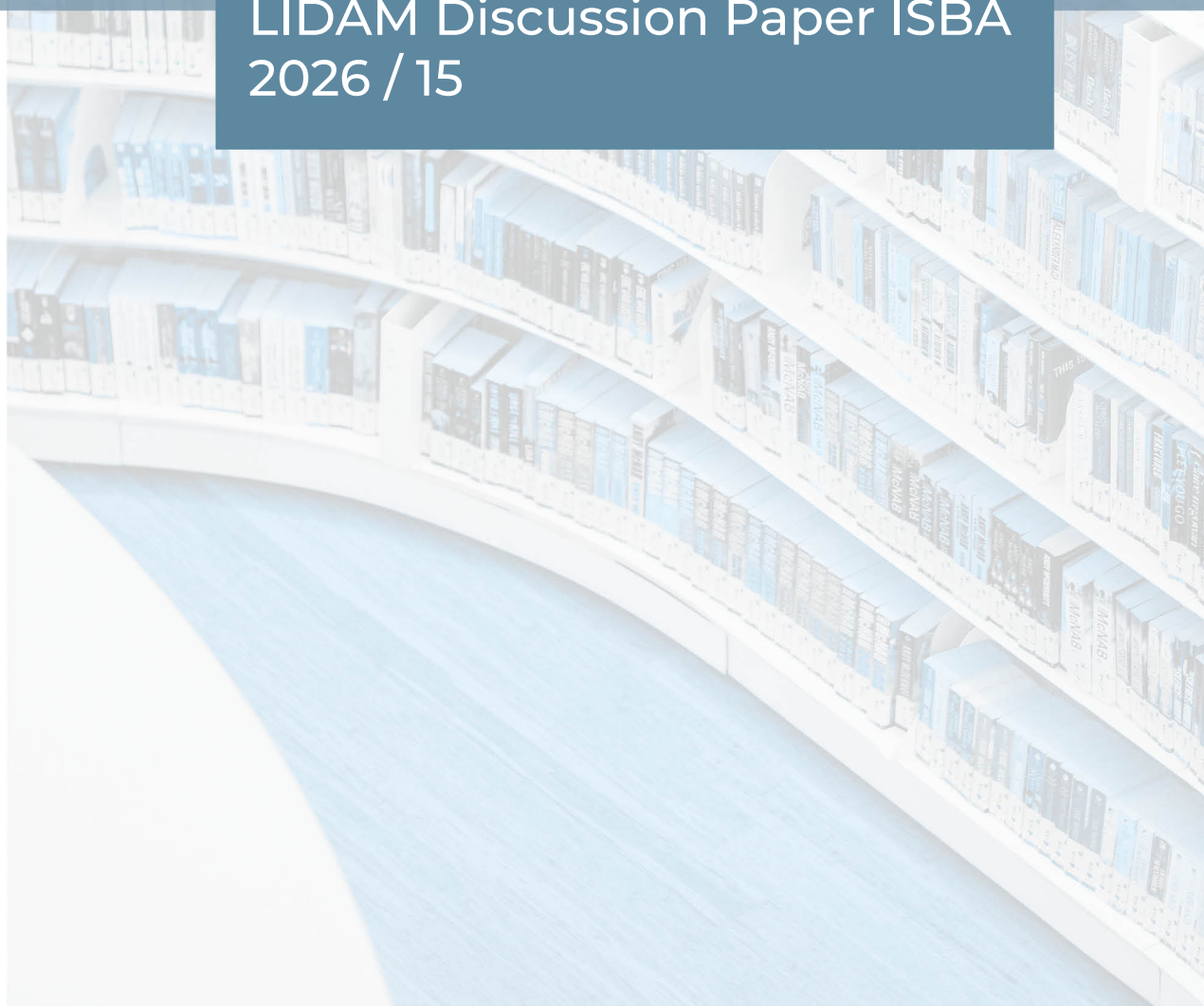
BALANCE AND FAIRNESS THROUGH MULTICALIBRATION IN NONLIFE INSURANCE PRICING

Michel Denuit, Marie Michaelides, Julien Trufin

LIDAM Discussion Paper ISBA
2026 / 15



UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

BALANCE AND FAIRNESS THROUGH MULTICALIBRATION IN NONLIFE INSURANCE PRICING

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
Louvain Institute of Data Analysis and Modeling
UCLouvain
Louvain-la-Neuve, Belgium

Marie Michaelides
School of Mathematical and Computer Sciences
Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh, United Kingdom

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Brussels, Belgium

March 18, 2026

Abstract

Autocalibration is known to be an important requirement for insurance premiums since it guarantees that premium income balances corresponding claims, on average, not only at portfolio level but also inside each group paying similar premiums. Also, fairness has become a major concern because unfair treatment may expose insurers to lawsuits or reputational damage. Translating fairness into conditional mean independence allows actuaries to combine autocalibration and fairness into the multicalibration concept. This paper studies the properties of multicalibration in an insurance context and proposes practical ways to implement it, through local regression or bias correction within groups including credibility adjustments. A case study based on motor insurance data illustrates the relevance of multicalibration in insurance pricing.

Keywords: Insurance pricing, autocalibration, fairness, sufficiency, multicalibration.

1 Introduction and motivation

Financial equilibrium is an important property in insurance pricing as it ensures that premiums match corresponding claims, on average. This property must hold at portfolio level (global balance) but also within meaningful groups of policyholders. Numerical evidence suggests that candidate premiums produced by machine learning tools, like boosted regression trees or neural networks, can depart substantially from observed claim totals. See, e.g., Denuit et al. (2019) and the references therein. Autocalibration has been proposed as a remedy by Denuit et al. (2021), Denuit and Trufin (2023, 2024), Lindholm et al. (2023), Wüthrich (2023) and Wüthrich and Ziegel (2024) since it ensures that the amounts collected from policyholders paying the same premium match the corresponding claim totals, on average.

Besides financial equilibrium, discrimination is another important issue in the private insurance industry. We refer the reader, e.g., to Frees and Huang (2023) and Charpentier (2024) for a general overview of the topic. The ban of gender-based discrimination within the European Union (EU) is emblematic in that respect. In 2011, the European Court of Justice concluded that any gender-based insurance discrimination must be prohibited. From December 21, 2012, all insurers operating in the EU are required to offer unisex premiums and benefits. In North-America, race is another example of protected feature in several states and provinces. The protected attributes listed in the EU Directives are particularly relevant for insurers operating in the EU. In addition to gender, this list includes some other features that are often used in insurance like age, health or disability status, for instance. As another example, Fröhlich and Williamson (2024) consider "having migrant background" as sensitive feature in their study of fair machine learning techniques.

The EU directive banning the use of gender in insurance risk classification imposes individual fairness. In general, this means that the amounts of premium charged to policyholders differing only on the prohibited feature must be equal. In this case, the sensitive feature is not included in the risk classification scheme and commercial premiums do not depend on the sensitive feature in an explicit way. However, indirect, or proxy discrimination may remain present when some rating factors included in the price list are correlated to the omitted sensitive feature. For instance, motor insurance premiums often differ according to the power of the vehicle or the distance traveled. If these features are correlated to gender, because men tend to drive more powerful cars over longer distances for example, then insurance premiums still indirectly discriminate according to gender. EU guidelines precisely consider this case, as explained later on.

