DISCRIMINATION-FREE INSURANCE PRICING

By

M. Lindholm, R. Richman, A. Tsanakas and M.V. Wüthrich

Abstract

We consider the following question: given information on individual policyholder characteristics, how can we ensure that insurance prices do not discriminate with respect to protected characteristics, such as gender? We address the issues of direct and indirect discrimination, the latter resulting from implicit learning of protected characteristics from nonprotected ones. We provide rigorous mathematical definitions for direct and indirect discrimination, and we introduce a simple formula for discrimination-free pricing, that avoids both direct and indirect discrimination. Our formula works in any statistical model. We demonstrate its application on a health insurance example, using a state-of-the-art generalized linear model and a neural network regression model. An important conclusion is that discrimination-free pricing in general requires collection of policyholders' discriminatory characteristics, posing potential challenges in relation to policyholder's privacy concerns.

Keywords

Causal inference, differentiation, direct discrimination, discriminatory covariates, indirect discrimination, individual policy characteristics, insurance pricing, proxy discrimination.

JEL codes: G22, J16, J15, C45, D49

1. Introduction

Motivation and context. We address the following fundamental question: given information on individual policyholder characteristics, how can we calculate insurance prices that do not discriminate with respect to protected characteristics, such as gender? This is a pertinent question in the context of

Astin Bulletin 52(1), 55–89. doi:10.1017/asb.2021.23

anti-discrimination legislation; for instance, current EU law requires gender neutral insurance pricing, see European Council (2004). This question has become even more pronounced with the emergence of big data and associated developments in complex algorithmic models, since such models may be able to infer discriminatory characteristics from other policyholder features. For an overview on antidiscrimination laws, we refer to Avraham *et al.* (2014) and Prince and Schwarcz (2019).

We aim at developing pricing formulas that are devoid of discrimination, while the insurer is still able to differentiate between policyholders with respect to nonprotected characteristics. Here, by "discrimination" we mean the provision of insurance prices that differentiate between policyholders on the basis of (legally) prohibited characteristics. For this, we assume that an insurer has access to policyholders' data that can be split into discriminatory (e.g., gender, ethnicity) and nondiscriminatory characteristics (e.g., age, smoking habits). When we refer to discriminatory characteristics, we are relying on legal and regulatory requirements, such as those in the EU, which prohibit insurers from using certain characteristics within their pricing framework. In such a legal context, the use of protected characteristics amounts to illegal discrimination, thus creating an imperative for insurance pricing models to avoid using them. For example, within the EU, the council directive (European Council, 2004) provides definitions of *direct* and *indirect discrimination*, motivating our technical arguments.

Direct discrimination can be easily understood and identified as the use of prohibited characteristics as rating factors. Indirect discrimination presents more of a challenge, because it can be thought of as the confluence of two distinct effects: (a) the implicit ability to infer protected characteristics from other (legitimately used) policyholder features and (b) a systematic disadvantage resulting for a group that is protected by a nondiscrimination provision (Tobler, 2008). These two concepts are interrelated but distinct, the former, also referred to as *proxy discrimination*, arises from correlation between protected and unprotected characteristics; the latter, *disparate impact*, from correlations between protected characteristics and actual insurance prices – we refer to Frees and Huang (2021) for a detailed discussion from an actuarial perspective. The pricing adjustment we propose explicitly addresses (a) however such an adjustment may be legally unnecessary if (b) is not additionally present. Both these effects disappear when discriminatory characteristics are statistically independent of nondiscriminatory ones, though this observation does not imply that (a) and (b) are mathematically or conceptually equivalent.

The development of ideas in this paper is drawn from an actuarial rather than a legal perspective. We do not make any claim about their correspondence to legal definitions of discrimination in particular jurisdictions and do not argue that the pricing adjustment as proposed in this paper should be applied in all circumstances. Our focus is to provide an explicit mathematical method to remove indirect discrimination – if it happens to exist – from insurance pricing models. We begin our arguments on the assumption that certain

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

characteristics have been prohibited and consider how pricing models can be adapted correspondingly. We say that

- A pricing model *avoids direct discrimination*, if none of the discriminatory features (characteristics) is used as a rating factor.
- A pricing model *avoids indirect discrimination*, if it avoids direct discrimination and, furthermore, the nondiscriminatory features are used in a way that does not allow implicit inference of discriminatory features from them.

To help clarify these concepts, we consider examples of directly and (potentially) indirectly discriminatory rating factors. In many jurisdictions, it is illegal to include the race/ethnicity of a policyholder within a pricing model, meaning that direct discrimination on the basis of race is illegal, even if race was (hypothetically) a good predictor of propensity to claim. There are other rating factors that are highly correlated with race, but which do not have much direct impact on the propensity to claim. For example, a policyholder's native language is highly correlated with race in parts of the world where certain languages are spoken only by members of a particular race, and including this rating factor within a pricing model will do little but act as a proxy for race. Hence, including this rating factor may lead to what we term indirect discrimination in this work.

Then, there are rating factors that may be both directly predictive of insurance claims as well as act as proxies for discriminatory characteristics. For example, using the presence of diabetes as a rating factor will be directly predictive of health insurance costs, but since certain racial or ethnic groups may be predisposed to develop diabetes, including diabetes as a rating factor may lead to this rating factor acting as a proxy for race, potentially leading to indirect discrimination. Our aim in this paper is to develop a method that is capable of removing both direct and indirect discrimination from pricing models, where these may exist, while maintaining the predictive nature of variables that do not directly discriminate against protected characteristics. Thus, we emphasize that by avoiding indirect discrimination we do *not* mean to suggest removing all variables that may allow implicit inference of discriminatory features from the model (e.g., diabetes), but instead to ensure that these variables, while still remaining within the predictive model, do not act as proxies for discriminatory characteristics.

Finally, we stress that when we talk about "inferring discriminatory features," we do not mean that an insurer necessarily has access to such data. Rather, such inference, as we will show in the sequel, takes place implicitly, via correlation between discriminatory and other features.

We illustrate indirect discrimination in the following example and will come back to this example in Section 2.2, below.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

Example 1. Assume that we have observed a health insurance product and obtained the following claim counts $(n_{i,j})_{i,j=0,1}$ and claim exposures $(e_{i,j})_{i,j=0,1}$:

$n_{i,j}$	Woman	Man	Row total	$e_{i,j}$	Woman	Man	Row total
Smoker	32	4	36	Smoker	133	24	157
Non-smoker	28	48	76	Non-smoker	131	301	432
Column total	60	52	112	Column total	264	325	589

where $i=1$ corresponds to "smoker" and $j=1$ corresponds to "woman". Based on the above contingency tables, we estimate the claim frequencies $\lambda_{i,j}$ by the empirical frequency $\widehat{\lambda}_{i,j}=n_{i,j}/e_{i,j}$. Assume now that gender is considered a discriminatory characteristic. In order to avoid direct discrimination, its explicit influence on the calculated insurance price needs to be removed. The standard way of doing this is to consider the aggregated estimators (row sums) $\widehat{\lambda}_{i,\bullet}=n_{i,\bullet}/e_{i,\bullet}=(n_{i,0}+n_{i,1})/(e_{i,0}+e_{i,1})$. This approach produces, for example, for smokers,

$$\widehat{\lambda}_{1,\bullet}=\frac{36}{157}=0.229.$$

The estimate $\widehat{\lambda}_{1,\bullet}$ (and a premium for smokers based on it), thus, can be calculated by completely ignoring policyholders' gender information. But one can note that an alternative representation of $\widehat{\lambda}_{1,\bullet}$ is

$$\begin{aligned}\widehat{\lambda}_{1,\bullet}&=\widehat{\lambda}_{1,1}\frac{e_{1,1}}{e_{1,1}+e_{1,0}}+\widehat{\lambda}_{1,0}\frac{e_{1,0}}{e_{1,1}+e_{1,0}}\\&=\widehat{\lambda}_{1,1}\widehat{\mathbb{P}}(\text{woman}\mid\text{smoker})+\widehat{\lambda}_{1,0}\widehat{\mathbb{P}}(\text{man}\mid\text{smoker}),\end{aligned}$$

where $\widehat{\mathbb{P}}$ refers to the empirical distribution obtained from the data. Hence, the estimate $\widehat{\lambda}_{1,\bullet}$ not only contains information about the influence of smoking on producing a claim, but via the conditional probabilities $\widehat{\mathbb{P}}(\text{gender}\mid\text{smoking habits})$ also about the propensity of smokers to be female or male. In our case, because smoking habits substantially differ between genders (a smoker is a woman with probability $133/157=85\%$, whereas a non-smoker is a woman with probability $131/432=30\%$). It is indeed the case that the above approach exploits the correlation between gender and smoking habits, which may give rise to indirect discrimination against females in the case that claims frequencies for females are higher than males, as they indeed are here; we come back to this in Example 8, below.

The numbers used in Example 1 are purely illustrative, though we note that the proportion of female smokers has been greater compared to the male population in for example, Sweden during the 2000s. A further discussion of implications of alternative statistical assumptions behind this example is given in Section 2.1, Remark 9. The example illustrates that avoiding direct discrimination does not necessarily entail also avoiding indirect discrimination. Consequently, just ignoring discriminatory features in the calculation of insurance prices does not generally yield discrimination-free prices. Hence, unawareness (or willful ignorance) of discriminatory features is not a solution to the problem of calculating discrimination-free insurance prices.

Finally, we are not arguing in this paper whether certain characteristics ought to be prohibited from a legal or ethical perspective. Indeed, there are varying views on this around the world; for example, gender is a permitted characteristic in insurance pricing in many jurisdictions outside of the EU. Also, there are circumstances, where apparently discriminatory characteristics may be used for pricing, if there is a "legitimate aim"; in the context of EU law see for example, Article 2(b) in European Council (2004). Furthermore, we do not aim to address insurance market and economic implications that may result from legally prohibiting the use of certain characteristics in insurance pricing. An example of such issues is potential "reverse discrimination," meaning that pricing without using all policyholder characteristics may imply (unwanted) cross-subsidies between groups of policyholders, with this in turn leading to adverse selection and other undesirable side effects. Moreover, excluding some rating factors from statistical models typically leads to a decrease of predictive performance.

Our contributions. First, we embed the ideas of direct and indirect discrimination into a mathematical context. The ideas and principles we develop are relevant to all situations where predictors are calculated on the basis of conditional expected values and, hence, they are applicable in all fields where discrimination is an important issue, for example, also in customer credit rating. Second, we give a rigorous probabilistic account of discrimination-free prices and their existence. We propose a simple pricing formula that avoids both potential direct and indirect discrimination. This adjustment will always remove the potential for indirect discrimination from prices, regardless of whether such indirect discrimination is present or not. Furthermore, while the formula only uses nondiscriminatory features as rating factors, it introduces an adjustment, which requires knowledge of policyholders' discriminatory features. Third, we justify discrimination-free prices using tools from causal inference. Fourth, we identify bias in aggregate portfolio prices as an unintended consequence of discrimination-free prices. While prices that can be written as conditional expectations under the physical probability measure naturally lead to an unbiased pricing system on a portfolio level, discrimination-free prices do not generally have this property. Therefore, we propose methods for bias correction. The bias corrections rely on the overall portfolio risk being assessed using all available characteristics, since it is only the step of allocating the overall price to individual contracts that potential discrimination can occur. Fifth, we illustrate how discrimination-free prices can be calculated in practice, using either machine learning algorithms or standard statistical methods like generalized linear models (GLMs).

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

Literature review. Although an issue of key relevance for insurance pricing, until recently relatively little attention has been paid to the issue of discrimination-free pricing within the actuarial literature. In a discussion of the implications of EU gender legislation, Guillén (2012) suggests that covariates highly correlated with gender can be used as proxies by insurance companies, which from our perspective may result in indirect discrimination. Focusing on the case of mortality pricing, Chen and Vigna (2017) criticize the industry practice of deriving unisex life tables by mixing the life tables for each gender on the grounds that this does not respect the principles of actuarial fairness, which is to say that the total unisex premiums charged for the portfolio are not equal to the total premiums charged using gender-specific life tables. They provide alternative approaches without this shortcoming; note that our proposed discrimination-free prices reproduce the pricing formulas of Chen and Vigna (2017). The implications of unisex pricing on insurer capital requirements in the context of Solvency II are examined in Chen *et al*. (2018), and an ALM approach to unisex pricing is taken in Burszas *et al*. (2018), where also the concept of "gender mix risk" is discussed. Market implications of unisex tariffs are discussed in Sass and Seifried (2014), see also De Jong and Ferris (2006) for a discussion of adverse selection stemming from restrictions on risk classification. A recent wide-ranging discussion of several issues connected with the topic of discrimination in insurance is found in Frees and Huang (2021), who also address the issue of indirect discrimination.

The issue of indirect discrimination occurring by ignoring discriminatory covariates has been discussed in Pope and Sydnor (2011) and Kusner *et al*. (2017). The procedure for discrimination-free pricing provided in Pope and Sydnor (2011) is essentially the same as in our proposal; this pricing rule is applied in the context of auto insurance pricing by Aseervatham *et al*. (2016). However, these authors do not provide a probabilistic justification for the prices used nor do they address the critical issue of a potential bias at portfolio level (and associated corrections).

