Insurance pricing: discrimination, causality, and fairness

Over the last decade there has been a surge in applying machine learning techniques in non-life insurance pricing. This is mainly due to cheaper data collection and storage, combined with new analysis methods for unstructured data and increased computational power. In parallel there have been rising concerns about data privacy and hidden implications of using "black-box" price predictions.

An obvious concern when using black-box models for pricing is that of implicit discrimination. As we will see below, this is *always* a potential issue, regardless of the model, and this is regulated in EU-law, see [4]. A related question is that of (algorithmic) fairness, and below it will be seen that these two concepts will often fail to agree.

Further, in order to adjust for implicit discrimination, discriminatory information needs to be collected (more on this below), which is a privacy concern in itself. More generally, this relates to the question of which covariates that are suitable to use for pricing. This in turn connects to discussions about covariates' causal effects and risk factors. However, below it will be seen that this is not essential for avoiding implicit discrimination.

THE ACTUARIALLY FAIR PREMIUM AND DISCRIMINATION

Let X denote the covariates (rating factors or policy features), and Y the claim cost that we try to predict. The actuarially fair premium, $\mu(x)$, is defined as $\mu(x) = E\,[\,Y \mid X = x]$ and can be interpreted as the best prediction of the future claim cost Y, given the specific policy features $X = x$. Charging the premium $\mu(x)$ to each policyholder will on average generate a total premium income equal to the expected claim cost.

How will pricing be affected when there are covariates, D, that are considered protected, such as sex or ethnicity? For example, EU regulation [4] stipulates that insurers are not allowed to price insurance policies based on sex, neither directly nor indirectly. Direct discrimination occurs when the price explicitly depends on D. Therefore, the actuarially fair, best-estimate, premium based on all information, $\mu(x,d) = E\,[\,Y \mid X = x,\ D = d]$, cannot be used by insurers, since it explicitly depends on D. The definition of indirect discrimination is more complex [4] and can be interpreted as reflecting two distinct ideas:

(i) when using the non-protected covariates X, adjustments should be made to ensure that D is not implicitly proxied by X;
(ii) the effect of the pricing practice should not lead to a disadvantage for either sex.

Property (i) is meant to prevent *proxy discrimination* by requiring that D cannot be learned from X; e.g. for some portfolios a policyholder's ethnicity may be accurately predicted from their postcode. For property (i) one can give a statistical definition [5]. Property (ii) is referred to as *disparate impact*, for which there are multiple alternative (and potentially conflicting) mathematical formulations and which has a specific meaning in US law. For the remainder we therefore focus on property (i), as precisely formulating property (ii) remains in part an open research question.




From left to right:
Prof. M. Lindholm is associate professor in Mathematical Statistics at the Department of Mathematics at Stockholm University.
R. Richman is Chief Actuary at Old Mutual Insure. He is a Fellow of the Institute and Faculty of Actuaries (IFoA) and the Actuarial Society of South Africa (ASSA).
Prof A. Tsanakas is professor in Risk Management at Bayes Business School, City, University of London, and Editor-in-Chief of the Annals of Actuarial Science.
Prof. M. Wüthrich is professor in the Department of Mathematics at ETH Zurich

ADJUSTING FOR POTENTIAL PROXY DISCRIMINATION

The premium $\mu(x)$ does not explicitly use the protected information D, and is hence not subject to direct discrimination. Nonetheless, we cannot be certain that it will not be affected by proxy discrimination. This is because, the calculation of $\mu(x)$ implicitly reflects the dependence between D and X. To see this more clearly, consider the situation where $D = 0$ ("man") or 1 ("woman"), for which we can write:

$$\mu(x) = \mu(x,d{=}0)p(d{=}0|x) + \mu(x,d{=}1)p(d{=}1\,|\,x) \qquad (1)$$

where $p(d|x)$ denotes the probability of a policyholder having the sex $D = d$, given the non-protected information $X = x$. This illustrates how standard covariates may carry information about the protected covariate D. At the extreme, if we could perfectly predict the sex based on non-protected information, then there would be no practical difference between direct and indirect discrimination.

We can however remove the potential for proxy discrimination, by modifying Eq. (1) to:

$$\mu^*(x)) = \mu(x,d{=}0)p^* + \mu(x,d{=}1)(1-p^*), \qquad (2)$$

where p^* is some probability between 0 and 1 for which the portfolio fraction of $D = 0$ is a natural choice. We call the adjusted premium from Eq. (2) the *discrimination-free insurance price* (DFIP), see also Ref. [5] where the DFIP is discussed in more generality and detail. By using $\mu^*(x)$ any potential dependence between X and D – and hence any proxy discrimination — has been removed, without requiring further assumptions.