This paper considers the case of a sensitive feature not subject to individual fairness by the law. The insurer may thus let commercial premiums vary according to the sensitive feature. Typical examples include protected variables like health or disability status that are used by insurers or proxy variables to a forbidden rating factor. Because the use of these features may expose the insurer to complaints from consumers organizations or formal investigation by the regulatory authorities, it is important to be able to prove that the price list is fair with respect to the sensitive feature under scrutiny, in some sense to be defined.

There are several approaches to fairness in the insurance and in the machine learning literature. Let us mention actuarial fairness, individual fairness (or fairness through unawareness after Dwork et al., 2012), counterfactual fairness (Kusner et al., 2017), and several group-

fairness criteria (independence, separation, and sufficiency after Barocas et al., 2019). In this paper, we follow Baumann and Loi (2023) who advocate the relevance of a weak version of sufficiency in insurance pricing. This group-fairness notion imposes conditional mean independence of the response and the sensitive feature, given the premium. Being based on conditional mean independence, sufficiency is intuitively appealing and nicely combines with autocalibration, resulting in the concept of multicalibration introduced by Hebert-Johnson et al. (2018). In words, a premium is multicalibrated with respect to a collection of groups partitioning the portfolio if it is autocalibrated within each group, in addition to being autocalibrated overall. The groups considered in this paper are formed based on the sensitive feature. The reinforcement of sufficiency into multicalibration is mentioned as an avenue for future research in Section 3.3 in Baumann and Loi (2023) while calibration conditional on gender and on salary has been implemented by Fissler et al. (2022) in their application to workers' compensation insurance. This paper develops these ideas and bridges fairness to autocalibration.

The remainder of this paper is organized as follows. The insurance ratemaking problem is formalized in Section 2. Section 3 recalls the autocalibration concept. Section 4 is devoted to fairness notions. The group-fairness concept based on conditional mean independence considered in this paper is introduced there. Section 5 unites autocalibration fairness through multicalibration while Section 6 explains how to implement multicalibration through multibalance correction. Numerical illustrations are given in Section 7. The final Section 8 summarizes the main contribution of this paper and discusses the results.

2 Nonlife insurance pricing

The machine learning literature mainly considers binary outcomes and algorithmic decision-making systems. Typical examples include criminal risk assessment, hiring, college admission, lending or social services intervention, for instance, corresponding to "yes-no" situations. A binary response is useful at underwriting stage in insurance where the insurer decides to cover the risk or to refuse it. However, claim-related responses (claim numbers, claim severities, and claims totals) enter the analysis when we consider insurance pricing and these responses are not binary (integer, continuous or mixed nature).

In this paper, we consider a general claim-related response $Y \geq 0$. To fix the ideas, it is useful to consider that Y corresponds to the annual claim frequency for a policyholder in the portfolio. In accordance with credibility theory, we denote as M the true mean value of Y, such that

$$\mathrm{E}[Y|M = \mu] = \mu.$$

The random variable M is assumed to have finite expectation. Notice that the last identity can be rewritten as $\mathrm{E}[Y|M] = M$, almost surely. Throughout the paper, every equality between random variables is assumed to hold with probability 1, or almost surely.

The true mean M remains unobserved and purely conceptual but is nevertheless useful to formalize the insurance ratemaking problem where M corresponds to the pure premium that should be charged by the insurer. This is because individual departures $Y - M$ cancel out in any large insurance portfolio where a law of large numbers applies. Charging M

corresponds to actuarial fairness, which is practically impossible because M cannot be observed. Contrary to algorithmic fairness problems, risk classification in insurance targets M which cannot be observed at individual level. At the collective level, the distribution of M reflects the composition of the insured population targeted by the insurance company under consideration.

Since M is unknown, the insurance company can only use information recorded in its database about each policyholder. This corresponds to the set of features $X_1, \ldots, X_p$ gathered in the random vector $\boldsymbol{X}$. Here, the information $\boldsymbol{X}$ is used as a substitute to M so that we assume that Y and $\boldsymbol{X}$ are independent given M. Actuarial pricing then aims to evaluate the pure premium as accurately as possible, based on $\boldsymbol{X}$. This means that the target is the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y given the available information $\boldsymbol{X}$. Henceforth, $\mu(\boldsymbol{X})$ is referred to as the best-estimate price after Lindholm et al. (2022). Notice that the conditional independence assumption allows us to write

$$\mu(\boldsymbol{X}) = \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X}, M]\big|\boldsymbol{X}\big] = \mathrm{E}[M|\boldsymbol{X}]$$

so that $\mu(\boldsymbol{X})$ also approximates the individual premium M based on the information available to the insurance company.

The unknown function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is approximated by a technical premium $\boldsymbol{x} \mapsto \pi_0(\boldsymbol{x})$. The technical premium is the most accurate approximation to $\mu(\boldsymbol{X})$, for internal use, only. It allows the insurance company to compute policyholder value, for instance, by comparing the anticipated cost $\pi_0(\boldsymbol{X})$ to the actual revenues generated by the contract. These revenues correspond to the commercial premium $\pi(\boldsymbol{X})$ actually charged for coverage (excluding taxes, commissions and general expenses). The premium structure $\pi(\cdot)$ is generally simplified compared to $\pi_0(\cdot)$, dropping some features or using them at a less granular level.

Henceforth, we assume that $\pi_0(\boldsymbol{X})$, $\pi(\boldsymbol{X})$, and $\mu(\boldsymbol{X})$ are continuous random variables admitting positive probability density functions over $(0, \infty)$. This eases exposition and is generally the case when there is at least one continuous feature contained in the available information $\boldsymbol{X}$ and the functions $\pi_0(\cdot)$ and $\pi(\cdot)$ are continuously increasing functions of real scores built from $\boldsymbol{X}$. Typically, a log-link is used and $\pi_0(\boldsymbol{x}) = \exp(\mathrm{score}_0(\boldsymbol{x}))$ where $\mathrm{score}_0(\boldsymbol{x})$ is obtained from boosting, random forests or neural networks. Also, $\pi(\boldsymbol{x})$ is the exponential of an additive score produced by a Generalized Additive Model for transparency and to comply with the multiplicative premium structure implemented in the majority of insurance commercial IT systems. Notice that we assume throughout the paper that commercial premiums vary according to rating factors related to risk, only, without non-risk based corrections (like price walking, for instance). For simplicity, we assume that the coverage period is 1 year. Exposure can be included in the analysis to account for shorter coverage periods, if needed, as is done in Section 7 for the numerical illustrations.

Fairness is meaningful only for the commercial premium $\pi(\boldsymbol{X})$. To figure out the situation, it is useful to think about the example of gender within EU, which never enters the commercial premium since its use is prohibited but which generally impacts on the technical premiums. The commercial premium can be regarded as the decision made by the insurance company about the policyholder while the technical premium corresponds to the prediction, as defined in Scantamburlo et al. (2025). Since $\pi_0(\cdot)$ is for internal use, only, there is no need

for it to be fair. On the contrary, it must be as accurate as possible so that the insurance company is able to counteract adverse selection and to monitor premium transfers induced within the portfolio by the application of the commercial premiums. Precisely, any departure $\pi(\boldsymbol{X}) - \pi_0(\boldsymbol{X})$ corresponds to a transfer to (if positive) or from (if negative) another risk profile.