We are aware of relatively few examples of causal inference applied within an insurance context. For renewals of insurance policies, some insurers seek to estimate policyholder demand elasticity by randomly varying renewal prices for a subset of policyholders (i.e., a form of randomized controlled trial is conducted) and estimating the impact on the probability of renewal. Once the demand elasticities have been estimated, a profit maximizing pricing policy can be established in a practice referred to as price optimization, see for example, Krikler *et al*. (2004). Within that context, Guelman and Guillén (2014) apply methods from causal inference to estimate demand elasticity functions from observational data collected by an insurer.

We emphasize that the issues discussed in this paper apply to many other industries; we refer to, for example, Fuster *et al*. (2018) where a credit rating application is considered. Their study focuses on evaluating the differential impact of prediction technologies on ethnic groups, rather than on a mathematical definition of discrimination.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

Organization of the paper. In Section 2, we discuss different kinds of insurance prices, comprising the *best-estimate price*, which considers all available information, the *unawareness price*, which avoids direct discrimination, and the *discrimination-free price*, which avoids both direct and indirect discrimination, whenever the latter exists. In particular, Subsection 2.3 gives mathematical descriptions of direct and indirect discrimination, which are based on a change of probability measure. Special cases of discrimination-free prices can be interpreted in terms of causal inference; this is discussed in Section 3. The bias that discrimination-free prices can induce at portfolio level is discussed in Section 4, along with proposals for bias mitigation. In Section 5, we describe the calculation of discrimination-free prices based on models estimated from data. This is explored in more detail in Section 6, where a numerical example is given, based on a synthetic health insurance portfolio. Concluding remarks are collected in Section 7.

2. Discrimination-free pricing

2.1. Definition of discrimination-free prices

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space with physical probability measure $\mathbb{P}$. For a given portfolio of insurance policies, let $\mathbf{D}$ denote the vector of *discriminatory covariates* (characteristics, features, explanatory variables) of a policyholder, and let $\mathbf{X}$ denote the vector of *nondiscriminatory covariates*. This split into $\mathbf{X}$ and $\mathbf{D}$ is exogenous, provided by, for example, a legislator. Further, we assume that $\mathbf{X}$ and $\mathbf{D}$ are random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$; the randomness of these covariate vectors represents variations between policyholders. A realization of $(\mathbf{X}, \mathbf{D})$ corresponds to choosing an insurance policy at random from the portfolio; a policyholder profile with specific characteristics is obtained by conditioning on $\mathbf{X} = \mathbf{x}, \mathbf{D} = \mathbf{d}$. For simplicity, we denote the marginal and conditional distributions of covariates under $\mathbb{P}$ by $\mathbf{X} \sim \mathbb{P}(\mathbf{x})$, $\mathbf{D} \sim \mathbb{P}(\mathbf{d})$ and $(\mathbf{D} \mid \mathbf{X} = \mathbf{x}) \sim \mathbb{P}(\mathbf{d} \mid \mathbf{x})$, respectively, thus, we use the same letter $\mathbb{P}$ for the (conditional) distribution functions of $\mathbf{X}$ and $\mathbf{D}$.

A policyholder claim is denoted by the random variable Y. The claim Y typically depends on (but is not fully determined by) both the discriminatory covariates $\mathbf{D}$ and the nondiscriminatory ones $\mathbf{X}$. Our aim is to price such a claim Y, with the resulting price being free from direct as well as indirect discrimination (where this exists), according to the arguments of Section 1. A technical description of these concepts will be given in Section 2.3, below.

In the sequel, it will be useful to assume $Y, \mathbf{X}, \mathbf{D} \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. This assumption is not crucial for defining discrimination-free prices, but it will allow us to give more intuitive interpretations in terms of orthogonal projections and minimal distances. Our notion of price will be based on conditional expectations of Y, when conditioning on different subsets of covariates. We first introduce a number of different prices that are important for the subsequent discussions and derivations.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

Definition 2 (best-estimate price). *The best-estimate price for Y w.r.t. $(\boldsymbol{X}, \boldsymbol{D})$ is defined by*

$$\mu(\mathbf{X}, \mathbf{D}) := \mathbb{E}[Y \mid \mathbf{X}, \mathbf{D}].$$

Remark 3.

(a) We call the price $\mu(\boldsymbol{X}, \boldsymbol{D})$ "best-estimate" because it minimizes the $\mathcal{L}^2$-distance of all $(\boldsymbol{X}, \boldsymbol{D})$-measurable prices to Y, that is, $\mu(\boldsymbol{X}, \boldsymbol{D})$ is the orthogonal projection of Y onto the sub-space generated by $(\boldsymbol{X}, \boldsymbol{D})$.

(b) In general, the best-estimate price is not discrimination-free, unless we are in the special case of $\mu(\boldsymbol{X}, \boldsymbol{D}) = \mu(\boldsymbol{X})$, implied by $\boldsymbol{X}$ being independent of $\boldsymbol{D}$.

(c) The best-estimate price is unbiased w.r.t. Y, that is,

$$\mu := \mathbb{E}[Y] = \mathbb{E}\left[\mu(\boldsymbol{X}, \boldsymbol{D})\right];$$

we use the tower property of conditional expectations, see Williams [26, Sec. 9.7]. Unbiasedness is important because it indicates that best-estimate prices achieve on average the correct price level for the portfolio.

An initial attempt at achieving discrimination-free prices arises through simply ignoring discriminatory covariates $\mathbf{D}$.

Definition 4 (unawareness price). *The unawareness price for Y w.r.t. $\boldsymbol{X}$ is defined by*

$$\mu(\mathbf{X}) := \mathbb{E}[Y \mid \mathbf{X}]. \tag{2.1}$$

Remark 5.

(a) As the price $\mu(\boldsymbol{X})$ does not depend explicitly on $\boldsymbol{D}$, it avoids direct discrimination. However, the unawareness price may produce indirect discrimination, as was discussed in Section 1; see also Kusner et al. (2017). Specifically, we can write the unawareness price as

$$\mu(\boldsymbol{X}) = \int_{\boldsymbol{d}} \mu(\boldsymbol{X}, \boldsymbol{d}) \; \mathrm{d}\mathbb{P}(\boldsymbol{d} \mid \boldsymbol{X}). \tag{2.2}$$

The potential for discrimination arises because the conditional probability $\mathbb{P}(\boldsymbol{d} \mid \boldsymbol{X})$ enables inference of discriminatory covariates $\boldsymbol{D}$ from nondiscriminatory ones $\boldsymbol{X}$. We stress that discrimination here is indirect: while $\boldsymbol{D}$ is not directly used in the pricing formula, it is potentially "proxied" by $\boldsymbol{X}$, if statistical dependence between $\boldsymbol{D}$ and $\boldsymbol{X}$ exists. This is precisely the situation discussed in Section 1. Indirect discrimination is avoided in the special case when $\boldsymbol{D}$ and $\boldsymbol{X}$ are independent, since then it holds that $\mathrm{d}\mathbb{P}(\boldsymbol{d} \mid \boldsymbol{X}) = \mathrm{d}\mathbb{P}(\boldsymbol{d})$.

(b) The price $\mu(\boldsymbol{X})$ minimizes the $\mathcal{L}^2$-distance to Y based solely on $\boldsymbol{X}$, that is, it is the best price w.r.t. information $\boldsymbol{X}$. At the same time, the price $\mu(\boldsymbol{X})$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

also minimizes the $\mathcal{L}^2$-distance to $\mu(\boldsymbol{X}, \boldsymbol{D})$, by a simple application of the Pythagorean theorem. Note that

$$||\mu(\boldsymbol{X}) - \mu(\boldsymbol{X}, \boldsymbol{D})||_2^2 = \mathbb{E}[\mathrm{Var}(\mu(\boldsymbol{X}, \boldsymbol{D}) \mid \boldsymbol{X})],$$

which intuitively should decrease with increasing dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$. Hence, the quality in the approximation of $\mu(\boldsymbol{X}, \boldsymbol{D})$ using $\mu(\boldsymbol{X})$ should be good if $\boldsymbol{D}$ essentially is a deterministic function of $\boldsymbol{X}$, that is, if the nondiscriminatory covariates $\boldsymbol{X}$ allow us to almost perfectly infer the discriminatory covariates $\boldsymbol{D}$.

(c) The unawareness price is unbiased, since

$$\mu = \mathbb{E}[Y] = \mathbb{E}\left[\mu(\boldsymbol{X})\right].$$

We now propose a price that is free of both direct and indirect discrimination.

Definition 6 (discrimination-free price). *A discrimination-free price for Y w.r.t. $\boldsymbol{X}$ is defined by*

$$h^*(\boldsymbol{X}) := \int_{\boldsymbol{d}} \mu(\boldsymbol{X}, \boldsymbol{d}) \ \mathrm{d}\mathbb{P}^*(\boldsymbol{d}), \tag{2.3}$$

where the distribution $\mathbb{P}^*(\boldsymbol{d})$ is defined on the same range as the marginal distribution of the discriminatory variables $\boldsymbol{D} \sim \mathbb{P}(\boldsymbol{d})$.

Remark 7.

(a) The discrimination-free price (2.3) is obtained by averaging best-estimate prices over discriminatory covariates, using a (potentially arbitrary) marginal distribution $\mathbb{P}^*(\boldsymbol{d})$. The crucial step here is the imposed marginalization w.r.t. $\boldsymbol{D}$, rather than the specific choice of $\mathbb{P}^*(\boldsymbol{d})$ (which can be $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$). Given that the price $h^*(\boldsymbol{X})$ does not explicitly depend on $\boldsymbol{D}$, it is obviously free from direct discrimination. We argue that the averaging construction proposed in (2.3) also removes all potential indirect discrimination. While (2.3) appears similar to (2.2), there is a key difference: discrimination-free prices do not in any way depend on the conditional distribution $\mathbb{P}(\boldsymbol{d} \mid \boldsymbol{X})$ – hence they do not use any inference of discriminatory covariates from nondiscriminatory ones. This will be further discussed in Section 2.3 and verified in the case study of Section 6. In the special case of $\boldsymbol{X}$ and $\boldsymbol{D}$ being independent and $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$, it follows that $h^*(\boldsymbol{X}) = \mu(\boldsymbol{X})$.

(b) Definition 6 is designed to remove the possible explanatory power that $\boldsymbol{X}$ may have for $\boldsymbol{D}$; it does not assume independence between $\boldsymbol{X}$ and $\boldsymbol{D}$ in the given portfolio. This point will be made more precise in Section 2.3, and in Section 2.4 we discuss existence of discrimination-free prices as well as alternative interpretations of $h^*(\boldsymbol{X})$.

(c) Definition 6 can also be motivated by arguments from causal inference. Specifically, formulas like (2.3) are used to quantify the direct causal

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

effect of $\boldsymbol{X}$ on Y; we discuss this in more detail in Section 3, below. We stress that although causal inference can in many situations serve as an alternative motivation of discrimination-free prices, the reasoning behind our Definition 6 does not rely on any causal assumptions. Further discussions of this are provided in Section 3. Furthermore, formula (2.3) using the special choice $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$ corresponds precisely to the partial dependence plot (PDP) introduced by Friedman (2001), see also Zhao and Hastie (2021).

(d) *Prices obtained using (2.3) will in general not be unbiased, since*

$$\mu = \mathbb{E}[Y] \neq \mathbb{E}[h^*(\boldsymbol{X})] = \int_{\boldsymbol{x},\boldsymbol{d}} \mu(\boldsymbol{x}, \boldsymbol{d}) \mathrm{d}\mathbb{P}^*(\boldsymbol{d}) \mathrm{d}\mathbb{P}(\boldsymbol{x}), \tag{2.4}$$

even for the special choice $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$. This observation motivates portfolio level price adjustments, which will be discussed in Section 4. We note that, in actuarial practice, such a bias is not necessarily a problem, as insurers are primarily interested in the relativities between different policyholders, which can be used to differentiate a baseline premium of the overall portfolio costs to individual policyholders. Still, a poor allocation principle may result in adverse selection.

(e) *Note that, given the potential arbitrariness of $\mathbb{P}^*$, calculation of discrimination-free prices only requires knowledge of the mapping $(\boldsymbol{x}, \boldsymbol{d}) \mapsto \mu(\boldsymbol{x}, \boldsymbol{d})$, where $\mu(\boldsymbol{x}, \boldsymbol{d})$ may be an (algorithmically derived implicit) regression function. Nevertheless, as pointed out in the previous remark, if one aims to correct a potential bias of $h^*(\boldsymbol{X})$, it is necessary to perform modeling and model calibration under the "real-world" probability measure $\mathbb{P}$.*

(f) *Given the construction (2.3), $\mathbb{P}^*(\boldsymbol{d})$ may be inferred from comparing best-estimate prices $\mu(\boldsymbol{X}, \boldsymbol{D})$ and observed discrimination-free prices $h^*(\boldsymbol{X})$.*

2.2. Choice of weighting distributions for discriminatory covariates

From Definition 6, it follows that the distribution $\mathbb{P}^*(\mathbf{d})$ can be chosen rather freely. A simple choice is $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$, that is, average in (2.3) w.r.t. the marginal distribution of the discriminatory characteristics in the portfolio. This choice is supported by causal inference arguments in Section 3. We denote this special case by

$$h(\mathbf{X}) := \int_{\mathbf{d}} \mu(\mathbf{X}, \mathbf{d}) \ \mathrm{d}\mathbb{P}(\mathbf{d}). \tag{2.5}$$

We illustrate how $h(\mathbf{X})$ is evaluated in the context of Example 1.

