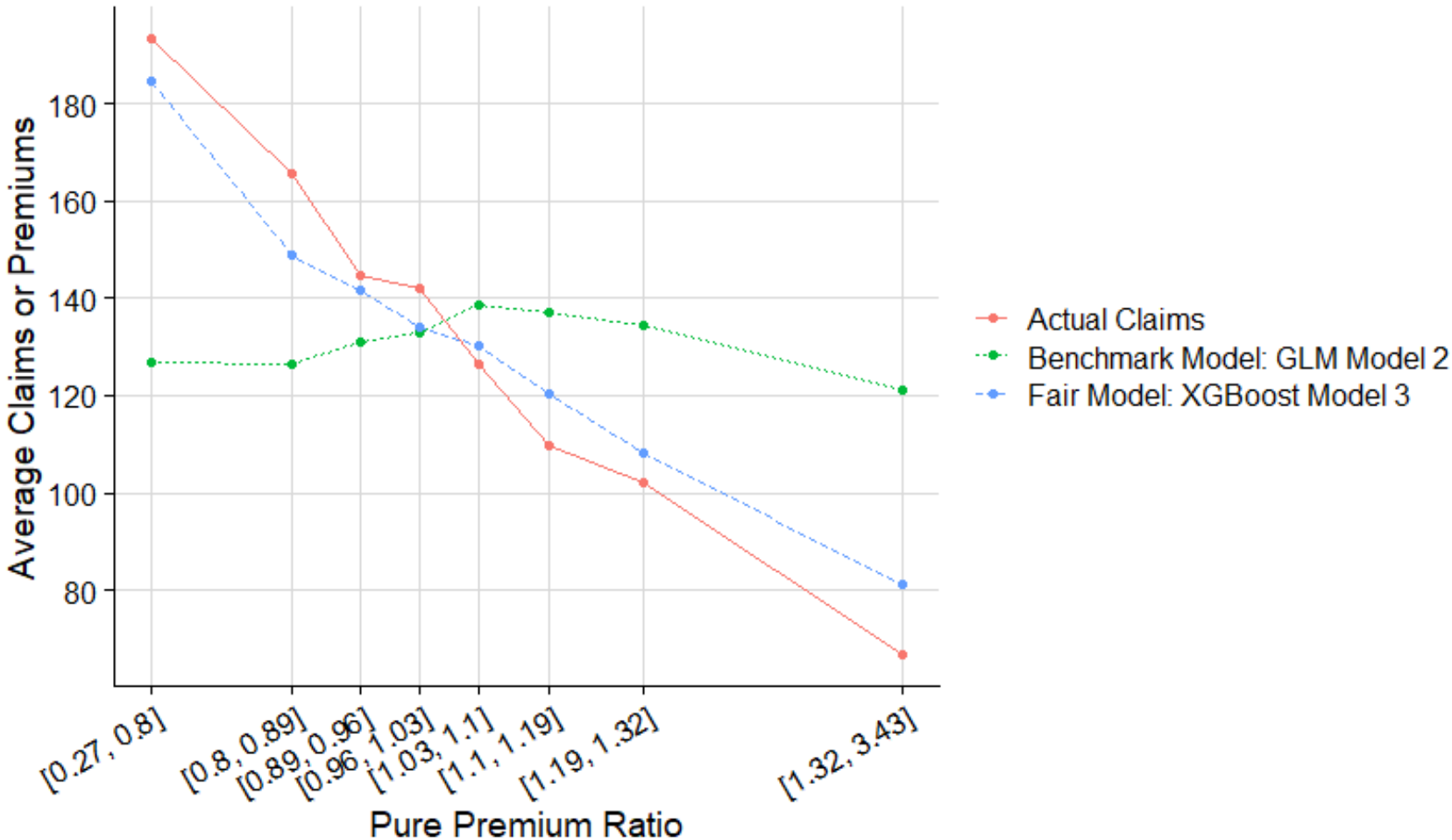
Double Lift Chart (GLM Model 3 vs. GLM Model 2)



Double Lift Charts by Gender (XGBoost Model 3 vs. GLM Model 2)



Double Lift Chart (XGBoost Model 3 vs. GLM Model 2)



Regulations

- No Regulation
- Restriction on the Use of a Protected Variable
- Prohibition on the Use of a Protected Variable
- Restriction on the Use of a Proxy Variable
- Prohibition on the Use of a Proxy Variable
- Disparate Impact Standard
- Community Rating
- Affirmative Action

Regulation Comparison

Table 4: Comparison between Different Regulations

Regulation	**Individual or Group Fairness**	**Representative Model**[24]
No Regulation	Neither	Model 1
Restriction on a Protected Variable	Neither	Model 1*
Prohibition on a Protected Variable	Individual	Model 2 or 5
Restriction on a Proxy Variable	Individual	Model 2*
Prohibition on a Proxy Variable	Individual	Model 2*[25]
Disparate Impact Standard	Group	Model 3 or 4
Community Rating	Group	Model 3 or 4
Affirmative Action	Neither	None

Summary

- We establish a connection among various insurance regulations, fairness criteria and anti-discrimination insurance pricing models.
- We show the fairness and accuracy trade-off of different insurance pricing models, impacted by the methods (GLM or XGBoost), anti-discrimination models, explanatory and legitimate variables.
- We analyse the impact of implementing different models on adverse selection and consumer behavior.
- Which fairness criteria/regulation should regulators use? It depends on the regulations and laws of specific lines of business and jurisdictions.
- Which insurance pricing model should insurers use? It depends on the regulations/laws of specific lines/jurisdictions, features of explanatory and protected variables, and model performance.

Thank You!

Comments and suggestion are welcome!

feihuang@unsw.edu.au and xi.xin@unsw.edu.au

References I

Consumer Reports. 2021. "Effects of Varying Education Level and Job Status on Online Auto Insurance Price Quotes." https://advocacy.consumerreports.org/wp-content/uploads/2021/01/Auto-Insurance-White-Paper-Report-FINAL1.26C.pdf.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." In *Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 797–806.

Dolman, Chris, and Dimitri Semenovich. 2019. "Algorithmic Fairness: Some Practical Considerations for Actuaries." *Actuaries Summit 2019*.

Dutang, Christophe, Arthur Charpentier, and Maintainer Christophe Dutang. 2015. "Package 'CASdatasets'."

References II

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–26.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–68.

Frees, Edward W, and Fei Huang. 2021. "The Discriminating (Pricing) Actuary." *North American Actuarial Journal, Forthcoming*.

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual Fairness." In *Advances in Neural Information Processing Systems*, 4066–76.

References III

Lindholm, Mathias, Ronald Richman, Andreas Tsanakas, and Mario V Wuthrich. 2020. "Discrimination-Free Insurance Pricing." *Available at SSRN*.

Pope, Devin G, and Justin R Sydnor. 2011. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy* 3 (3): 206–31.

Verma, Sahil, and Julia Rubin. 2018. "Fairness Definitions Explained." In *2018 Ieee/Acm International Workshop on Software Fairness (Fairware)*, 1–7. IEEE.

Wu, Yongkai, Lu Zhang, and Xintao Wu. 2019. "Counterfactual Fairness: Unidentification, Bound and Algorithm." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. "Learning Fair Representations." In *International Conference on Machine Learning*, 325–33. PMLR.