The paper concentrates on $\pi(\cdot)$ for which financial equilibrium and fairness both make sense. It is nevertheless worth mentioning that the fairness notion studied in this paper improves the performances of any candidate premium. Hence, it is also meaningful for technical premiums to improve accuracy as measured by any Bregman loss, including deviance.

3 Autocalibration

In general, autocalibration ensures that the average response in a neighborhood defined with the help of the autocalibrated predictor under consideration matches the predictor value. See, e.g., Krüger and Ziegel (2021). Under autocalibration, the amounts collected from policyholders paying the same premium match the corresponding claim totals, on average. Subsidizing transfers between policyholders are thus avoided, helping to prevent adverse selection effects.

This concept is formally defined next, in the context of insurance.

Definition 3.1. The commercial premium $\pi(\cdot)$ is said to be autocalibrated if

$$\mathrm{E}[Y|\pi(\boldsymbol{X}) = p] = p \text{ for all premium amount } p. \tag{3.1}$$

Autocalibration is an important property in insurance pricing because it ensures that every group of policyholders paying the same premium is on average self-financing. In other words, (3.1) guarantees that the premium paid by policyholders charged the same amount of premium p exactly covers the expected claims of that group.

Notice that (3.1) can be equivalently rewritten as

$$\mathrm{E}[Y - \pi(\boldsymbol{X})|\pi(\boldsymbol{X}) = p] = \mathrm{E}[\mu(\boldsymbol{X}) - \pi(\boldsymbol{X})|\pi(\boldsymbol{X}) = p] = 0 \text{ for all premium amount } p.$$

This shows that the net cash-flows $Y - \pi(\boldsymbol{X})$ as well as the pricing errors $\mu(\boldsymbol{X}) - \pi(\boldsymbol{X})$ cancel out within groups of policyholders paying the same amount of premium, on average. Autocalibration thus appears to be a very appealing requirement for candidate premiums. From a conceptual point of view, (3.1) can also be written in terms of the true mean M as $\mathrm{E}[M|\pi(\boldsymbol{X})] = \pi(\boldsymbol{X})$ which ensures that the commercial premium also captures the premium that should be charged, on average. Notice that we can equivalently condition with respect to the score in (3.1) when $\pi(\boldsymbol{X})$ is obtained by exponentiating an additive score.

4 Fairness notion

4.1 Sensitive feature

Assume now that there is a sensitive feature S recorded in the database. Its use is allowed but requires particular attention because it exposes the insurer to possible complaints about

unfair treatment. In this paper, we assume that $S = X_{j_0}$ for some $j_0 \in \{1, 2, \ldots, p\}$ to ease exposition; see Remark 4.2 below for a discussion. This means that the insurer actually uses S for premium calculation but wants to demonstrate that it is used in a fair and responsible way. Typical examples for S include sensitive features like health or disability status, or socio-economic profile, for instance. In the situation where individual fairness is imposed by the law, like with gender in the EU for instance, S may correspond to a feature which is correlated to the prohibited one, as shown in the next example.

Example 4.1. Considering the ban of gender-based discrimination in the EU, the individual fairness imposed by the law means that the premium $\pi(\cdot)$ must fulfill the constraint $\pi(\boldsymbol{x}, \text{man}) = \pi(\boldsymbol{x}, \text{woman}) = \pi(\boldsymbol{x})$ for all risk profiles $\boldsymbol{x}$. In words, this means that a man and a woman with the same rating factors $\boldsymbol{x}$ must pay the same premium. If this condition is not fulfilled then direct discrimination results. This requirement is thus similar to the notions of "fairness through unawareness" or "blindness", except that the law partly prohibits the use of proxies. According to Article 17 of the "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)", the use of true risk factors that might be correlated with gender remains permitted. Examples are given to illustrate the extent of the prohibition. In a nutshell, the insurer is not allowed to include in its price list a feature without impact on claims but correlated to gender. Hence, the use of proxy variables as substitutes for gender remains permitted, even if it generates indirect discrimination. Here, S would be a rating factor X_{j_0} strongly correlated to gender while being a true risk factor if the insurance company wishes to demonstrate that it complies not only with the letter but also with the spirit of the anti-discrimination law.

Remark 4.2. Even if the paper concentrates on the case $S = X_{j_0}$ for some $j_0 \in \{1, 2, \ldots, p\}$, the sensitive feature S might also result from a combination of several features in $\boldsymbol{X}$, like disability and area, for instance, to ensure equal access to insurance for disabled people whatever their living area. A similar analysis can be performed in these more complicated cases where $S = \psi(\boldsymbol{X})$ for some given function ψ. Notice that S may even be another prediction or premium. For instance, $\psi(\boldsymbol{X})$ could be the prediction of policyholder's gender based on the rating factors $\boldsymbol{X}$.

Remark 4.3. The case where S is not contained in, or deduced from $\boldsymbol{X}$ is not relevant for practice under the fairness notion considered in this paper. Indeed, this situation would mean either that the use of S is prohibited by the law or that the insurer refuses to let premiums depend on S. If we except the case where Y and S are independent given $\pi(\boldsymbol{X})$, any fair premium (in the sense of multicalibration introduced below) necessarily depends on S so that it conflicts with the ban imposed by the law or the commercial decision by the insurer.

4.2 Sufficiency

While individual fairness requires that individuals who only differ with respect to the prohibited feature must be treated equally, group fairness is a global notion resulting from the comparison of groups defined by the sensitive feature S. Group fairness requires that benefits or harms resulting from the adoption of a pricing structure $\pi(\cdot)$ are distributed fairly

across these groups, where the fair distribution is defined from an appropriate group-fairness criterion. Classical group-fairness criteria include independence, separation and sufficiency. See, e.g., Baumann and Loi (2023) and Charpentier (2024) for a discussion in the context of insurance.

In general, there is no consensus about the fairness criterion and different contexts may call for different criteria. Moreover, fairness criteria are often incompatible so that one criterion must be selected. In accordance with the analysis conducted by Baumann and Loi (2023), sufficiency appears to be an appropriate fairness condition in the context of private insurance. In its strict sense, sufficiency corresponds to conditional independence of Y and S given $\pi(\boldsymbol{X})$. This is a rather strong requirement that rules out demographic parity which requires independence of $\pi(\boldsymbol{X})$ and S. According to Charpentier (2024, Proposition 9.3), if a candidate premium $\pi(\cdot)$ satisfies independence and sufficiency with respect to S then Y and S are independent. Therefore, unless S does not influence Y, no premium $\pi(\cdot)$ can simultaneously fulfill demographic parity and sufficiency.

In this paper, we consider the weaker version of sufficiency retained by Baumann and Loi (2023), defined with the help of conditional mean independence of Y and S given $\pi(\boldsymbol{X})$. The conditional mean independence concept is used in econometrics to decide whether additional features are needed in a regression model. It turns out that it also possesses an intuitive interpretation in the context of insurance premium. This leads to the following definition.