Example 8. In Example 1, we argued that aggregated estimators (row sums) $\widehat{\lambda}_{i,\bullet}$ are discriminatory because gender can be inferred from smoking habits.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

The price $h(\mathbf{X})$ removes this effect by replacing the conditional probability $\mathbb{P}$(gender | smoking habits) by $\mathbb{P}$(gender). This implies that the frequency estimate for smokers $\widehat{\lambda}_{1,\bullet}$ is replaced by

$$\begin{aligned}\widetilde{\lambda}_{1,\bullet} &= \widehat{\lambda}_{1,1}\widehat{\mathbb{P}}(\text{woman}) + \widehat{\lambda}_{1,0}\widehat{\mathbb{P}}(\text{man}) \\ &= \frac{32}{133}\cdot\frac{264}{589} + \frac{4}{24}\cdot\frac{325}{589} \\ &= 0.200 < 0.229 = \widehat{\lambda}_{1,\bullet}.\end{aligned} \tag{2.6}$$

Similarly, for non-smokers

$$\widetilde{\lambda}_{0,\bullet} = \widehat{\lambda}_{0,1}\widehat{\mathbb{P}}(\text{woman}) + \widehat{\lambda}_{0,0}\widehat{\mathbb{P}}(\text{man}) = 0.184. \tag{2.7}$$

We demonstrate the potential portfolio bias that discrimination-free prices induce. The total cost of the portfolio, under best-estimate prices, is equal to the observed total claim of 112. For discrimination-free prices, the total cost is given by

$$\widetilde{\lambda}_{1,\bullet}(e_{1,1} + e_{1,0}) + \widetilde{\lambda}_{0,\bullet}(e_{0,1} + e_{0,0}) = 110.77 < 112.$$

This indicates that the discrimination-free price $h(\mathbf{X})$ leads to an under-pricing of the overall portfolio in the present situation.

Recall that there is some flexibility in the selection of $\mathbb{P}^*(\mathbf{d})$. In this simple example, with $\mathbf{D}$ being a binary classification variable, we can directly choose $\mathbb{P}^*$(woman) and $\mathbb{P}^*$(man) in a way that eliminates the portfolio bias. Specifically, we set

$$\widetilde{\lambda}^*_{i,\bullet} = \widehat{\lambda}_{i,1}\mathbb{P}^*(\text{woman}) + \widehat{\lambda}_{i,0}\mathbb{P}^*(\text{man}), \qquad \text{for } i = 0, 1,$$

and require for the resulting overall portfolio price that it holds

$$\widetilde{\lambda}^*_{1,\bullet}(e_{1,1} + e_{1,0}) + \widetilde{\lambda}^*_{0,\bullet}(e_{0,1} + e_{0,0}) = 112.$$

The resulting choice is $\mathbb{P}^*(\text{woman}) = 48.3\% > 44.8\% = \mathbb{P}(\text{woman})$.

Finally, we note that in this example, switching to discrimination-free prices leads to a reduction in the share of the portfolio costs covered by women. Women cause $60/112 = 53.6\%$ of the total costs which is exactly the share of the total costs that women have to pay under best-estimate pricing (assuming that the prices coincide with the claims caused). If we use the unawareness price by simply dropping the gender variable, women cover 47.8% of the total costs. If we charge the discrimination-free price (2.6)-(2.7), women cover 45.7% of all costs, thus, less than under the unawareness price. This exactly reflects the potential for indirect discrimination in the unawareness price: women have on average higher costs than men, and the allocation of these excess costs is bigger to the sub-population where women are more prevalent compared to the population distribution $\mathbb{P}(\mathbf{d})$, that is, we learn $\mathbf{D}$ from $\mathbf{X}$ through the portfolio distribution.

Remark 9. *While in Examples 1 and 8, the potential indirect discrimination was against women, one can easily swap the "woman" and "man" labels, so that such*

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

indirect discrimination is against males. This indicates that the notion of discrimination used here (as well as the proposed pricing adjustment) does not reflect (or indeed seek to correct for) historical or current injustices. A more subtle impact arises if, ceteris paribus, the frequency of smokers in the female population was actually lower than that for males. In such a case, unawareness prices would actually understate the impact of smoking, as this would be "masked" by males' otherwise lower propensity to claim; on the contrary, discrimination-free prices would become more sensitive with respect to the specific risk posed by smoking. This idea is further developed in the detailed numerical example presented later in the paper; see last paragraph of Section 6.1 and Figure 3.

Furthermore, it is useful to consider the extrema of discrimination-free prices. Consider the following prices:

$$h^{(+)}(\mathbf{X}) := \sup_{\mathbb{P}^*} \int_{\mathbf{d}} \mu(\mathbf{X}, \mathbf{d}) \, d\mathbb{P}^*(\mathbf{d}),$$
$$h^{(-)}(\mathbf{X}) := \inf_{\mathbb{P}^*} \int_{\mathbf{d}} \mu(\mathbf{X}, \mathbf{d}) \, d\mathbb{P}^*(\mathbf{d}).$$

Here, $h^{(+)}(\mathbf{X})$ and $h^{(-)}(\mathbf{X})$ correspond to the essential supremum and infimum over $\mathbf{d}$ in the range of $\mathbf{D}$, respectively. Thus, for nondiscriminatory covariates $\mathbf{X} = \mathbf{x}$, this immediately gives us

$$h^{(-)}(\mathbf{x}) \leq h^*(\mathbf{x}), h(\mathbf{x}), \mu(\mathbf{x}) \leq h^{(+)}(\mathbf{x}).$$

Moreover, for the bias property we get the following relationship

$$\int_{\mathbf{x}} h^{(-)}(\mathbf{x}) d\mathbb{P}(\mathbf{x}) \leq \mathbb{E}[h^*(\mathbf{X})], \mu \leq \int_{\mathbf{x}} h^{(+)}(\mathbf{x}) d\mathbb{P}(\mathbf{x}).$$

By definition $h^{(+)}(\mathbf{x})$ corresponds to the "worst" (or most "prudent") price and has been discussed in the context of unisex pricing in Chen and Vigna (2017).

As seen in Example 8, the discrimination-free price (2.3) is generally biased. An alternative possibility for the choice of $\mathbb{P}^*(\mathbf{d})$ is to additionally require unbiasedness in (2.4). In the simple case of a binary discriminatory covariate like gender in Example 8, this reduced to choosing a suitable $\mathbb{P}^*$(woman). A more general construction of unbiased prices via choices of $\mathbb{P}^*(\mathbf{d})$ is presented in Section 4.

A special case corresponds to an additive best-estimate price, in the sense that $\mu(\mathbf{X}, \mathbf{D}) = \mu_1(\mathbf{X}) + \mu_2(\mathbf{D})$. Then, the simple choice $\mathbb{P}^*(\mathbf{d}) = \mathbb{P}(\mathbf{d})$ is appealing, as it provides an unbiased price. Note that

$$h(\mathbf{X}) = \int_{\mathbf{d}} \mu_1(\mathbf{X}) \, d\mathbb{P}(\mathbf{d}) + \int_{\mathbf{d}} \mu_2(\mathbf{d}) \, d\mathbb{P}(\mathbf{d}) = \mu_1(\mathbf{X}) + \mathbb{E}[\mu_2(\mathbf{D})],$$

which implies

$$\mathbb{E}[h(\mathbf{X})] = \mathbb{E}[\mu_1(\mathbf{X})] + \mathbb{E}[\mu_2(\mathbf{D})] = \mathbb{E}[\mu(\mathbf{X}, \mathbf{D})] = \mu.$$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

2.3. Revisiting direct and indirect discrimination

In this section, following the development of our ideas so far, we provide more technical definitions of prices that avoid direct and indirect discrimination, where the latter may exist.

Choose an arbitrary probability measure $\mathbb{P}^*$ on the measurable space $(\Omega, \mathcal{F})$ such that $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^*)$. Choose a (sub-)vector $\mathbf{Z}$ of the covariates $(\mathbf{X}, \mathbf{D})$ and define the $(\mathbb{P}^*, \mathbf{Z})$-conditional-expectation price by

$$\mu^*(\mathbf{Z}) := \mathbb{E}^*[Y \mid \mathbf{Z}],$$

where $\mathbb{E}^*$ denotes the expectation under $\mathbb{P}^*$.

Definition 10. *A price avoids direct discrimination, if it can be written as*

$$\mu^*(\boldsymbol{Z}) = \mathbb{E}^*[Y \mid \boldsymbol{Z}],$$

where $\boldsymbol{Z}$ is $\sigma(\boldsymbol{X})$-measurable, and where the expectation is taken w.r.t. a probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F})$ such that $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^*)$.

Remark 11.

(a) Definition 10 says that a price avoids direct discrimination if it can be written as a measurable function of the nondiscriminatory covariates $\boldsymbol{X}$. For $\boldsymbol{Z} = \boldsymbol{X}$ we receive maximal use of nondiscriminatory information (relative to $\mathbb{P}^*$), therefore, we typically work with $\boldsymbol{Z} = \boldsymbol{X}$.

(b) The choice $\mathbb{P}^* = \mathbb{P}$ (and $\boldsymbol{Z} = \boldsymbol{X}$) provides the unawareness price $\mu(\boldsymbol{X})$ of Definition 4 which, thus, avoids direct discrimination.

(c) Importantly, under the choice $\mathbb{P}^* = \mathbb{P}$, the unawareness price $\mu(\boldsymbol{X})$ can be calculated without explicit knowledge of $\mu(\boldsymbol{X}, \boldsymbol{D})$ – hence it does not require collection of discriminatory policyholder information. This also applies if we need to estimate $\mu(\boldsymbol{X})$ from data, see (5.3) below.

Now, indirect discrimination can be defined.

Definition 12. *A price $\mu^*(\boldsymbol{Z})$ that avoids direct discrimination is said to avoid indirect discrimination if $\boldsymbol{Z}$ and $\boldsymbol{D}$ are independent under $\mathbb{P}^*$.*

Independence under $\mathbb{P}^*$ effects the decoupling of discriminatory covariates from nondiscriminatory ones, for specific policyholders. Thus, according to Definition 12, a price that avoids indirect discrimination satisfies

$$\mu^*(\mathbf{Z}) = \int_{\mathbf{d}} \mu^*(\mathbf{Z}, \mathbf{d})\ d\mathbb{P}^*(\mathbf{d} \mid \mathbf{Z}) = \int_{\mathbf{d}} \mu^*(\mathbf{Z}, \mathbf{d})\ d\mathbb{P}^*(\mathbf{d}), \tag{2.8}$$

where $\mu^*(\mathbf{Z}, \mathbf{d}) = \mathbb{E}^*[Y \mid \mathbf{Z}, \mathbf{D} = \mathbf{d}]$.

Remark 13.

(a) From Definition 12, it is clear that avoiding indirect discrimination requires avoiding direct discrimination. As indirect discrimination relates to covariates in $\boldsymbol{X}$ acting as proxies for (elements of) $\boldsymbol{D}$, it is not

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

meaningful to talk about indirect discrimination, when $\boldsymbol{D}$ *is used directly in pricing.*

(b) *The independence in Definition 12 is an artifice of the introduced probability measure* $\mathbb{P}^*$ *under which insurance is priced and does not generally reflect the actual observed dependence between* $\boldsymbol{X}$ *and* $\boldsymbol{D}$.

(c) *For* $\boldsymbol{Z}=\boldsymbol{X}$, *the calculation that avoids indirect discrimination is based on the knowledge of* $\mu^*(\boldsymbol{X},\boldsymbol{D})$, *see (2.8) – hence it requires collection of discriminatory policyholder information. In fact, one of the most critical problems in practice is that discriminatory information is often incomplete, for example, about ethnicity, which may result in indirect discrimination.*

(d) *In statistical applications we usually use the conditional probability* $\mathbb{P}(y\mid \boldsymbol{X},\boldsymbol{D})$ *to model a claim* Y, *given the covariates* $(\boldsymbol{X},\boldsymbol{D})$. *The reason for this choice is that* Y, *given* $(\boldsymbol{X},\boldsymbol{D})$, *is observed under the real world measure* $\mathbb{P}$, *which allows for direct estimation of the regression function, see Section 5 below,*

$$(\boldsymbol{x},\boldsymbol{d})\mapsto \mu(\boldsymbol{x},\boldsymbol{d}).$$

We could choose the measure $\mathbb{P}^*$ *in a way that preserves the (causal) structure of how the covariates impact the response, that is, let* $\mathbb{P}^*(y\mid \boldsymbol{x},\boldsymbol{d})=\mathbb{P}(y\mid \boldsymbol{x},\boldsymbol{d})$. *This then motivates the choice*

$$\mathrm{d}\mathbb{P}^*(y,\boldsymbol{x},\boldsymbol{d})=\mathrm{d}\mathbb{P}(y\mid \boldsymbol{x},\boldsymbol{d})\ \mathrm{d}\mathbb{P}^*(\boldsymbol{x})\ \mathrm{d}\mathbb{P}^*(\boldsymbol{d}),$$

for $\boldsymbol{Z}=\boldsymbol{X}$ *in Definition 12. In view of (2.8), this results in the discrimination-free price*

$$\mu^*(\boldsymbol{X})=\int_{\boldsymbol{d}}\mu(\boldsymbol{X},\boldsymbol{d})\ \mathrm{d}\mathbb{P}^*(\boldsymbol{d}\mid \boldsymbol{X})=\int_{\boldsymbol{d}}\mu(\boldsymbol{X},\boldsymbol{d})\ \mathrm{d}\mathbb{P}^*(\boldsymbol{d})=h^*(\boldsymbol{X}).$$

Thus, the discrimination-free price of Definition 6 does neither allow for potential direct nor for indirect discrimination.