CONCLUSION AND OUTLOOK

Through the pricing method of Eq (2), we have proposed a way to adjust actuarially fair prices in order to address proxy discrimination. Importantly for practice, the calculation of DFIP is model-agnostic, in the sense that it can be derived as an adjustment to *any* pricing model, from GLMs to complex machine learning models. Nonetheless, in order to determine the DFIP in Eq. (2), it is necessary to have access to the more detailed price $\mu(x,d)$, which can only be estimated using protected data. Thus, to appropriately quantify the materiality of proxy discrimination and correct for it, the collection of some protected information is needed. This requirement may raise privacy concerns and a technical solution is discussed in Ref. [6].

A further consideration is about which types of covariates should be used in the first place, when calculating an insurance price. Some policy features may be classified as *risk factors*, if they have a *direct causal effect on* claims, see e.g. [3, 1, 8]. Standard rating factors are not necessarily assumed to causally impact Y; instead they are characterised by *statistical association* with Y. In Ref. [5] it is shown that, following certain causal assumptions, the DFIP from Eq. (2) will coincide with the expected direct causal effect of X on Y. However, in real-world applications it is rarely the case that all risk factors are observed or that their causal interrelations are fully understood. In these situations, the causal effect can likely not be assessed, and the causal connection to Eq. (2) is lost. Nonetheless, also in such more complex settings, the DFIP will still correctly adjust for proxy discrimination. Furthermore, a requirement to use *only* risk factors with a direct causal effect on claims will likely reduce the number of covariates that are available, see [3]. This, despite its conceptual appeal, will also incur a cost in terms of predictive accuracy.

Finally, arguments around discrimination relate to notions of algorithmic fairness, which has attracted considerable attention in the machine learning literature. For example, there is persistent concern that machine learning algorithms discriminate against sub-populations, in applications ranging from e.g., mortgage lending to facial recognition [2]. Algorithmic fairness is typically defined in terms of statistical properties of predictors. For example, in order to satisfy *demographic parity*, a predictor $\hat{\mu}(X)$ should be independent of D. This means that there should be no statistical association between risk predictions and protected characteristics. This is a very penal requirement, because in situations where there is some statistical association between the nonprotected covariates X and the protected covariates D, the prices are not allowed to include any information from X. Requiring that prices are statistically independent from protected covariates implies that in some portfolios, e.g. where policyholders from one demographic group are more likely to engage in high risk behaviours (e.g. smoking), it becomes impossible to apply risk pricing. We note that this is not the case with the DFIP, which allows risk pricing based on non-protected covariates, only adjusting for their proxying effect. More generally, this illustrates potential conflicts between adjusting for proxy discrimination, while trying to satisfy common algorithmic fairness conditions. For more, examples see [7]. ■

References

[1] Araiza Iturria C.A., Hardy M., Marriott P.A. (2022) Discrimination-Free Premium Under a Causal Framework. *Available at SSRN*: https://ssrn.com/abstract=4079068

[2] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

[3] EIOPA (2021) Artificial Intelligence Governance Principles: Towards Ethical and Trustworthy Artificial Intelligence in the European Insurance Sector: A Report from EIOPA's Consultative Expert Group on Digital Ethics in Insurance. *Available at*: https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa-ai-governance-principles-june-2021.pdf

[4] European Council (2004) COUNCIL DIRECTIVE 2004/113/EC – implementing the principle of equal treatment between men and women in the access to and supply of goods and services. *Official Journal of the European Union*, L 373, pp. 37 — 43

[5] Lindholm M., Richman R., Tsanakas A., Wüthrich M.V. (2022) Discrimination-Free Insurance Pricing. *ASTIN Bulletin*, 52(1), pp. 55 — 89

[6] Lindholm M., Richman R., Tsanakas A., Wüthrich M.V. (2022) A Multi-Task Network Approach for Calculating Discrimination-Free Insurance Prices. *Available at SSRN*: https://ssrn.com/abstract=4155585

[7] Lindholm M., Richman R., Tsanakas A., Wüthrich M.V. (2022) A Discussion of Discrimination and Fairness in Insurance Pricing. *Available at SSRN*: https://ssrn.com/abstract=4207310

[8] Pearl J. (2009) Causal inference in statistics: An overview. *Statistics Surveys*, 3, pp. 96 —146

This article has earlier been published in The European Actuary in March 2023 (https://actuary.eu/wp-content/uploads/2023/02/TEA-33-MAR-DEF.pdf)