Definition 4.4. A premium $\pi(\cdot)$ is said to be fair with respect to a sensitive feature $S = X_{j_0}$ valued in $\mathcal{S}$ if

$$\mathrm{E}[Y|\pi(\boldsymbol{X}) = p, S = s] = \mathrm{E}[Y|\pi(\boldsymbol{X}) = p, S = s'] \text{ for all } p \text{ and } s, s' \in \mathcal{S}. \tag{4.1}$$

The requirement (4.1) is intuitively appealing. It just means that the expected benefits are equal across groups created by S provided the same amount of premium is paid. This argument may be used by the insurer as a proof that the use of S in premium calculation is responsible and protects policyholders' rights because the return from the insurance operation (measured by the amount Y paid by the insurer) is the same, on average, as long as the premium is identical.

If $Y \in \{0, 1\}$ then (4.1) corresponds to the conditional independence of Y and S given $\pi(\boldsymbol{X})$. However, for more general responses like claim frequencies and severities, (4.1) is a weaker requirement. Conditional mean independence can be tested from data. The reader is referred to Lundborg (2022) for a survey and to Zhang et al. (2025) for recent developments on testing procedures.

5 Multicalibration

5.1 Definition

Given that autocalibration is a central concept in insurance pricing, rooted in financial equilibrium, it is thus natural to combine it with conditional mean independence (4.1) retained as group-fairness notion. It turns out that this corresponds to the concept of multicalibration proposed in the machine learning literature by Hebert-Johnson et al. (2018). This is formally defined next.

Definition 5.1. The commercial premium $\pi(\cdot)$ is said to be multicalibrated with respect to a sensitive feature $S = X_{j_0}$ valued in $\mathcal{S}$ if

$$\mathrm{E}[Y|\pi(\boldsymbol{X}) = p, S = s] = \mathrm{E}[Y|\pi(\boldsymbol{X}) = p, S = s'] = p \text{ for all } p \text{ and } s, s' \in \mathcal{S}. \tag{5.1}$$

Compared to (4.1), we see that (5.1) not only imposes the equality of conditional expectations, but also that they coincide with the premium amount p involved in the conditions. Notice that (5.1) can be equivalently rewritten as

$$\mathrm{E}[Y - \pi(\boldsymbol{X})|\pi(\boldsymbol{X}) = p, S = s] = 0 \text{ for all } p \text{ and } s \in \mathcal{S}.$$

The net losses $Y - \pi(\boldsymbol{X})$ thus average out in any group of policyholders paying the same amount of premium whatever the value s of the sensitive feature S. This expresses the compensation of claims with premiums in these groups and recalls the balance equations at the heart of the method of marginal totals. Also, (5.1) can be rewritten as

$$\mathrm{E}[\mu(\boldsymbol{X}) - \pi(\boldsymbol{X})|\pi(\boldsymbol{X}) = p, S = s] = 0 \text{ for all } p \text{ and } s \in \mathcal{S}.$$

This shows that the pricing errors $\mu(\boldsymbol{X}) - \pi(\boldsymbol{X})$ cancel out on average within groups of policyholders paying the same amount of premium and having the same value s of the sensitive feature S. Multicalibration thus appears to be a very appealing requirement for commercial premiums.

5.2 Properties

Let us first establish that multicalibration reinforces autocalibration.

Property 5.2. *(i) Any multicalibrated premium $\pi(\cdot)$ with respect to $S = X_{j_0}$ is also autocalibrated.*

(ii) If $\pi(\cdot)$ is autocalibrated then conditional mean independence (4.1) *implies multicalibration with respect to $S = X_{j_0}$.*

Proof. For any multicalibrated premium, we have

$$\mathrm{E}[Y|\pi(\boldsymbol{X})] = \mathrm{E}\big[\underbrace{\mathrm{E}[Y|\pi(\boldsymbol{X}), S]}_{=\pi(\boldsymbol{X})}\big|\pi(\boldsymbol{X})\big] = \pi(\boldsymbol{X})$$

so that it is also autocalibrated, as announced under item (i). Considering item (ii), notice that (4.1) together with autocalibration imply

$$\mathrm{E}[Y|\pi(\boldsymbol{X}), S] = \mathrm{E}[Y|\pi(\boldsymbol{X})] = \pi(\boldsymbol{X}).$$

This ends the proof. □

All the properties derived for autocalibrated premiums thus also hold for multicalibrated premiums. In particular, multicalibrated predictors never overfit. This has been formalized for autocalibrated predictors with the help of the convex order. Recall from Denuit et al.

(2005, Section 3.4) that given two random variables Z and T, with finite expectations, T is said to be smaller than Z in the convex order (denoted as $T \preceq_{\text{cx}} Z$) if

$$\mathrm{E}[g(T)] \leq \mathrm{E}[g(Z)] \text{ for all convex functions } g, \tag{5.2}$$

provided the expectations exist. In words, $T \preceq_{\text{cx}} Z$ means that T is less dispersed than Z while T and Z have the same expected value. In particular, $\mathrm{Var}[T] \leq \mathrm{Var}[Z]$.

According to Lemma 3.1 from Wüthrich (2023), we know that $\pi(\boldsymbol{X}) \preceq_{\text{cx}} \mu(\boldsymbol{X})$ holds for any autocalibrated $\pi(\cdot)$. Therefore, any multicalibrated premium $\pi(\cdot)$ with respect to $S = X_{j_0}$ satisfies

$$\pi(\boldsymbol{X}) \preceq_{\text{cx}} \mu(\boldsymbol{X}) \preceq_{\text{cx}} Y.$$

This shows that multicalibration prevents overfitting, in the sense that an autocalibrated premium cannot be more variable than $\mu(\boldsymbol{X})$ nor Y.

We know that $\mu(\boldsymbol{X})$ is autocalibrated as

$$\begin{aligned} \mathrm{E}[Y|\mu(\boldsymbol{X})] &= \mathrm{E}\big[\mathrm{E}[Y|\mu(\boldsymbol{X}), \boldsymbol{X}]\big|\mu(\boldsymbol{X})\big] \\ &= \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X}]\big|\mu(\boldsymbol{X})\big] = \mu(\boldsymbol{X}). \end{aligned}$$

The next result shows that $\mu(\boldsymbol{X})$ is also multicalibrated with respect to $S = X_j$ for every $j \in \{1, \ldots, p\}$.