(e) *Linking to Remark 7(e), in practice, we need to know (calibrate under) the real world measure* $\mathbb{P}$ *in order to study unbiasedness w.r.t.* $\mu=\mathbb{E}[Y]$. *Since the actual portfolio that we hold is described by* $\boldsymbol{Z}\sim\mathbb{P}(\boldsymbol{z})$, *we need to average discrimination-free prices* $\mu^*(\boldsymbol{Z})$ *w.r.t. the same population* $\mathbb{P}(\boldsymbol{z})$ *to see whether we receive unbiasedness of discrimination-free prices on the actual portfolio.*

2.4. Existence of discrimination-free prices

We have not yet discussed existence of discrimination-free prices according to Definition 6 and the possibility of avoiding indirect discrimination according to Definition 12. This is done in the present section.

We emphasize that properties of available data (and the related statistical models) play a crucial role in our considerations:

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

- Indirect discrimination may be the result of incomplete discriminatory information, see Remark 13(c).
- Indirect discrimination may be the result of nonexistent or insufficient information of certain parts of the population.

In this section, we discuss the second item that can enter in different ways. A first one is that not all parts of the population are equally well represented in the development of the statistical model. For instance, there is research in image recognition to discover malignant melanoma (skin cancer). If this research is mainly based on images of people with light complexion, the corresponding model will likely fail to discover malignant melanoma for people with dark complexion. This is a form of discrimination resulting from *insufficient data* of certain parts of the population. In our situation, this may result in poor best-estimate prices $\mu(\mathbf{X}, \mathbf{D})$ for certain covariate combinations. Note that the quality of the estimation of best-estimate prices directly impacts discrimination-free prices.

In the current section, we rather focus on *nonexistent data* of certain parts of the population. The meaning and implications of nonexistent data are going to be discussed in more detail. We start with an example. Assume that the discriminatory covariates **D** correspond to gender and the nondiscriminatory ones **X** to education. Education could be in the ordinal form "secondary school degree," "high school degree" or "university degree," but information about education could also be received in the following categorical form "Catholic college degree," "public college degree" or "girls college degree." Per definition the last label "girls college degree" contains as only gender "female". This implies that

$$\mathbb{P}(\mathbf{X} = \text{girls college degree}, \mathbf{D} = \text{man}) = 0,$$

thus, the event $A = \{\mathbf{X} = \text{girls college degree}, \mathbf{D} = \text{man}\} \in \mathcal{F}$ is a null set w.r.t. $\mathbb{P}$. In many cases, we do not model responses Y on null sets. Therefore, neither Y on A may be specified in our model nor the conditional expectation $\mu(\text{girls college degree, man}) = \mathbb{E}[Y \mid A]$ may be determined. But this implies that we cannot evaluate the discrimination-free price

$$h^*(\mathbf{X}) = \int_{\mathbf{d}} \mu(\mathbf{X}, \mathbf{d})\ d\mathbb{P}^*(\mathbf{d}),$$

if $\mathbb{P}^*(\mathbf{d})$ has positive probability mass on both genders. In the current situation, the problem may be solved by setting $\mathbb{P}^*(\mathbf{D} = \text{woman}) = 1$ which gives the discrimination-free price $h^*(\mathbf{X}) = \mu(\mathbf{X}, \text{woman})$.

If the education information **X** has an additional level "boys college degree", the above solution will not work because we have a second $\mathbb{P}$-null set $B = \{\mathbf{X} = \text{boys college degree}, \mathbf{D} = \text{woman}\} \in \mathcal{F}$ which makes it impossible to choose a distribution $\mathbb{P}^*(\mathbf{d})$ such that the discrimination-free price $h^*(\mathbf{X})$ is well-defined.

The simple solution to this problem is to drop the education information, that is, choose a smaller covariate set. This is equivalent to choosing a true

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

subset $\mathbf{Z}$ of $\mathbf{X}$ in Definition 12. In practice, we often try to inter- or extrapolate the model assumptions for Y. This is reasonable if unavailable information corresponds to numerical variables (and responses have some smoothness in these covariates). In certain cases, it may also be justified for categorical variables by, for example, postulating a multiplicative influence structure of covariates, say, women are $x\%$ better than men regardless of the attended college. This is similar to a GLM approach where gender may be reflected by a single parameter on the canonical scale. In our situation such an assumption can be made, but it cannot be verified because of a missing control group.

Proposition 14. *Assume there exists a product measure* $\mathbb{P}^*(\boldsymbol{x})\mathbb{P}^*(\boldsymbol{d})$ *on* $(\Omega, \mathcal{F})$ *which is absolutely continuous w.r.t. the probability measure* $\mathbb{P}(\boldsymbol{x}, \boldsymbol{d})$ *of the covariates* $(\boldsymbol{X}, \boldsymbol{D})$*. Then, there exists a price* $\mu^*(\boldsymbol{X})$ *that avoids indirect discrimination.*

Proof. Absolute continuity implies that every $\mathbb{P}(\mathbf{x}, \mathbf{d})$-null set is also a $\mathbb{P}^*(\mathbf{x})\mathbb{P}^*(\mathbf{d})$-null set. Therefore, $\mu(\mathbf{X}, \mathbf{D})$ is well-defined on all sets where $(\mathbf{X}, \mathbf{D})$ has positive $\mathbb{P}^*(\mathbf{x})\mathbb{P}^*(\mathbf{d})$-probability mass. Since the latter is a product measure, we can calculate the discrimination-free price $h^*(\mathbf{X})$ by integrating $\mu(\mathbf{X}, \mathbf{d})$ over $d\mathbb{P}^*(\mathbf{d} \mid \mathbf{X}) = d\mathbb{P}^*(\mathbf{d})$, see also (2.8). This completes the proof. □

3. Causal inference and discrimination

The purpose of this section is to discuss the discrimination-free prices of Definition 6 in a causal inference setting. Discrimination-free prices given by Definition 6 hold without recourse to any causal relationships between variables. Nonetheless, there is a nice motivation of discrimination-free pricing in a causal inference context which provides additional insight. We give these arguments in a pedagogical and somewhat informal way; for a rigorous treatment we refer to Hernán and Robins (2020), Pearl (2009) and Pearl *et al.* (2016), Ch.3.1).

The starting point of causal inference is a hypothesis of variable relationships, which may be described in terms of a directed graph $\mathfrak{G}$. The graph $\mathfrak{G}$ consists of a set of *nodes* corresponding to the different variables and *directed edges* – “arrows” – indicating directions of potential influence between the variables. This informal definition is most easily understood by an example such as the one given in Figure 1 (left), involving the variables $(Y, \mathbf{X}, \mathbf{D})$ introduced above in the context of insurance pricing. The graph $\mathfrak{G}$ in Figure 1 (left) is an example of a directed acyclic graph (DAG), meaning that the graph does not contain any loops (for a precise definition, see [21, Chapter 1.4]). Figure 1 (left) corresponds to a situation where the discriminatory characteristics $\mathbf{D}$ may influence Y both directly, but also indirectly via $\mathbf{X}$.

Figure 1 (left) already captures a large number of realistic insurance pricing situations. For instance, in view of Example 1, we may identify smoking habits by $\mathbf{X}$ and the gender by the discriminatory factors $\mathbf{D}$. Differences in smoking habits between men and women can be expressed by a directed edge

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

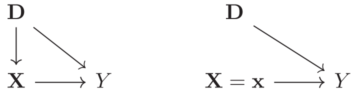

FIGURE 1: (left) Causal diagram described by $\mathfrak{G}$; (right) causal diagram altered according to the intervention $\mathbf{X}=\mathbf{x}$.

$\mathbf{D}\to\mathbf{X}$, while intrinsic differences between men and women when it comes to health outcomes are described by $\mathbf{D}\to Y$. Moreover, smoking in itself may cause health problems, $\mathbf{X}\to Y$, this is exactly expressed by the directed edges in Figure 1 (left).

Since the directed edges in the DAG $\mathfrak{G}$ do not act fully deterministically, we endow $\mathfrak{G}$ with a probability measure $\mathbb{P}$ that describes the randomness involved. Here, we consider a Markovian measure, which, colloquially speaking, means that all nodes in Figure 1 (left) are complemented with independent noisy background variables (Pearl *et al.*, 2016, Chapter 3.2.1). In such a Markovian setting, let, for a general DAG $\mathfrak{G}$, $\mathbf{Z}=(Z_1,\ldots,Z_p)$ be the vector containing all variables (e.g., $\mathbf{Z}=(Y,\mathbf{X},\mathbf{D})$) and let $\mathbf{V}_i$ denote the set of "parent" variables of Z_i (that have a directed edge attached pointing directly to Z_i). Furthermore, in this section, we denote by $\mathbb{p}(\mathbf{z})$ the probability density or mass function of $\mathbf{Z}$. Then, on the Markovian DAG, it holds that (see e.g., Theorem 1 in Pearl (2009))

$$\mathbb{p}(z_1,\ldots,z_p)=\prod_i \mathbb{p}(z_i\mid \mathbf{v}_i). \tag{3.1}$$

In the simple example of Figure 1 (left), identity (3.1) leads to decomposition

$$\mathbb{p}(y,\boldsymbol{x},\mathbf{d})=\mathbb{p}(y\mid \boldsymbol{x},\mathbf{d})\mathbb{p}(\boldsymbol{x}\mid\mathbf{d})\ \ \mathbb{p}(\mathbf{d}),$$

which, of course, is nothing but Bayes' rule.

With this modeling setup in place, one way to approach nondiscriminatory pricing is to ask the following:

> *Given that a policyholder has the set of characteristics* $\mathbf{X}=\mathbf{x}$, *what is the expected value of Y, after removing all causal, direct or indirect, effects of discriminatory covariates* $\mathbf{D}$*?*

In the context of causal inference, to answer such a question, we need to carry out a so-called *intervention* $\mathbf{X}=\mathbf{x}$. An intervention amounts to "fixing" $\mathbf{X}$ to the particular value $\mathbf{x}$, which leads to impacts of $\mathbf{X}$ on Y only via directed edges starting in $\mathbf{X}$, and by removing all possible impacts on $\mathbf{X}$ from other variables. That is, the intervention will be executed without any influence from states of the other variables. This operation is illustrated on the right-hand side of Figure 1, where we remove all directed edges to $\mathbf{X}$ and set the value of $\mathbf{X}$ to $\mathbf{x}$. Removing any potential edge from $\mathbf{D}$ to $\mathbf{X}$ allows us to consider only the (direct) causal effect of setting $\mathbf{X}=\mathbf{x}$ on Y. This operation is intrinsically different to conditioning. When conditioning on $\mathbf{X}=\mathbf{x}$, the distribution of $\mathbf{D}$ is generally affected; but in the modified graph on the right-hand side of Figure 1, changes in $\mathbf{x}$ do not influence $\mathbf{D}$ and vice versa. This is precisely the desired effect of removing the implicit inference of discriminatory covariates from nondiscriminatory ones, in correspondence to Remark 13(b). The above intervention of removing all directed edges to $\mathbf{X}$ and of fixing $\mathbf{X}=\mathbf{x}$ is denoted by the so-called do-operator "do($\mathbf{X}=\mathbf{x}$)" in causal inference (Pearl *et al.*, 2016, Chapter 3.2.1).

In order to formalize the intervention do ($\mathbf{X}=\mathbf{x}$), let $\mathfrak{G}^*$ denote the modified DAG where all edges pointing to $\mathbf{X}$ have been removed, for example, as on the right-hand side of Figure 1. Next, we need to specify the probability measure operating on the graph $\mathfrak{G}^*$, which will *not* be the conditional measure $\mathbb{P}(\mathbf{z} \mid \mathbf{x})$. To that effect, let $\mathcal{X}$ denote the indices in $\mathbf{Z}$ corresponding to $\mathbf{X}$ in a Markov DAG $\mathfrak{G}$, and let $\mathbf{Z}^*$ be the vector consisting of all Z_i, $i \notin \mathcal{X}$. Then, on $\mathfrak{G}^*$, using (3.1), $\mathbb{p}_{\mathfrak{G}^*}$ must satisfy:

$$\mathbb{p}_{\mathfrak{G}^*}(\mathbf{z}^*, \mathbf{x}) = \mathbb{p}_{\mathfrak{G}^*}(\mathbf{x}) \prod_{i \notin \mathcal{X}} \mathbb{p}_{\mathfrak{G}^*}(z_i \mid \mathbf{v}_i), \tag{3.2}$$

since, on $\mathfrak{G}^*$, the influence from parents of $\mathbf{X}$ has been removed. In particular, it follows that

$$\mathbb{p}_{\mathfrak{G}^*}(\mathbf{z}^* \mid \mathbf{x}) = \prod_{i \notin \mathcal{X}} \mathbb{p}_{\mathfrak{G}^*}(z_i \mid \mathbf{v}_i).$$

Furthermore, since $\mathfrak{G}^*$ is a modified version of $\mathfrak{G}$ where only those edges pointing to $\mathbf{X}$ have been removed, it holds that $\mathbb{p}_{\mathfrak{G}^*}(z_i \mid \mathbf{v}_i) = \mathbb{p}(z_i \mid \mathbf{v}_i)$, $i \notin \mathcal{X}$, that is, the remaining causal relations have not been modified. Putting everything together, we arrive at the following definition of do ($\mathbf{X}=\mathbf{x}$):

$$\mathbb{p}(\mathbf{z}^* \mid \text{do}\,(\mathbf{X}=\mathbf{x})) := \mathbb{p}_{\mathfrak{G}^*}(\mathbf{z}^* \mid \mathbf{x}) = \prod_{i \notin \mathcal{X}} \mathbb{p}(z_i \mid \mathbf{v}_i), \tag{3.3}$$

which is known as the *truncated factorization formula*, see for example, Corollary 1 in Pearl (2009).