Property 5.3. *The best-estimate price $\mu(\boldsymbol{X})$ is multicalibrated with respect to any $S = X_j$, $j = 1, 2, \ldots, p$.*

Proof. It suffices to write

$$\begin{aligned} \mathrm{E}[Y|\mu(\boldsymbol{X}), S] &= \mathrm{E}\big[\mathrm{E}[Y|\mu(\boldsymbol{X}), S, \boldsymbol{X}]\big|\mu(\boldsymbol{X}), S\big] \\ &= \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X}]\big|\mu(\boldsymbol{X}), S\big] \\ &= \mu(\boldsymbol{X}), \end{aligned}$$

which ends the proof. □

Since pure premiums correspond to conditional expectations, they can thus be consistently estimated only if the expected loss is minimum for the mean response. The class of Bregman loss functions is known to be strictly consistent for the (conditional) mean functional so that they are the only meaningful loss functions in the context of insurance pricing. For a convex function $\ell(\cdot)$, recall that the Bregman loss function $L(\cdot, \cdot)$ is defined as

$$L(y, m) = \ell(y) - \ell(m) - \ell'(m)(y - m) \tag{5.3}$$

where ℓ' denotes the subgradient of the convex function ℓ. A premium $\pi_2(\cdot)$ outperforms $\pi_1(\cdot)$ for the Bregman loss function L in (5.3) if $\mathrm{E}[L(Y, \widehat{\pi}_2(\boldsymbol{X}))] \leq \mathrm{E}[L(Y, \widehat{\pi}_1(\boldsymbol{X}))]$. Since multicalibrated predictors are also autocalibrated, we know from Theorem 3.1 in Krüger and Ziegel (2021) that given two multicalibrated premiums $\pi_1(\cdot)$ and $\pi_2(\cdot)$, the inequality $\mathrm{E}[L(Y, \pi_2(\boldsymbol{X}))] \leq \mathrm{E}[L(Y, \pi_1(\boldsymbol{X}))]$ holds true for every Bregman loss function L in (5.3) if, and only if, $\pi_1(\boldsymbol{X}) \preceq_{\text{cx}} \pi_2(\boldsymbol{X})$. This shows that the convex order is the appropriate tool to compare the relative performances of multicalibrated premiums.

6 Multibalance correction

Balance correction has been proposed as a practical way to restore autocalibration by Denuit et al. (2021). Precisely, the balance-corrected version $\pi_{\mathrm{bc}}(\cdot)$ of the candidate premiums $\pi(\cdot)$ is defined as

$$\pi_{\mathrm{bc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X})]. \tag{6.1}$$

The new premium $\pi_{\mathrm{bc}}(\cdot)$ obtained from (6.1) is always autocalibrated as shown in Proposition 3.3 in Wüthrich (2023) without the monotonicity condition assumed in Denuit et al. (2021). The intuition behind (6.1) is that $\pi(\cdot)$ is informative but not necessarily on the right scale, so that financial equilibrium defining autocalibration is violated. The conditional expectation (6.1) corrects the candidate premium by retaining the order induced by $\pi(\cdot)$ but averaging insurance losses corresponding to the same value of $\pi(\boldsymbol{X})$.

Inspired from balance correction (6.1) restoring autocalibration, let us introduce multibalance correction restoring multicalibration.

Definition 6.1. The multibalance-corrected version $\pi_{\mathrm{mbc}}(\cdot)$ of the premium $\pi(\cdot)$ with respect to S is defined as

$$\pi_{\mathrm{mbc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S]. \tag{6.2}$$

The next result shows that the premium $\pi_{\mathrm{mbc}}(\cdot)$ obtained by multibalance correction (6.2) is always multicalibrated with respect to S.

Proposition 6.2. *The multibalance-corrected version $\pi_{\mathrm{mbc}}(\cdot)$ of the premium $\pi(\cdot)$ defined in* (6.2) *is multicalibrated with respect to $S = X_{j_0}$.*

Proof. The announced result follows from

$$\begin{aligned}
\mathrm{E}[Y|\pi_{\mathrm{mbc}}(\boldsymbol{X}), S] &= \mathrm{E}\big[Y\big|\mathrm{E}[Y|\pi(\boldsymbol{X}), S], S\big] \\
&= \mathrm{E}\Big[\mathrm{E}\big[Y\big|\mathrm{E}[Y|\pi(\boldsymbol{X}), S], S, \pi(\boldsymbol{X})\big]\Big|\mathrm{E}[Y|\pi(\boldsymbol{X}), S], S\Big] \\
&= \mathrm{E}\Big[\mathrm{E}\big[Y\big|S, \pi(\boldsymbol{X})\big]\Big|\mathrm{E}[Y|\pi(\boldsymbol{X}), S], S\Big] \\
&= \mathrm{E}\big[\pi_{\mathrm{mbc}}(\boldsymbol{X})\big|\pi_{\mathrm{mbc}}(\boldsymbol{X}), S\big] \\
&= \pi_{\mathrm{mbc}}(\boldsymbol{X}),
\end{aligned}$$

which ends the proof. □

It turns out that multibalance correction improves the performance of the premium with respect to any Bregman loss function. This is precisely stated next.

Proposition 6.3. *For any premium π, we have*

$$\mathrm{E}\big[L\big(Y, \pi(\boldsymbol{X})\big)\big] \geq \mathrm{E}\big[L\big(Y, \pi_{\mathrm{bc}}(\boldsymbol{X})\big)\big] \geq \mathrm{E}\big[L\big(Y, \pi_{\mathrm{mbc}}(\boldsymbol{X})\big)\big] \geq \mathrm{E}\big[L\big(Y, \mu(\boldsymbol{X})\big)\big]$$

for all loss functions (5.3).

Proof. Since $\mathrm{E}[Y|\pi(\boldsymbol{X})] \preceq_{\mathrm{cx}} \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$, we have $\pi_{\mathrm{bc}}(\boldsymbol{X}) \preceq_{\mathrm{cx}} \pi_{\mathrm{mbc}}(\boldsymbol{X}) \preceq_{\mathrm{cx}} \mu(\boldsymbol{X})$. The announced result then follows from Proposition 4.5 in Denuit and Trufin (2024) and the equivalence of Bregman dominance and convex order for autocalibrated predictors. □

Remark 6.4. The multibalance-corrected premium obtained from a candidate premium $\pi(\cdot)$ does not necessarily coincide with the multibalance-corrected premium obtained from its balance-corrected version $\pi_{\mathrm{bc}}(\cdot)$. Indeed,

$$\begin{aligned}\mathrm{E}[Y|\pi_{\mathrm{bc}}(\boldsymbol{X}),S] &= \mathrm{E}\big[\mathrm{E}[Y|\pi(\boldsymbol{X}),\pi_{\mathrm{bc}}(\boldsymbol{X}),S]\big|\pi_{\mathrm{bc}}(\boldsymbol{X}),S\big]\\ &= \mathrm{E}\big[\mathrm{E}[Y|\pi(\boldsymbol{X}),S]\big|\pi_{\mathrm{bc}}(\boldsymbol{X}),S\big]\\ &= \mathrm{E}\big[\pi_{\mathrm{mbc}}(\boldsymbol{X})\big|\pi_{\mathrm{bc}}(\boldsymbol{X}),S\big].\end{aligned}$$

Hence, the multibalance-corrected premium based on $\pi_{\mathrm{bc}}(\cdot)$ is obtained by averaging the multibalance-corrected premium based on $\pi(\cdot)$ over the level sets of $\pi_{\mathrm{bc}}(\cdot)$, for each value of S. As a consequence, the two multibalance-corrected premiums may differ, since this averaging can remove information relevant for predicting Y once S is given. They coincide if, and only if, conditioning with respect to $\pi_{\mathrm{bc}}(\boldsymbol{X})$ retains all the predictive information carried by $\pi(\boldsymbol{X})$ for each value of S (that is, if and only if $\pi(\boldsymbol{X})$ is measurable with respect to $\sigma(\pi_{\mathrm{bc}}(\boldsymbol{X}),S)$). A simple sufficient condition for this to hold is that the mapping $p \mapsto \mathrm{E}[Y|\pi(\boldsymbol{X})=p]$ be one-to-one on the range of $\pi(\boldsymbol{X})$, so that $\pi(\boldsymbol{X})$ can be recovered from $\pi_{\mathrm{bc}}(\boldsymbol{X})$. In particular, this condition is satisfied when the mapping $p \mapsto \mathrm{E}[Y|\pi(\boldsymbol{X})=p]$ is strictly increasing, which means that $\pi(\boldsymbol{X})$ induces the same ordering of risks as $\mu(\boldsymbol{X})=\mathrm{E}[Y|\boldsymbol{X}]$.

To conclude this section, we consider the case discussed in Remark 4.3, where the sensitive feature S is excluded from the set of admissible rating factors, either because of regulation or of a commercial decision made by the insurer. The next result shows how multibalance correction provides an illustration of the structural limitation highlighted in Remark 4.3.

Proposition 6.5. *Assume that S is excluded from the set of admissible rating factors $\boldsymbol{X}$. Then, for any admissible premium $\pi(\cdot)$, the following statements hold.*

(i) The corresponding premium $\pi_{\mathrm{mbc}}(\cdot)$ is multicalibrated with respect to S, in the sense that

$$\mathrm{E}[Y|\pi_{\mathrm{mbc}}(\boldsymbol{X},S),S] = \pi_{\mathrm{mbc}}(\boldsymbol{X},S).$$

(ii) In general, $\pi_{\mathrm{mbc}}(\cdot)$ depends on S and therefore does not define an admissible premium.

(iii) The multibalance-corrected premium $\pi_{\mathrm{mbc}}(\cdot)$ does not depend on S if, and only if,

$$\mathrm{E}[Y|\pi(\boldsymbol{X}),S] = \mathrm{E}[Y|\pi(\boldsymbol{X})],$$

that is, if, and only if, Y and S are conditionally mean-independent, given the premium $\pi(\boldsymbol{X})$.

(iv) The multibalance correction can be written as

$$\pi_{\mathrm{mbc}}(\boldsymbol{X},S) = \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X},S]\big|\pi(\boldsymbol{X}),S\big].$$

In particular, $\pi_{\mathrm{mbc}}(\cdot)$ coincides with $\mathrm{E}[Y|\boldsymbol{X},S]$ if, and only if, $\mathrm{E}[Y|\boldsymbol{X},S]$ is measurable with respect to $\sigma(\pi(\boldsymbol{X}),S)$.

Proof. By definition of the multibalance correction, $\pi_{\text{mbc}}(\boldsymbol{X}, S) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$. Therefore,

$$\begin{aligned}\mathrm{E}[Y|\pi_{\text{mbc}}(\boldsymbol{X}, S), S] &= \mathrm{E}\big[\mathrm{E}[Y|\pi(\boldsymbol{X}), \pi_{\text{mbc}}(\boldsymbol{X}, S), S]\big|\pi_{\text{mbc}}(\boldsymbol{X}, S), S\big] \\ &= \mathrm{E}\big[\mathrm{E}[Y|\pi(\boldsymbol{X}), S]\big|\pi_{\text{mbc}}(\boldsymbol{X}, S), S\big] \\ &= \mathrm{E}[\pi_{\text{mbc}}(\boldsymbol{X}, S)|\pi_{\text{mbc}}(\boldsymbol{X}, S), S] \\ &= \pi_{\text{mbc}}(\boldsymbol{X}, S).\end{aligned}$$

This proves the statement under item (i).

Considering items (ii)-(iii), notice that $\pi_{\text{mbc}}(\boldsymbol{X}, S) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$ generally depends on S. Hence, $\pi_{\text{mbc}}(\cdot)$ is independent of S if, and only if,

$$\mathrm{E}[Y|\pi(\boldsymbol{X}), S] = \mathrm{E}[Y|\pi(\boldsymbol{X})],$$

which proves the statements under items (ii)-(iii).

Finally,

$$\begin{aligned}\mathrm{E}[Y|\pi(\boldsymbol{X}), S] &= \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X}, \pi(\boldsymbol{X}), S]\big|\pi(\boldsymbol{X}), S\big] \\ &= \mathrm{E}\big[\mathrm{E}[Y|\boldsymbol{X}, S]\big|\pi(\boldsymbol{X}), S\big]\end{aligned}$$

proves the statement under item (iv) and ends the proof. □

Proposition 6.5 provides an illustration of Remark 4.3. When S is excluded from the set of admissible rating factors $\boldsymbol{X}$, multicalibration can only be achieved in the exceptional case where Y and S are conditionally mean-independent, given the premium. Otherwise, enforcing multicalibration necessarily leads to premiums that depend on S and are thus inadmissible.

7 Practical implementation of multicalibration

This section proposes practical ways to make a candidate premium $\pi(\cdot)$ multicalibrated with respect to a sensitive feature S included in $\boldsymbol{X}$. The approaches differ according to the format of S.

7.1 Categorical sensitive feature

Assume that the sensitive feature S is categorical, with levels $\ell \in \mathcal{S}$. In principle, multicalibration could be implemented by making $\pi(\cdot)$ autocalibrated separately in every group $S = \ell$, by replacing $\pi(\cdot)$ with its balance-corrected version $\pi_{\text{bc}}(\cdot)$ for every level of S. However, this simple approach is problematic if exposures are limited for some groups, so that non-parametric smoothing does not work.

This section presents two empirical approaches to multicalibration. The first one implements multicalibration through an explicit multibalance correction, while the second one enforces multicalibration directly through an iterative bias-correction procedure. Throughout this section, we consider a dataset of observations $(N_i, \boldsymbol{X}_i, S_i, w_i)$, $i = 1, \ldots, n$, where N_i denotes the claim count, $S_i = X_{i,j_0} \in \mathcal{S}$ the categorical sensitive feature and $w_i \geq 0$ known weights (typically exposures). In practice, we use the observed claim frequency $Y_i = \frac{N_i}{w_i}$ as response variable, obtained by dividing the claim count by the exposure.

7.1.1 Direct empirical multibalance correction via groupwise regression

Multibalance correction amounts to restoring balance separately within each group defined by the sensitive feature S. That is, instead of enforcing autocalibration globally, the balance condition is imposed conditionally on $S = \ell$, for every $\ell \in \mathcal{S}$. This principle can be implemented in several ways, depending on how the conditional mean function $\mathrm{E}[Y|\pi(\boldsymbol{X}), S = \ell]$ is estimated within each group. Examples include local linear regression as in Ciatto et al. (2023) or isotonic regression as in Wüthrich (2023).

To fix ideas, we describe here an implementation based on isotonic regression. For each level $\ell \in \mathcal{S}$, consider the conditional mean function

$$m_\ell(p) = \mathrm{E}\big[Y|\pi(\boldsymbol{X}) = p, S = \ell\big], \qquad p > 0.$$

An empirical implementation of multibalance correction is obtained by estimating $m_\ell(\cdot)$ nonparametrically within each group $S = \ell$, under a monotonicity constraint in p. Specifically, we define $\widehat{m}_\ell$ as the solution of the weighted isotonic regression problem

$$\widehat{m}_\ell \in \underset{m \text{ non-decreasing}}{\operatorname{argmin}} \sum_{i:\, S_i = \ell} w_i \big(Y_i - m(\pi_i)\big)^2, \qquad \ell \in \mathcal{S}, \tag{7.1}$$

restricting the search to non-decreasing conditional mean function, that is, such that $p \leq p'$ implies $m(p) \leq m(p')$ on the range of $\{\pi_i : S_i = \ell\}$. The resulting empirical multibalance-corrected premium is then obtained by the plug-in rule

$$\widehat{\pi}_{\mathrm{mbc}}(\boldsymbol{X}_i) = \widehat{m}_{S_i}\big(\pi(\boldsymbol{X}_i)\big), \qquad i = 1, \ldots, n, \tag{7.2}$$

where $\widehat{m}_\ell(\cdot)$ is the solution to (7.1).