Returning to our example, set $\mathbf{Z}^* = (Y, \mathbf{D})$. From (3.3) it directly follows that (since, in the modified graph $\mathfrak{G}^*$, $\mathbf{D}$ has no parents)

$$\mathbb{p}(y, \mathbf{d} \mid \text{do}\,(\mathbf{X}=\mathbf{x})) = \mathbb{p}(y \mid \mathbf{d}, \mathbf{x})\ \mathbb{p}(\mathbf{d}).$$

After marginalizing over $\mathbf{d}$, we then obtain the distribution of Y following the intervention do ($\mathbf{X}=\mathbf{x}$):

$$\mathbb{P}(y \mid \text{do}(\mathbf{X}=\mathbf{x})) = \int_{\mathbf{d}} \mathbb{P}(y \mid \mathbf{x}, \mathbf{d})\ d\mathbb{P}(\mathbf{d}). \tag{3.4}$$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

Finally, one can define a price that only takes into account the causal effect of $\mathbf{X}$ on Y by considering $\mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}=\mathbf{x})]$, where the expectation is calculated with respect to $\mathbb{P}(y \mid \mathrm{do}(\mathbf{X}=\mathbf{x}))$. The next result is a direct consequence.

Proposition 15. *Consider the Markovian DAG $(\mathfrak{G}, \mathbb{P})$ defined by the left-hand side of Figure 1. It then holds that*

$$\mathbb{E}[Y \mid \mathrm{do}(\boldsymbol{X}=\boldsymbol{x})] = \int_{\boldsymbol{d}} \mu(\boldsymbol{x}, \boldsymbol{d}) \;\, \mathrm{d}\mathbb{P}(\boldsymbol{d}) = h(\boldsymbol{x}),$$

where $h(\boldsymbol{x})$ was defined by (2.5).

Remark 16.

(a) *Proposition 15 justifies the discrimination-free price $h(\boldsymbol{X})$ of Equation (2.5) under specific Markovian DAG assumptions, motivating the choice $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$ in Definition 6. While we find the assumptions underlying Proposition 15 reasonable in an insurance context, violating those assumptions will undermine the causal interpretation of discrimination-free prices. Nonetheless, these assumptions are not needed in order for $h(\boldsymbol{X})$ to produce discrimination-free prices, in the spirit of Section 2.3, which "breaks" the statistical dependence between $\boldsymbol{X}$ and $\boldsymbol{D}$. However, it is interesting to see that our discrimination-free pricing framework exactly corresponds to the do-operator "$\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})$" in the causal inference setting of Figure 1.*

(b) *It is possible to extend the covariate relations described by Figure 1 to more general situations, for instance, by including unmeasured characteristics (latent variable) $\boldsymbol{U}$. For ways to deal with these more general situations, we refer to Pearl et al. (2016) and Lauritzen (1996, Chapter 3.2.2).*

4. Attribution of total portfolio premium to individual policies

The difficulty that we still have to deal with is that, in general, a discrimination-free price has a bias, see (2.4) and Example 8. This bias needs to be corrected because otherwise the premium for the entire portfolio may not be at the appropriate level. There is no canonical way of correcting for this potential bias; moreover, the requirement that the bias correction should be discrimination-free excludes complex cost allocation mechanisms.

The portfolio bias of the $\mathbb{P}^*$-discrimination-free price is defined by

$$B^* := \mu - \mathbb{E}[h^*(\mathbf{X})] = \mathbb{E}[Y] - \int_{\mathbf{x},\mathbf{d}} \mu(\mathbf{x}, \mathbf{d}) \;\, \mathrm{d}\mathbb{P}^*(\mathbf{d}) \;\, \mathrm{d}\mathbb{P}(\mathbf{x}).$$

Simple bias corrections arise from taking rather different positions. An egalitarian position is taken by distributing the portfolio bias B^* uniformly across

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

the entire portfolio, regardless of any nondiscriminatory covariates $\mathbf{X}$. This motivates the *uniformly adjusted* $\mathbb{P}^*$*-discrimination-free price* defined by

$$\pi^{*,u}(\mathbf{X}) := h^*(\mathbf{X}) + B^*. \tag{4.1}$$

Moreover, if we do not consider any covariates (neither discriminatory nor nondiscriminatory ones), we are back in the situation of a homogeneous situation where we charge the same (constant) premium μ to every policyholder. A drawback of the uniformly adjusted price (4.1) is that it may result in negative prices for certain covariate values $\mathbf{X}$.

A different position is to allocate the bias B^* by differentiating w.r.t. $\mathbf{X}$ in a still discrimination-free fashion (avoiding any inference of $\mathbf{D}$ from $\mathbf{X}$). A natural way is to allocate the total premium proportionally to $h^*(\mathbf{X})$, resulting in the *proportionally adjusted* $\mathbb{P}^*$*-discrimination-free price*

$$\pi^{*,p}(\mathbf{X}) := h^*(\mathbf{X}) \frac{\mu}{\mu - B^*}. \tag{4.2}$$

In the remainder of this section, we discuss a more sophisticated approach that chooses the distribution $\mathbb{P}^*(\mathbf{d})$ specifically such that the discrimination-free price $h^*(\mathbf{X})$ is unbiased, that is, $B^* = 0$. A simple illustration was given in Example 8. In general, there will be many such distributions that may satisfy this condition, and an additional criterion for choosing $\mathbb{P}^*(\mathbf{d})$ is needed.

A standard criterion is to chose the measure $\mathbb{P}^*$, such that the distribution $\mathbb{P}^*(\mathbf{d})$ is as close as possible to the physical distribution $\mathbb{P}(\mathbf{d})$, subject to the resulting discrimination-free price $h^*(\mathbf{X})$ being unbiased. To proceed, first note that, given independence of $(\mathbf{X}, \mathbf{D})$ under $\mathbb{P}^*$, it holds that

$$\mathbb{E}[h^*(\mathbf{X})] = \mathbb{E}^*[\zeta(\mathbf{D})],$$

where $\zeta(\mathbf{t}) = \mathbb{E}[\mu(\mathbf{X}, \mathbf{t})]$. When the relative entropy (Kullback–Leibler divergence) is chosen to quantify distance between distributions, we work out $\mathbb{P}^*(\mathbf{d})$ as the solution to the following problem:

$$\min_{\mathbb{Q}(\mathbf{d})} \mathbb{E}\left[\frac{d\mathbb{Q}(\mathbf{d})}{d\mathbb{P}(\mathbf{d})} \log\left(\frac{d\mathbb{Q}(\mathbf{d})}{d\mathbb{P}(\mathbf{d})}\right)\right], \quad \text{such that} \quad \mathbb{E}^*[\zeta(\mathbf{D})] = \mu. \tag{4.3}$$

Following standard results (see Breuer and Csiszár, 2013; Csiszár, 1975) for precise statement and conditions), the solution takes the form:

$$\mathbb{P}^*(\mathbf{d}) = \mathbb{E}\left[\mathbf{1}_{\{\mathbf{D} \le \mathbf{d}\}} \frac{e^{\beta\zeta(\mathbf{D})}}{\mathbb{E}[e^{\beta\zeta(\mathbf{D})}]}\right],$$

where the parameter β is suitably chosen such that the constraint $\mathbb{E}^*[\zeta(\mathbf{D})] = \mu$ is fulfilled. Note that, in view of Section 2.4, we need to assume existence of distributions $\mathbb{P}^*(\mathbf{d})$ that fulfill the constraint in (4.3).

Hence, the premium for a policyholder with nondiscriminatory covariates $\mathbf{X} = \mathbf{x}$ is defined by

$$\pi^{*,KL}(\mathbf{x}) := h^*(\mathbf{x}) = \mathbb{E}\left[\mu(\mathbf{x}, \mathbf{D}) \frac{e^{\beta\zeta(\mathbf{D})}}{\mathbb{E}[e^{\beta\zeta(\mathbf{D})}]}\right]. \tag{4.4}$$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

To ease the interpretation of this formula, let $\mathbf{D} = D$ be one-dimensional and $\mu(\mathbf{x}, d) \geq 0$ be increasing in d. Then, for $\beta > 0$, we have

$$\begin{aligned}\pi^{*,KL}(\mathbf{x}) &= \mathbb{E}\left[\mu(\mathbf{x}, D)\frac{e^{\beta\zeta(D)}}{\mathbb{E}[e^{\beta\zeta(D)}]}\right] \\ &= \mathbb{E}\left[\mu(\mathbf{x}, D)\right] + \mathbb{C}\mathrm{ov}\left[\mu(\mathbf{x}, D), \frac{e^{\beta\zeta(D)}}{\mathbb{E}[e^{\beta\zeta(D)}]}\right] \\ &\geq \mathbb{E}\left[\mu(\mathbf{x}, D)\right] = h(\mathbf{x}),\end{aligned}$$

which corresponds to the situation where the choice $\mathbb{P}^* = \mathbb{P}$ would produce a negative bias (under-pricing). The calculation of $\pi^{*,KL}(\mathbf{x})$ assigns a higher premium to policyholders with covariates $\mathbf{X} = \mathbf{x}$ such that $\mu(\mathbf{x}, D)$ is more volatile, as can be seen in approximation (4.5) below. This represents policies for which lack of information on discriminatory covariates matters more, in the sense that there is a higher sensitivity to the uncertainty induced by not using the discriminatory factor D. One can thus view the bias correction in $\pi^{*,KL}(\mathbf{x})$ as an implicit discrimination-free risk load.

For β close to zero, a Taylor series expansion of $\pi^{*,KL}(\mathbf{x})$ gives the approximation

$$\begin{aligned}h^*(\mathbf{x}) &\approx \mathbb{E}\left[\mu(\mathbf{x}, \mathbf{D})\right] + \beta\mathbb{C}\mathrm{ov}\left[\mu(\mathbf{x}, \mathbf{D}), \zeta(\mathbf{D})\right] \\ &= \mathbb{E}\left[\mu(\mathbf{x}, \mathbf{D})\right] + \beta\sqrt{\mathbb{V}\mathrm{ar}[\mu(\mathbf{x}, \mathbf{D})]\mathbb{V}\mathrm{ar}[\zeta(\mathbf{D})]}\mathbb{C}\mathrm{orr}\left[\mu(\mathbf{x}, \mathbf{D}), \zeta(\mathbf{D})\right].\end{aligned} \tag{4.5}$$

5. Estimated prices

All previous discussion and derivations of discrimination-free prices and indirect discrimination were conducted under the assumption that the "true" probabilistic model underlying the portfolio $(Y, \mathbf{X}, \mathbf{D})$ is known, represented by the physical measure $\mathbb{P}$. In practice, an *estimated model* is used because, typically, the data generating mechanism is unknown.

Specifically, one starts from data

$$\mathcal{S} = \{(y_1, \mathbf{x}_1, \mathbf{d}_1), \ldots, (y_n, \mathbf{x}_n, \mathbf{d}_n)\},$$

assuming that $(y_i, \mathbf{x}_i, \mathbf{d}_i)$ are i.i.d. realisations of $(Y, \mathbf{X}, \mathbf{D}) \sim \mathbb{P}$. As the data are generated under $\mathbb{P}$, we cannot estimate discrimination-free prices $h^*(\mathbf{X})$ directly under $\mathbb{P}^*$. Instead, we need to estimate best-estimate prices first under $\mathbb{P}$, and then we can derive discrimination-free prices by averaging out $\mathbf{d}$ with respect to the chosen distribution $\mathbb{P}^*(\mathbf{d})$.

Consequently, a regression model (in the broader sense) is chosen

$$\widehat{\mu}\colon(\mathbf{x}, \mathbf{d}) \mapsto \widehat{\mu}(\mathbf{x}, \mathbf{d}) = \widehat{\mu}(\mathbf{x}, \mathbf{d}; \boldsymbol{\theta}), \tag{5.1}$$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

which typically differs from the (true) best-estimate price functional $(\mathbf{x}, \mathbf{d}) \mapsto \mu(\mathbf{x}, \mathbf{d})$, given in Definition 2, but which should mimic $\mu(\mathbf{x}, \mathbf{d})$ in the best possible way. One may specify a fixed functional form for $\widehat{\mu}$ in (5.1) or, in a wider sense, one can specify an algorithm that generates the mapping (5.1) from the data $\mathcal{S}$. In either case, $\widehat{\mu}$ will still depend on unknown parameters $\boldsymbol{\theta}$ that have to be estimated from the data $\mathcal{S}$ (using a given objective function) yielding estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{S})$.

The resulting $\mathcal{S}$-calibrated regression function

$$(\mathbf{x}, \mathbf{d}) \mapsto \widehat{\mu}(\mathbf{x}, \mathbf{d}; \widehat{\boldsymbol{\theta}}), \tag{5.2}$$

then provides the approximation to the best-estimate price functional $(\mathbf{x}, \mathbf{d}) \mapsto \mu(\mathbf{x}, \mathbf{d})$. Note that (5.2) provides an estimate of the best-estimate price and, obviously, this estimate is, generally, discriminatory because it explicitly considers the discriminatory covariate values $\mathbf{d}$. Moreover, since we use the data $\mathcal{S}$ which have been generated under the physical measure $\mathbb{P}$, the regression function (5.2) also needs to be understood under the physical measure $\mathbb{P}$, we refer to Remark 13(d).

The unawareness price functional $\mathbf{x} \mapsto \mu(\mathbf{x})$ can be approximated in an analogous manner by just dropping $\mathbf{d}$ in (5.1) and (5.2), resulting in an estimated regression function

$$\mathbf{x} \mapsto \bar{\mu}(\mathbf{x}; \widehat{\boldsymbol{\vartheta}}), \tag{5.3}$$

where the functional forms $\widehat{\mu}$ and $\bar{\mu}$ may differ as well as their parameters $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$, respectively. We emphasize that typically $\bar{\mu}(\,\cdot\,; \widehat{\boldsymbol{\vartheta}})$ may indirectly discriminate w.r.t. $\mathbf{d}$ because in the estimation process of $\widehat{\boldsymbol{\vartheta}}$, we implicitly use covariate combinations $(\mathbf{x}_i, \mathbf{d}_i)$ which (empirically) contain the dependencies $\mathbb{P}(\mathbf{d} \mid \mathbf{x})$ that may allow for inference of $\mathbf{D}$ from $\mathbf{X}$. The estimated unawareness price $\bar{\mu}(\mathbf{x}; \widehat{\boldsymbol{\vartheta}})$ can also be interpreted as an approximation to

$$\mathbb{E}[\widehat{\mu}(\mathbf{X}, \mathbf{D}; \widehat{\boldsymbol{\theta}}) \mid \mathbf{X} = \mathbf{x}; \mathcal{S}],$$

using the tower property of conditional expectations argument for $\mathbf{D}$ (under the physical measure $\mathbb{P}$).

Typically, also $\mathbb{P}(\mathbf{d})$ is not known. Assuming $\mathbf{D}$ is discrete, $\mathbb{P}(\mathbf{d})$ can be estimated by the empirical probabilities $n_{\mathbf{d}}/n$ (observed relative frequency of the discriminatory covariate $\mathbf{d}$ in $\mathcal{S}$). This generates the discrimination-free price

$$\widehat{h}(\mathbf{x}) = \sum_{\mathbf{d}} \widehat{\mu}(\mathbf{x}, \mathbf{d}; \widehat{\boldsymbol{\theta}}) \frac{n_{\mathbf{d}}}{n}, \tag{5.4}$$

where we use the estimated best-estimate price functional (5.2); if $\mathbf{D}$ is continuous, we would use its empirical distribution function, which results in a discrete formula similar to (5.4). The price (5.4) is discrimination-free in the sense of Definition 6, that is, the discrimination-free property is not affected by the fact that we work with an estimated model. While potential estimation error may result in prices $\widehat{h}(\mathbf{x})$ that are not very close to $h(\mathbf{x})$, the property of nondiscrimination is preserved within the selected model; we explore this in more detail in Section 6. When choosing the structure of the regression function $\widehat{\mu}$ in (5.1), we should require existence of the discrimination-free price (5.4) in the sense of Proposition 14.

Finally, we note that one may attempt, in the light of Section 3, to estimate a graphical model (see e.g., Hernán and Robins, (2020)), which would provide discrimination-free prices in a more direct way. However, we do not pursue this direction for two reasons. First, because actuarial pricing models typically comprise a large number of covariates (e.g., more than 50 is typical for direct motor insurance pricing), which could make construction, estimation, and validation of an appropriate graphical model challenging. Second, we do not make any claim about causality in the context of specific actuarial applications; we merely note that our proposal is in line with concepts from causal inference, if particular conditions are fulfilled.

6. Numerical illustration

6.1. Model and alternative pricing rules

We present a simple health insurance example that demonstrates our approach of discrimination-free insurance pricing. This example satisfies the causal relations of Figure 1 and, thus, it can also be understood in a causal inference context.

Let $\mathbf{D} = D$ correspond to the single discriminatory characteristic "gender", that is, $D \in \{\text{woman}, \text{man}\}$. Furthermore, let $\mathbf{X} = (X_1, X_2)'$, where $X_1 \in \{15, \ldots, 80\}$ denotes the age of the policyholder, and $X_2 \in \{\text{non-smoker}, \text{smoker}\}$; below we assume that smoking habits are gender related. We consider three different types of health costs: birthing-related health costs only affecting women between ages 20 and 40 (type 1), cancer-related health costs with a higher frequency for smokers and also for women (type 2), and health costs due to other disabilities (type 3). For simplicity, we only consider claim counts, assuming deterministic claim costs for the three different claim types. We assume independence between individuals, all having the same exposure (= 1). Moreover, we assume that the claim counts for the different claim types are described by independent Poisson GLMs with canonical (i.e., log-) link function. The three different types of claims are governed by the following log-frequencies (regression functions):

$$\log \lambda_1(\mathbf{X}, D) := \alpha_0 + \alpha_1 1_{\{X_1 \in [20,40]\}} 1_{\{D=\text{woman}\}}, \tag{6.1}$$

$$\log \lambda_2(\mathbf{X}, D) := \beta_0 + \beta_1 X_1 + \beta_2 1_{\{X_2=\text{smoker}\}} + \beta_3 1_{\{D=\text{woman}\}}, \tag{6.2}$$

$$\log \lambda_3(\mathbf{X}, D) := \gamma_0 + \gamma_1 X_1, \tag{6.3}$$

based on the joint nondiscriminatory and discriminatory covariates $(\mathbf{X}, D)$. The deterministic claim costs of the different claim types are given by $(c_1, c_2, c_3) = (0.5, 0.9, 0.1)$ for claims of type 1, type 2, and type 3, respectively.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

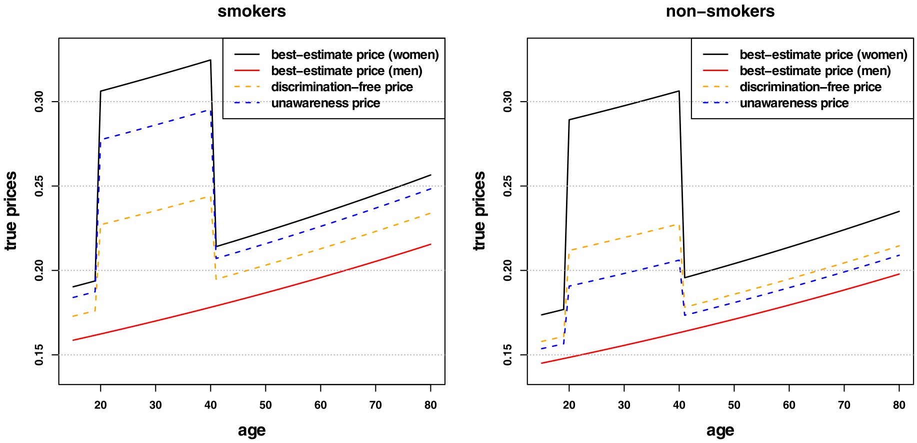

FIGURE 2: True model: (left) smokers and (right) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

The best-estimate price (considering all covariates) of Definition 2 is given by

$$\mu(\mathbf{X}, D) = c_1\lambda_1(\mathbf{X}, D) + c_2\lambda_2(\mathbf{X}, D) + c_3\lambda_3(\mathbf{X}, D).$$

This best-estimate price is illustrated in Figure 2 for the parameter values $(\alpha_0, \alpha_1) = (-40, 38.5)$, $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2, 0.004, 0.1, 0.2)$, and $(\gamma_0, \gamma_1) = (-2, 0.01)$. The plots on the left-hand side of Figure 2 refer to smokers ($X_2 = \text{smoker}$), while those on the right-hand side to non-smokers ($X_2 = \text{non-smoker}$). The solid black lines give the best-estimate prices $\mu(\mathbf{X}, D)$ for women and the solid red lines for men. Obviously, by using D as a rating factor, these best-estimate prices discriminate between genders.

Next, we calculate the discrimination-free price of Definition 6 for $\mathbb{P}^*(d) = \mathbb{P}(d)$, see (2.5), motivated by Proposition 15. It is given by

$$h(\mathbf{X}) = \sum_{d \in \{\text{woman, man}\}} (c_1\lambda_1(\mathbf{X}, d) + c_2\lambda_2(\mathbf{X}, d) + c_3\lambda_3(\mathbf{X}, d))\mathbb{P}(D = d).$$

For the calculation of this discrimination-free price, we need the gender proportions within our population. We set $\mathbb{P}(D = \text{woman}) = 0.45$. The orange dotted lines in Figure 2 provide the resulting discrimination-free prices for smokers (left) and non-smokers (right). Note that these are identical for men and women, that is, all price differences can be described solely by different ages X_1 and smoking habits X_2, irrespective of gender D. Moreover, the smoking habits do not reveal information about the gender; note that in the exposition so far, it has not been necessary to describe how smoking habits vary by gender, that is, interpreted in a causal inference setting, we have not used any arrow $D \rightarrow \mathbf{X}$, see Section 3.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

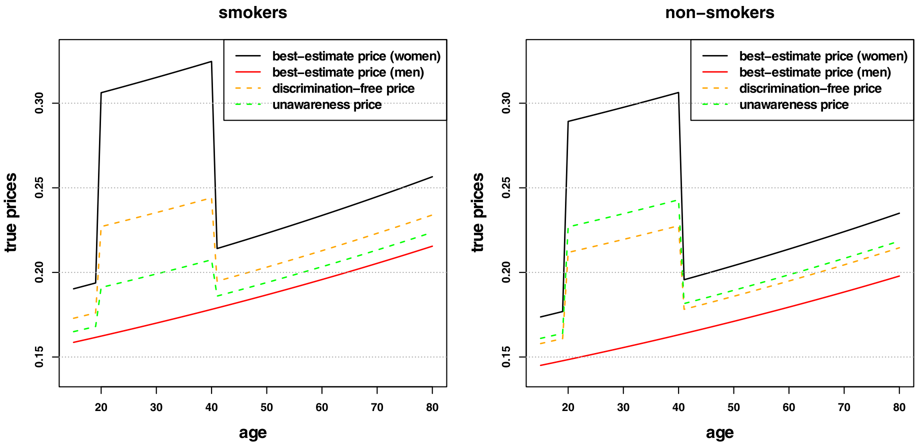

FIGURE 3: True model: (left) smokers and (right) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted green lines show the unawareness prices, for an alternative assumption on $\mathbb{P}(D = \text{woman} \mid \text{smoker})$.

We compare this discrimination-free price to the unawareness price obtained by simply dropping the gender covariate D from the calculations (Definition 4). Thus, we calculate

$$\mu(\mathbf{X}) = c_1\mathbb{E}[\lambda_1(\mathbf{X}, D) \mid \mathbf{X}] + c_2\mathbb{E}[\lambda_2(\mathbf{X}, D) \mid \mathbf{X}] + c_3\mathbb{E}[\lambda_3(\mathbf{X}, D) \mid \mathbf{X}].$$

The calculation of the unawareness price requires *additional information* about the following conditional probabilities

$$\mathbb{P}(D = d \mid \mathbf{X} = \mathbf{x}) = \frac{\mathbb{P}(D = d, \mathbf{X} = \mathbf{x})}{\mathbb{P}(\mathbf{X} = \mathbf{x})} = \frac{\mathbb{P}(D = d, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)}, \tag{6.4}$$

the last equality making the assumption that the age variable X_1 is independent from the random vector (X_2, D). In addition, we set $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.8$ and $\mathbb{P}(X_2 = \text{smoker}) = 0.3$. The former assumption tells us that smokers are more likely women; this is similar to Example 1. As a consequence, X_2 has explanatory power to predict the gender D, and the unawareness price may therefore be indirectly discriminatory against women. These unawareness prices are illustrated by the blue dotted lines in Figure 2. The blue dotted line lies above the discrimination-free price (orange) for smokers (Figure 2, left) and below for non-smokers (right). Thus, the unawareness price implicitly allocates a higher price to women because smokers are more likely women in our example, or in other words, the portfolio distribution allows us to infer the more likely gender from smoking habits.

Since there is no particular reason to assume a population where the proportion of smokers is greater amongst women, the potential for indirect gender discrimination is easily verified by an alternative assumption, namely, that smokers are more likely men, say, $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.2$. The resulting prices are plotted by the dotted green lines in Figure 3. We observe

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

that unawareness prices for smokers are below the discrimination-free ones (orange dotted line), with the reverse holding for non-smokers. That is, in this case women may again be indirectly discriminated against through their (non-)smoking habits, serving as a proxy for the explanatory variable of gender. This scenario demonstrates that the adjustment underlying discrimination-free prices does *not* undermine the direct causal impact (in the sense of Section 3) of smoking on prices, given that under discrimination-free prices the price for smokers *increases*, compared to unawareness prices. In fact, when $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.2$, unawareness prices "mask" the impact of smoking. In other words, when smoking is allowed to act as a proxy for gender, the sensitivity of prices to smoking reduces. This is because, for smokers, the unawareness price includes the implicit inference that the policyholder is a man, who, other things being equal, is less likely to claim than a woman.

The break-even point is $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.45 = \mathbb{P}(D = \text{woman})$ because in this case D and X_2 are independent, which prevents indirect discrimination through the portfolio distribution, and the unawareness price and the discrimination-free price are equal.

6.2. Application on estimated models

The previous discussion has been based on the knowledge of the model generating the data. We now address the more realistic situation where the model needs to be estimated. To this effect, we simulate data from $(\mathbf{X}, D, Y) \sim \mathbb{P}$ consistently with the given model assumptions, and subsequently calibrate a neural network regression model to the simulated data.