This construction is the direct analogue of balance correction for autocalibration implemented via isotonic regression by Wüthrich (2023), except that the recalibration is performed separately within each level of the categorical sensitive feature S. By construction, the premium $\widehat{\pi}_{\mathrm{mbc}}(\cdot)$ in (7.2) is an empirical implementation of the multibalance-corrected premium

$$\pi_{\mathrm{mbc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S],$$

and therefore achieves multicalibration by enforcing balance conditionally on S.

When some levels $\ell \in \mathcal{S}$ carry limited exposure, the fully nonparametric estimation in (7.1) may be unstable, yielding step functions with high variance and poor out-of-sample behavior. The next section therefore introduces a regularized procedure that relaxes the fully non-parametric group-wise approach. Rather than estimating conditional mean functions independently within each level of S, the procedure enforces balance on a finite collection of (π, S)-cells and stabilizes local corrections by borrowing strength across groups through exposure-driven shrinkage.

7.1.2 Regularized empirical enforcement of multicalibration via iterative bias correction

We now introduce a regularized empirical procedure that enforces multicalibration directly, without explicitly applying the multibalance correction. Starting from a baseline premium,

the procedure modifies the premium iteratively so that, at the chosen level of discretization, average residuals vanish within each group defined by the current premium level and the sensitive feature. Exposure-driven shrinkage is used to stabilize local corrections when data are sparse.

Binning and residual biases. Fix an integer $K \geq 1$. Starting from an initial premium $\pi^{(0)}(\cdot) = \pi(\cdot)$, consider at iteration $j \geq 0$ a partition of the range of $\pi^{(j)}(\boldsymbol{X})$ into K disjoint intervals $\{B_1^{(j)}, \ldots, B_K^{(j)}\}$ (for instance empirical quantile bins). For each observation $i = 1, \ldots, n$, define the residual

$$R_i^{(j)} = Y_i - \pi^{(j)}(\boldsymbol{X}_i).$$

For each bin $k \in \{1, \ldots, K\}$ and each level $\ell \in \mathcal{S}$, define the index set

$$I_{k\ell}^{(j)} = \big\{ i : \ \pi^{(j)}(\boldsymbol{X}_i) \in B_k^{(j)}, \ S_i = \ell \big\},$$

together with the corresponding exposure

$$w_{k\ell}^{(j)} = \sum_{i \in I_{k\ell}^{(j)}} w_i, \qquad w_{k\bullet}^{(j)} = \sum_{\ell \in \mathcal{S}} w_{k\ell}^{(j)}.$$

Whenever $w_{k\ell}^{(j)} > 0$, define the cell-wise and pooled bin-wise mean residuals as

$$\widehat{b}_{k\ell}^{(j)} = \frac{1}{w_{k\ell}^{(j)}} \sum_{i \in I_{k\ell}^{(j)}} w_i R_i^{(j)}, \qquad \widehat{b}_k^{(j)} = \frac{1}{w_{k\bullet}^{(j)}} \sum_{\ell \in \mathcal{S}} \sum_{i \in I_{k\ell}^{(j)}} w_i R_i^{(j)}.$$

The quantity $\widehat{b}_{k\ell}^{(j)}$ is the exposure-weighted sample analogue of $\mathrm{E}\Big[Y - \pi^{(j)}(\boldsymbol{X}) \,\Big|\, \pi^{(j)}(\boldsymbol{X}) \in B_k^{(j)}, \ S = \ell \Big]$.

Exposure-driven shrinkage. When $w_{k\ell}^{(j)}$ is small, the estimate $\widehat{b}_{k\ell}^{(j)}$ is unstable. To stabilize the correction, we shrink $\widehat{b}_{k\ell}^{(j)}$ toward the pooled bin-level quantity $\widehat{b}_k^{(j)}$ according to

$$\widetilde{b}_{k\ell}^{(j)} = z_{k\ell}^{(j)} \widehat{b}_{k\ell}^{(j)} + \big(1 - z_{k\ell}^{(j)}\big) \widehat{b}_k^{(j)}, \qquad z_{k\ell}^{(j)} = \frac{w_{k\ell}^{(j)}}{w_{k\ell}^{(j)} + c}, \tag{7.3}$$

where $c > 0$ is a tuning parameter controlling the amount of pooling across groups.

Update step. Let $\eta \in (0, 1]$ be a step size. For each observation i such that $\pi^{(j)}(\boldsymbol{X}_i) \in B_k^{(j)}$ and $S_i = \ell$, update

$$\pi^{(j+1)}(\boldsymbol{X}_i) = \pi^{(j)}(\boldsymbol{X}_i) + \eta \, \widetilde{b}_{k\ell}^{(j)}. \tag{7.4}$$

This update adjusts the premium by an exposure-regularized estimate of the residual bias in the corresponding (π, S)-cell.

Stopping criterion. Iterations are stopped when the remaining correction becomes negligible. A convenient criterion is to retain the solution such that

$$\max_{k,\ell:\, w_{k\ell}^{(j)}>0} \frac{\left|\eta\, \overline{\widetilde{b}}_{k\ell}^{(j)}\right|}{\overline{\pi}_{k\ell}^{(j)}} \leq \delta,$$

where $\overline{\pi}_{k\ell}^{(j)}$ and $\overline{\widetilde{b}}_{k\ell}^{(j)}$ denote exposure-weighted averages of $\pi^{(j)}(\boldsymbol{X}_i)$ and $\widetilde{b}_{k\ell}^{(j)}$ over $i \in I_{k\ell}^{(j)}$, and $\delta > 0$ is a tolerance parameter selected by the actuary.

Interpretation. At convergence, the procedure yields a premium $\pi^{(J)}(\cdot)$ such that, for each bin $B_k^{(J)}$ and each level $\ell \in \mathcal{S}$, the exposure-weighted mean residual within the corresponding cell is close to zero. Equivalently, at the resolution induced by the binning scheme, no systematic average deviation between Y and $\pi^{(J)}(\boldsymbol{X})$ remains within groups defined by $(\pi^{(J)}(\boldsymbol{X}), S)$.

This property is the discrete counterpart of the multicalibration requirement $\mathrm{E}[Y|\pi^{(J)}(\boldsymbol{X}) = p,\ S = \ell] = p$, $p > 0$, $\ell \in \mathcal{S}$, with exact conditioning replaced by conditioning on bin membership. Because the bins are updated along the iterations and are defined using the current premium $\pi^{(j)}(\cdot)$, the resulting premium $\pi^{(J)}(\cdot)$ should be understood as enforcing multicalibration through a fixed-point property, rather than as the result of a single application of the multibalance correction.