Specifically, we choose a health insurance portfolio of size $n = 100,000$ and simulate claim counts from the Poisson GLMs (6.1), (6.2), and (6.3), with the choice $\mathbb{P}(D = \text{woman} \mid X_2 = \text{smoker}) = 0.8$. An age distribution for X_1 is also needed for the simulation – the chosen probability weights are shown in Figure 4. We assume that age X_1 is independent from gender D and smoking habits X_2, as in (6.4).

Listing 1 gives an excerpt of the simulated data. We have the three covariates X_1 (age), X_2 (smoking habit), and D (gender) on lines 5–7, and lines 2–4 illustrate the numbers of claims N_1, N_2, and N_3, separated by claim types. The proportion of women in this simulated data is 0.4505, which is close to the true value of $\mathbb{P}(D = \text{woman}) = 0.45$. Our first aim is to fit a regression model to this data, under the assumptions that individual policies are independent, and that the different claim types are independent and Poisson distributed. Beside this, we do not make any structural assumption about the regression functions, but we try to infer them from the data using neural networks. The independence assumption between the claim counts N_1, N_2 and N_3 motivates modeling them separately. Thus, we will fit three different neural networks to model λ_1, λ_2, and λ_3, respectively. As we do not use any prior knowledge on the data generating process, we will feed all covariates (X_1, X_2, D) to each of the three networks.

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

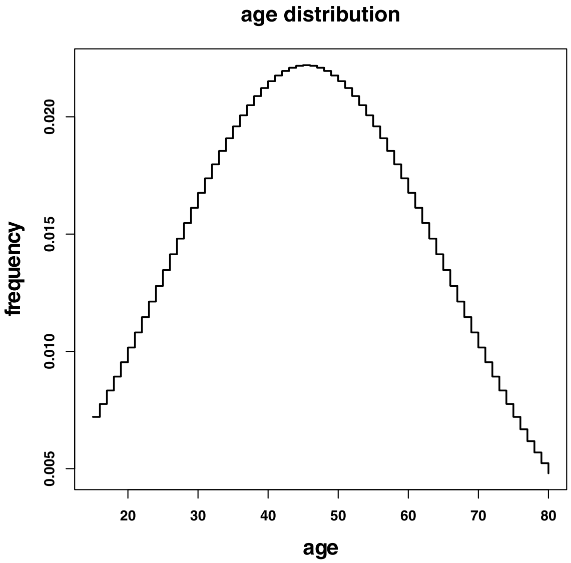

FIGURE 4: The age frequency used for both genders and smoking habits to simulate the data.

LISTING 1.

SIMULATED HEALTH INSURANCE DATA.

```
 ‘data.frame’ : 100000 obs. of 6 variables :
$ N1: int 0 0 0 0 0 0 0 0 0 0 ...
$ N2: int 0 0 1 0 0 1 0 0 2 0 ...
$ N3: int 0 1 0 0 1 0 0 0 0 0 ...
$ X1: num 36 57 70 49 63 27 41 58 16 34 ...
$ X2: num 0 0 1 0 0 1 0 0 1 1 ...
$ D : num 0 1 1 0 0 1 0 0 1 1 ...
```

LISTING 2.

NEURAL NETWORK ARCHITECTURE USED TO INFER λ_1, λ_2 AND λ_3.

```
Design <- layer_input (shape = c(3), dtype = ’float32 ’, name = ’Design ’)
#
Network = Design %>%
    layer_dense (units =15, activation = ’relu’, name = ’hidden1’) %>%
    layer_dense (units =15, activation = ’relu’, name = ’hidden2’) %>%
    layer_dense (units =1, activation = ’exponential’, name = ’Network ’)
#
model   <- keras_model (inputs = c(Design), outputs = c(Network))
model   %>% compile (loss = ’poisson’, optimizer = ’adam’)
```

Listing 2 illustrates the chosen neural network architecture, using the R library keras, with which the three regression functions (6.1)-(6.3) are estimated. We choose neural networks of depth 2 having 15 neurons in both

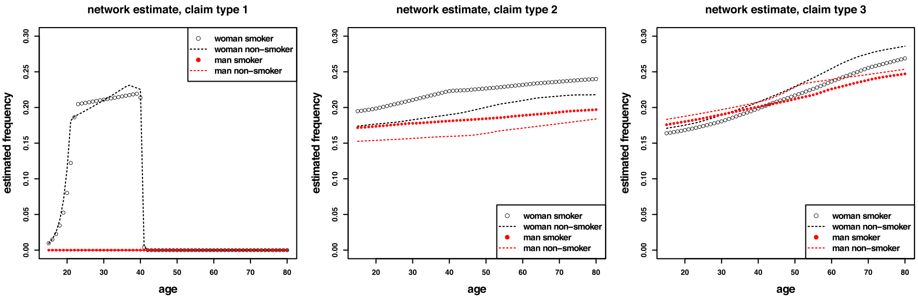

FIGURE 5: Estimated regression functions $\widehat{\lambda}_1(\mathbf{X}, D)$ (left), $\widehat{\lambda}_2(\mathbf{X}, D)$ (middle), and $\widehat{\lambda}_3(\mathbf{X}, D)$ (right) using the neural network architecture of Listing 2.

hidden layers, the rectified linear unit (ReLU) activation function, and the canonical link under the Poisson assumption. Moreover, we select the Poisson deviance loss as our objective function. This network involves 316 weights that need to be calibrated. We train these weights of the three networks over 1000 epochs on batches of size 20,000.

Figure 5 illustrates the estimates $\widehat{\lambda}_1(\mathbf{X}, D)$, $\widehat{\lambda}_2(\mathbf{X}, D)$, and $\widehat{\lambda}_3(\mathbf{X}, D)$ of the three regression functions (6.1), (6.2), and (6.3), respectively, obtained by fitting the three neural networks. The left-hand side of that figure gives claim type 1 which is birthing related. We see a rather accurate shape, with smoking habits correctly ignored and men not affected by these claims. Figure 5 (middle) gives the cancer related frequencies. Also here we receive the same order w.r.t. gender and smoking habits as in (6.2). Finally, the right-hand side illustrates all remaining claims. As, by (6.3) claim frequencies should not depend on gender and smoking habits, the variation between lines indicates that the regression model captures a spurious effect.

Using these estimated frequencies, we calculate the estimated best-estimate price (5.2)

$$\widehat{\mu}(\mathbf{X}, D;\widehat{\boldsymbol{\theta}}) = c_1\widehat{\lambda}_1(\mathbf{X}, D) + c_2\widehat{\lambda}_2(\mathbf{X}, D) + c_3\widehat{\lambda}_3(\mathbf{X}, D),$$

and its discrimination-free counterpart (5.4)

$$\widehat{h}(\mathbf{x}) = \sum_d \widehat{\mu}(\mathbf{x}, d;\widehat{\boldsymbol{\theta}})\frac{n_d}{n},$$

with empirical proportions $n_{\text{woman}}/n = 1 - n_{\text{man}}/n = 0.4505$. These prices are illustrated in Figure 6: black lines give best-estimate prices for women, red lines for men, and with the orange dotted lines showing the discrimination-free counterparts. Comparing Figures 2 and 6, we conclude that the resulting true prices and estimated prices are rather similar. Of course, by construction the resulting discrimination-free price is gender neutral within the estimated model, and in our case close to the theoretical one.

We indicate what happens if we drop the gender variable D from the very beginning, that is, if we train the networks only on the covariates $\mathbf{X} = (X_1, X_2)$

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

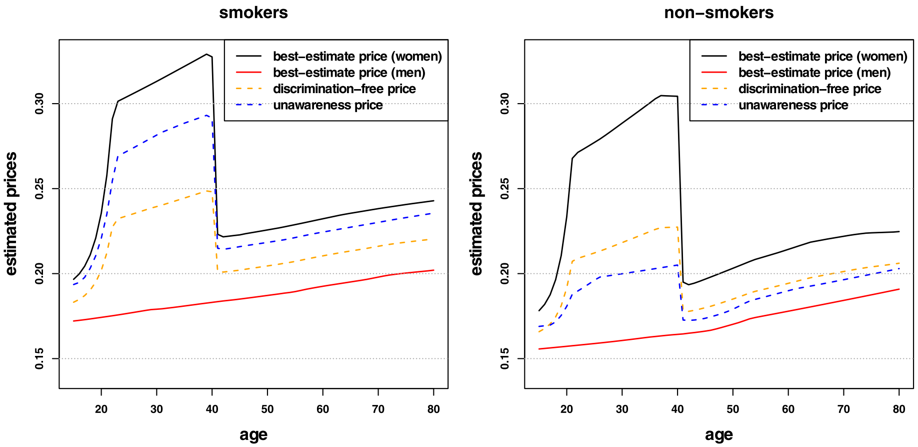

FIGURE 6: Estimated neural network model: (left) smokers and (right) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

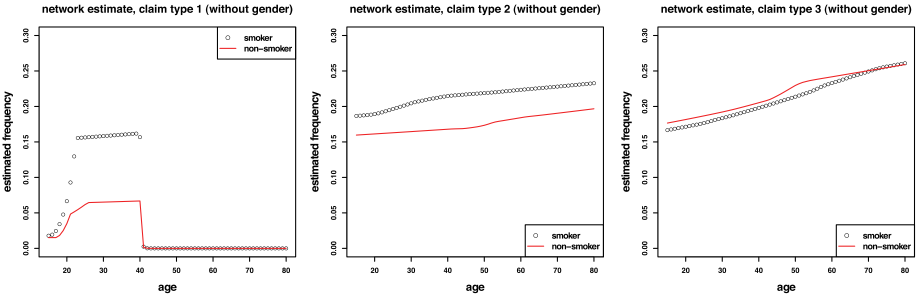

FIGURE 7: Estimated regression functions $\widehat{\lambda}_1(\mathbf{X})$ (left), $\widehat{\lambda}_2(\mathbf{X})$ (middle), and $\widehat{\lambda}_3(\mathbf{X})$ (right) using neural networks and ignoring the gender information D.

as considered in (5.3). We choose exactly the same network architecture as in Listing 2 except that we modify the input dimension on lines 1 from 3 for $(\mathbf{X}, D)$ to 2 for $\mathbf{X}$. This network involves 301 weights that need to be trained. The resulting estimated regression functions $\widehat{\lambda}_1(\mathbf{X})$, $\widehat{\lambda}_2(\mathbf{X})$, and $\widehat{\lambda}_3(\mathbf{X})$, ignoring gender information D, are illustrated in Figure 7. The left-hand side shows that we can no longer distinguish between gender; however, smokers are more heavily punished for birthing related costs, which is an undesired indirect discrimination effect against women because they are more often among the group of smokers (note that the y-scales in Figures 5 and 7 are the same). Finally, merging the different claim types provides the estimated unawareness prices (when first dropping D) as illustrated by the blue dotted lines in Figure 6, which can be compared with the blue dotted lines in Figure 2.

In our next analysis, we illustrate that the (non-)discrimination property does not depend on the quality of the regression model (5.1) chosen. We

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

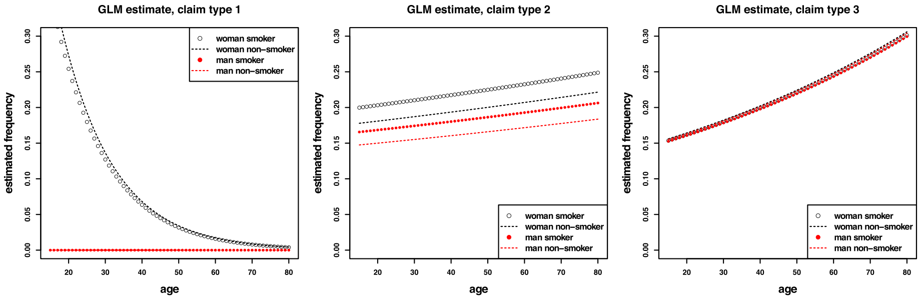

FIGURE 8: GLM estimated regression functions $\widehat{\lambda}_1^{\text{GLM}}(\mathbf{X}, D)$ (left), $\widehat{\lambda}_2^{\text{GLM}}(\mathbf{X}, D)$ (middle) and $\widehat{\lambda}_3^{\text{GLM}}(\mathbf{X}, D)$ (right).

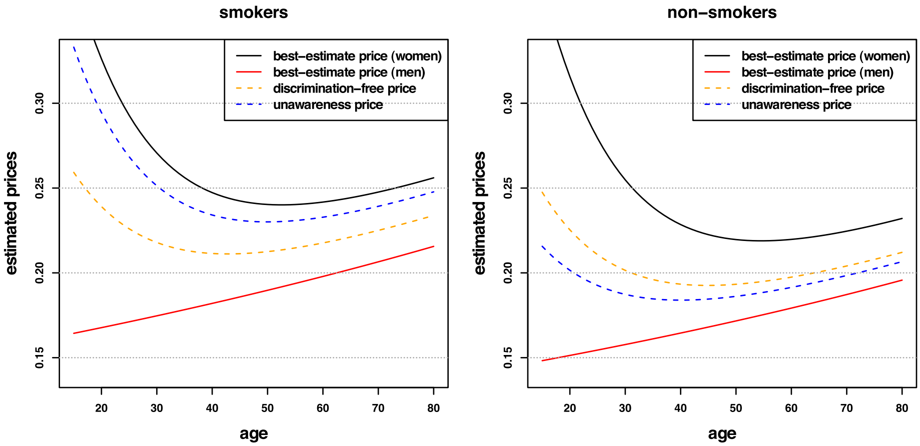

FIGURE 9: Estimated GLM: (left) smokers and (right) non-smokers with solid black and red lines giving the best-estimate prices for women and men, respectively. The dotted orange lines show the discrimination-free prices and the dotted blue lines show the unawareness prices.

choose a poor regression model (compared to the neural network above) by just assuming GLMs for $j = 1, 2, 3$

$$(\mathbf{x}, d) \mapsto \log \widehat{\lambda}_j^{\text{GLM}}(\mathbf{x}, d) = \theta_0^{(j)} + \theta_1^{(j)} x_1 + \theta_2^{(j)} 1_{\{x_2 = \text{smoker}\}} + \theta_3^{(j)} 1_{\{d = \text{woman}\}}. \tag{6.5}$$

This model will perform well for $j = 2, 3$, see (6.2)-(6.3), but it will perform poorly for $j = 1$, see (6.1). This is because such a model has difficulties capturing the highly nonlinear birthing-related effects, as seen in Figure 8 (left).

In Figure 9, we present the resulting best-estimate prices (black/red), unawareness prices (blue), and discrimination-free prices (orange), as estimated using the GLM. The first observation is that the resulting prices are a poor approximation to the true prices of Figure 2, the latter assuming full knowledge of the true model. However, the general discrimination behavior is the same in both figures, namely, that the unawareness price discriminates

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

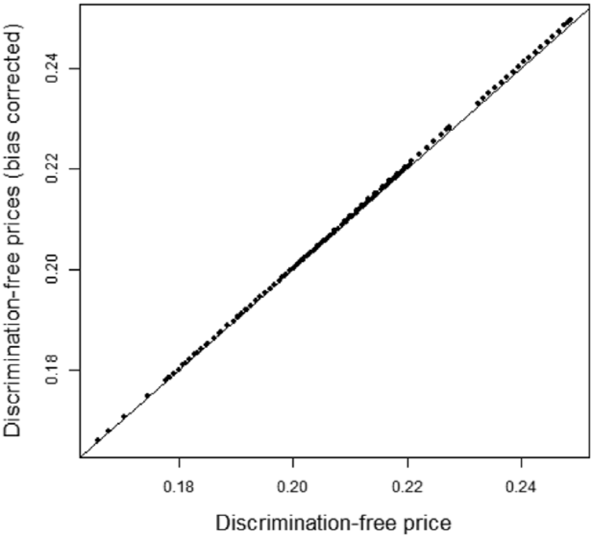

FIGURE 10: Bias-corrected discrimination-free prices $\widehat{h}^*(\mathbf{x})$ against unadjusted discrimination-free prices $\widehat{h}(\mathbf{x})$.

indirectly by learning the gender D from smoking habits X_2. This is illustrated by the relative positioning of blue and orange dotted lines, with smokers more heavily charged for birthing related costs due to the fact that smokers are more likely women.

In our last step, we consider the issue of correcting the bias introduced by discrimination-free pricing. The average predicted cost per policyholder and the average discrimination-free price are, respectively:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(\mathbf{x}_i, d_i; \widehat{\boldsymbol{\theta}}) = 0.2054$$

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{h}(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}) = 0.2050.$$

Thus, we have a small negative bias of approximately 0.2% of μ. We correct for this bias through an appropriate choice of $\mathbb{P}^*(D)$, as discussed in Section 4, yielding a bias-corrected price

$$\widehat{h}^*(\mathbf{x}) = \sum_{d} \widehat{\mu}(\mathbf{x}, d; \widehat{\boldsymbol{\theta}})\mathbb{P}^*(D = d).$$

As the discriminatory variable D has only two states, there is no need to use the complex formula (4.4); by setting $\frac{1}{n}\sum_{i=1}^{n} \widehat{h}^*(\mathbf{x}_i) = \mu$, one can directly obtain $\mathbb{P}^*(D = \text{woman}) = 0.4564$, which is slightly higher than the empirical portfolio proportion $n_{\text{woman}}/n = 0.4505$. In Figure 10, we display the bias-corrected discrimination-free prices $\widehat{h}^*(\mathbf{x})$ against the unadjusted discrimination-free

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

prices $\widehat{h}(\mathbf{x})$. We see that bias correction does not lead to any substantial price distortion in our example.

Remark. *There is one issue that has not been considered so far and which has been mentioned in the EU legislation (European Commission, 2012), footnote (1) to Article 2.2(14) – life and health underwriting. Namely, we have implicitly assumed that the measurements of the nondiscriminatory covariates are independent of the discriminatory characteristics. If we think of gender as a discriminatory covariate, this is not necessarily the case because, for instance, the waist to hip ratios naturally live on different scales for different genders, but they may still have the same impact on health related questions. This implies that nondiscriminatory covariates may need pre-processing w.r.t. discriminatory ones, such that the resulting measurements for different discriminatory characteristics are comparable.*

7. Concluding remarks

We conclude that the aim of this paper has been to provide:

(a) an actuarial formulation of discrimination-free prices;

(b) a demonstration that the omission of discriminatory information may lead to indirect discrimination in prices;

(c) a proposal for a simple formula that generates discrimination-free prices which works regardless of the choice of the underlying model;

(d) methods that ensure unbiasedness of discrimination-free prices at the portfolio level (the same considerations apply when transforming an actuarial tariff into a commercial one); and

(e) a discussion on the role of available data in obtaining discrimination-free prices.

The starting point to this paper has been an actuarial one. We have intentionally avoided a discussion on “fairness,” and, consequently, how fairness may be measured. For more on these topics, we refer to Kusner *et al*. (2017) and the references therein. Moreover, we have also not commented on which factors should be viewed as discriminatory – this is a societal decision that goes far beyond our actuarial discussion, see for example, Avraham *et al*. (2014). We (only) provide tools to implement such decisions.

We mention important points that have not been studied in this paper and which need further scientific research. First, discrimination-free pricing may have systemic implications, be they adverse or beneficial. For example, gender neutral pricing of motor insurance may result in cheaper premiums for more dangerous (male) drivers and vice versa, with the resulting incentives leading to a deterioration of aggregate driving behavior. On the other

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

hand, removing gender from car insurance pricing, arguably calls for including other covariates that better represent the risks being priced – ultimately the driving behavior. This is something within reach using telematics data, notwithstanding associated privacy concerns. Another example relates to the use of post-code information, which often correlates with ethnicity. Here, discrimination-free pricing can prevent further penalization of ethnic groups that have suffered historical injustices. The role of insurance in engineering socially beneficial outcomes is yet another discussion we cannot engage with in this paper. Another point worth commenting is whether discrimination-free pricing negatively impacts portfolio mixes (by adverse selection). Such impacts may result in a worse risk landscape of the industry, higher capital demands and, likely, higher premiums for the whole society.

An issue worth stressing once again is that, in order to be able to calculate discrimination-free prices, one needs to have access to *all discriminatory characteristics* – otherwise, it is not possible to properly adjust for the influence of such characteristics. When it comes to gender, the availability of such data may be feasible, but if we wanted to adjust for, for example, religious beliefs or sexual orientation, such information is in general not readily available. Customers may perceive it as peculiar and intrusive to be approached with questions concerning this type of apparently irrelevant (and sensitive) information. A concrete example is discussed in De Jong and Ferries (2006), where sexual preference is discussed as a risk factor relating to AIDS; the authors also highlight the danger of obtaining untruthful answers to questions around sensitive information, undermining the reliability of collected data. More broadly, collecting data on prohibited characteristics, as well as measuring their predictive power, could itself be legally contested (Prince and Schwarcz, 2019).

A key position taken in the present paper concerns the role of the overall price prediction at portfolio level. We have argued that the aggregate price for the portfolio may be calculated using all available information, including discriminatory covariates. Given this, it is the allocation of this overall cost that may introduce discrimination, and the discrimination-free pricing may be thought of as generating an allocation that avoids this. From this perspective, we know from the start that the allocation is biased w.r.t. the underlying (best-estimate) portfolio risk profile. It is, hence, of interest to analyze how this biased risk profile will affect the performance of the overall portfolio price prediction.

The argumentation used in the present paper has focused directly on how to obtain a discrimination-free price. This has led us to a procedure that tells us how to adjust the best-estimate price to arrive at a discrimination-free price. In a statistical sense, this could be seen as a "discrimination-free point estimate." A different line of thought instead could be that we try to develop a full statistical model that is discrimination-free, that is, sacrificing predictive performance by appropriately disregarding direct and indirect discrimination, this would result in a full statistical model that provides discrimination-free

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press

responses. An example of this approach in a life insurance context are the gender neutral intensities discussed in Chen and Vigna (2017). The main reason for considering prices directly is that we believe that this approach is closer to actuarial thinking, and because maximal predictive accuracy is a desirable feature in risk management, that is, we may use the full model for risk management purposes, but charge insurance prices according to its discrimination-free counterpart.

Acknowledgments

The authors would like to thank (in alphabetical order) the following for constructive remarks on an earlier version of the manuscript: Alois Gisler, Andreas Lagerås, Arvid Sjölander, Dirk Tasche, and Ruodu Wang.

References

Aseervatham, V., Lex, C. and Spindler, M. (2016) How do unisex rating regulations affect gender differences in insurance premiums? *The Geneva Papers on Risk and Insurance-Issues and Practice* **41**(1), 128–160.

Avraham, R., Logue, K. D. and Schwarcz, D. B. (2014) Understanding insurance anti-discrimination laws. *Southern California Law Review* **87**(2), 195–274.

Breuer, T. and Csiszár, I. (2013) Systematic stress tests with entropic plausibility constraints. *Journal of Banking & Finance* **37**(5), 1552–1559.

Bruszas, S., Kaschützke, B., Maurer, R. and Siegelin, I. (2018) Unisex pricing of German participating life annuities—Boon or bane for customer and insurance company? *Insurance: Mathematics and Economics* **78**, 230–245.

Chen, A., Guillén, M. and Vigna, E. (2018) Solvency requirement in a unisex mortality model. *ASTIN Bulletin: The Journal of the IAA* **48**(3), 1219–1243.

Chen, A. and Vigna, E. (2017) A unisex stochastic mortality model to comply with EU Gender Directive. *Insurance: Mathematics and Economics* **73**, 124–136.

Csiszár, I. (1975) *I*-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* **3**(1), 146–158.

De Jong, P. and Ferris, S. (2006) Adverse selection spirals. *ASTIN Bulletin: The Journal of the IAA* **36**(2), 589–628.

European Commission (2012). Guidelines on the application of COUNCIL DIRECTIVE 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). *Official Journal of the European Union* C11, 1–11.

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, 37–43.

Frees, E. W. and Huang, F. (2021) The discriminating (pricing) actuary. *North American Actuarial Journal*, to appear.

Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2018). Predictably unequal? the effects of machine learning on credit markets. Available at SSRN: https://ssrn.com/abstract=3072038 (Downloaded on Feb. 21, 2020).

Guelman, L. and Guillén, M. (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications* **41**(2), 387–396.

Guillén, M. (2012). Sexless and beautiful data: from quantity to quality. *Annals of Actuarial Science* **6**(2), 231–234.

HERNÁN, M. A. and ROBINS, J. M. (2020) *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

KRIKLER, S., DOLBERGER, D. and ECKEL, J. (2004) Method and tools for insurance price and revenue optimisation. *Journal of Financial Services Marketing* **9**(1), 68–79.

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017) Counterfactual fairness. In Advances in Neural Information Processing Systems, pp. 4066–4076.

LAURITZEN, S. L. (1996) *Graphical Models*. Oxford Science Publications.

PEARL, J. (2009) Causal inference in statistics: An overview. *Statistics Surveys* **3**, 96–146.

PEARL, J., GLYMOUR, M. and JEWELL, N. P. (2016) *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

POPE, D. G. and SYDNOR, J. R. (2011) Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy* **3**(3), 206–231.

PRINCE, A. E. R. and SCHWARCZ, D. (2019) Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review* **105**, 1257.

SASS, J. and SEIFRIED, F. T. (2014) Insurance markets and unisex tariffs: Is the European Court of Justice improving or destroying welfare? *Scandinavian Actuarial Journal* **2014**(3), 228–254.

TOBLER, C. (2008) Limits and potential of the concept of indirect discrimination. Directorate-General for Employment, Social Affairs and Equal Opportunities, Unit G2, European Commission.

WILLIAMS, D. (1991) *Probability with Martingales*. Cambridge University Press.

ZHAO, Q. and HASTIE, T. (2021) Causal interpretations of black-box models. *Journal of Business & Economic Statistics* **39**(1), 272–281.

M. LINDHOLM
Department of Mathematics, Division of Mathematical Statistics,
Stockholm University, Stockholm 106 91, Sweden
E-mail: lindholm@math.su.se

R. RICHMAN
Old Mutual Insure & University of the Witwatersrand,
Johannesburg 2192, Republic of South Africa
E-mail: ronald.richman@ominsure.co.za

A. TSANAKAS
Bayes Business School, City, University of London, 106 Bunhill
Row, London EC1Y 8TZ, United Kingdom
E-mail: A.Tsanakas.1@city.ac.uk

M.V. WÜTHRICH (CORRESPONDING AUTHOR)
RiskLab, Department of Mathematics, ETH Zurich, Zurich 8092,
Switzerland
E-mail: mario.wuethrich@math.ethz.ch

https://doi.org/10.1017/asb.2021.23 Published online by Cambridge University Press