In contrast, if the bins are fixed using the baseline premium $\pi(\cdot)$ and a single update is performed, the procedure reduces to a discretized empirical approximation of the multibalance-corrected premium $\pi_{\mathrm{mbc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$, at the chosen binning resolution.

On the role of credibility weights. In the classical Bühlmann–Straub framework, credibility arises from an explicit stochastic data-generating structure. A latent parameter Θ governs repeated observations at individual risk level, leading to the well-known variance decomposition into process variance and structural variance.

In the present multicalibration setting, no such structural latent parameter is postulated. The procedure is entirely cross-sectional and algorithmic. For a given premium bin k and group ℓ, the quantity

$$b_{k\ell} = \mathrm{E}[Y - \pi(X)|\pi(X) \in B_k,\ S = \ell]$$

may be viewed as a latent systematic bias parameter attached to the cell (k, ℓ), but this parameter is induced by the current premium rather than by an underlying stochastic heterogeneity. The observations within each cell simply play the role of repeated measurements used to estimate this conditional mean residual.

The analogy with credibility theory is therefore algebraic rather than structural. As in the Bühlmann–Straub model, the empirical estimator $\widehat{b}_{k\ell}$ is noisy when exposure $w_{k\ell}$ is small. A variance decomposition of the form

$$\mathrm{Var}[\widehat{b}_{k\ell}] = \frac{\mathrm{E}[\sigma_{k\ell}^2]}{w_{k\ell}} + \mathrm{Var}[b_{k\ell}]$$

motivates shrinking the cell-specific estimator toward the pooled bin-level quantity $\widehat{b}_k$, leading to the weight

$$z_{k\ell} = \frac{w_{k\ell}}{w_{k\ell} + c}.$$

The resulting correction

$$\widetilde{b}_{k\ell} = z_{k\ell}\widehat{b}_{k\ell} + (1 - z_{k\ell})\widehat{b}_k$$

should therefore be interpreted primarily as an exposure-driven regularization device. It stabilizes local bias estimates in low-exposure cells while allowing more granular adjustments where sufficient data are available.

In contrast with classical credibility theory, however, we do not assume an explicit hierarchical model with a stochastic latent parameter Θ. The credibility weight here provides a principled and actuarially interpretable shrinkage mechanism, but the multicalibration procedure itself remains a post-processing algorithm rather than the consequence of a fully specified probabilistic credibility model.

7.1.3 Numerical illustrations

We apply the proposed multicalibration procedure to the data set used by Noll et al. (2018). This data set contains a French motor third-party liability (MTPL) insurance portfolio comprising of 677 991 entries, each corresponding to a policy. Twelve variables are observed for each of these policies, listed in Table 7.1. We refer to Noll et al. (2018) for a detailed description of the complete data set.

Table 7.1: Description of the variables in the `freMTPL2freq` dataset.

Variable	Description
`IDpol`	Policy number.
`ClaimNb`	Number of claims.
`Exposure`	Total exposure in yearly units.
`Area`	Area code (A,B,C,D,E or F).
`Region`	Region of France where the policy is registered, categorical variable with 22 levels.
`Density`	Population density (number of inhabitants per square km) of the city where the policyholder lives.
`BonusMalus`	Bonus-malus level, ranging between 50 and 230 (reference level: 100).
`DrivAge`	Age of the driver of the vehicle, in years.
`VehAge`	Age of the vehicle, in years.
`VehGas`	Type of fuel used for the vehicle (diesel or regular gas).
`VehPower`	Power of the car, categorical variable ranging from 4 to 15.
`VehBrand`	Brand of the car, categorical variable with 11 levels.

The outcome variable Y is the observed claim frequency, defined as $Y = \frac{N}{w}$, where $N = \texttt{ClaimNb}$ and $w = \texttt{Exposure}$ measured in policy-years. For the multicalibration analysis, we consider a discrete sensitive attribute

$$S^{\mathrm{d}} = \texttt{bin\_VehAge},$$

obtained by grouping `VehAge` into three bins of approximately equal size: $(0, 3]$, $(3, 9]$ and > 9. The rating variables are

$$\boldsymbol{X} = \{\texttt{Area}, \texttt{Region}, \texttt{Density}, \texttt{BonusMalus}, \texttt{DrivAge}, \texttt{VehAge}, \texttt{VehGas}, \texttt{VehBrand}\},$$

while `bin_VehAge` is used solely for calibration.

The data are randomly split into training (60%), validation (20%), and test (20%) sets. Baseline premiums are obtained from a Poisson GAM with log-link and exposure offset. Precisely, we assume that

$$N_i|\boldsymbol{X}_i \sim \mathrm{Poisson}\big(w_i\lambda(\boldsymbol{X}_i)\big),$$

with

$$\ln \lambda(\boldsymbol{X}_i) = \beta_0 + \sum_m f_m(X_{i,m}).$$

The corresponding frequency premium is then given by

$$\pi(\boldsymbol{X}_i) = \lambda(\boldsymbol{X}_i) = \mathrm{E}[Y_i|\boldsymbol{X}_i].$$

Although $\pi(\cdot)$ is calibrated on average, it displays systematic residual bias across premium levels and across vehicle-age groups.

Starting from $\pi(\cdot)$, we consider five premium constructions. The first is the unawareness premium itself. The second corresponds to autocalibration obtained as a limiting case of the iterative multicalibration procedure when the credibility parameter c is taken to be very large. In that regime,

$$z_{k\ell} = \frac{w_{k\ell}}{w_{k\ell} + c} \approx 0,$$

so that only the bin-level bias enters the update, enforcing autocalibration. As a non-iterative benchmark targeting the same objective, we also consider classical balance correction $\pi_{\mathrm{bc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi_0(\boldsymbol{X})]$ based on weighted isotonic regression, Full multicalibration is obtained for finite c, combining bin-level and cell-level biases through exposure-driven shrinkage so as to enforce $\mathrm{E}[Y - \pi(\boldsymbol{X})|\mathrm{bin}, S] \approx 0$. Finally, multibalance correction is defined by $\pi_{\mathrm{mbc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$, estimated via weighted isotonic regression within each vehicle-age group. Throughout, we use $K = 10$ premium bins, step size $\eta = 0.2$, and stopping threshold $\delta = 0.01$. Residual bias is computed as exposure-weighted averages within each premium bin and vehicle-age group.

Figure 7.1 illustrates the residual structure. The baseline premiums exhibit a clear monotone distortion across premium bins, with pronounced under-pricing in the highest bins.

After autocalibration (iterative or isotonic), this global pattern disappears and the overall residual curve becomes essentially flat. However, significant differences remain across vehicle-age groups within the same premium bin, indicating persistent conditional bias with respect to S, as shown in the plot on the second row.

Applying full multicalibration or multibalance correction removes these group-specific distortions. In both cases, the colored curves align closely around zero across all bins. Multicalibration achieves this through iterative bias updates with exposure-driven shrinkage, producing smooth residual profiles. Multibalance correction enforces conditional calibration directly via group-wise isotonic regression. Minor local oscillations reflect the step-wise nature of the estimator and the absence of cross-group regularization. Overall, both methods successfully restore conditional balance while preserving premium-level calibration, with the iterative approach yielding slightly more regular patterns.

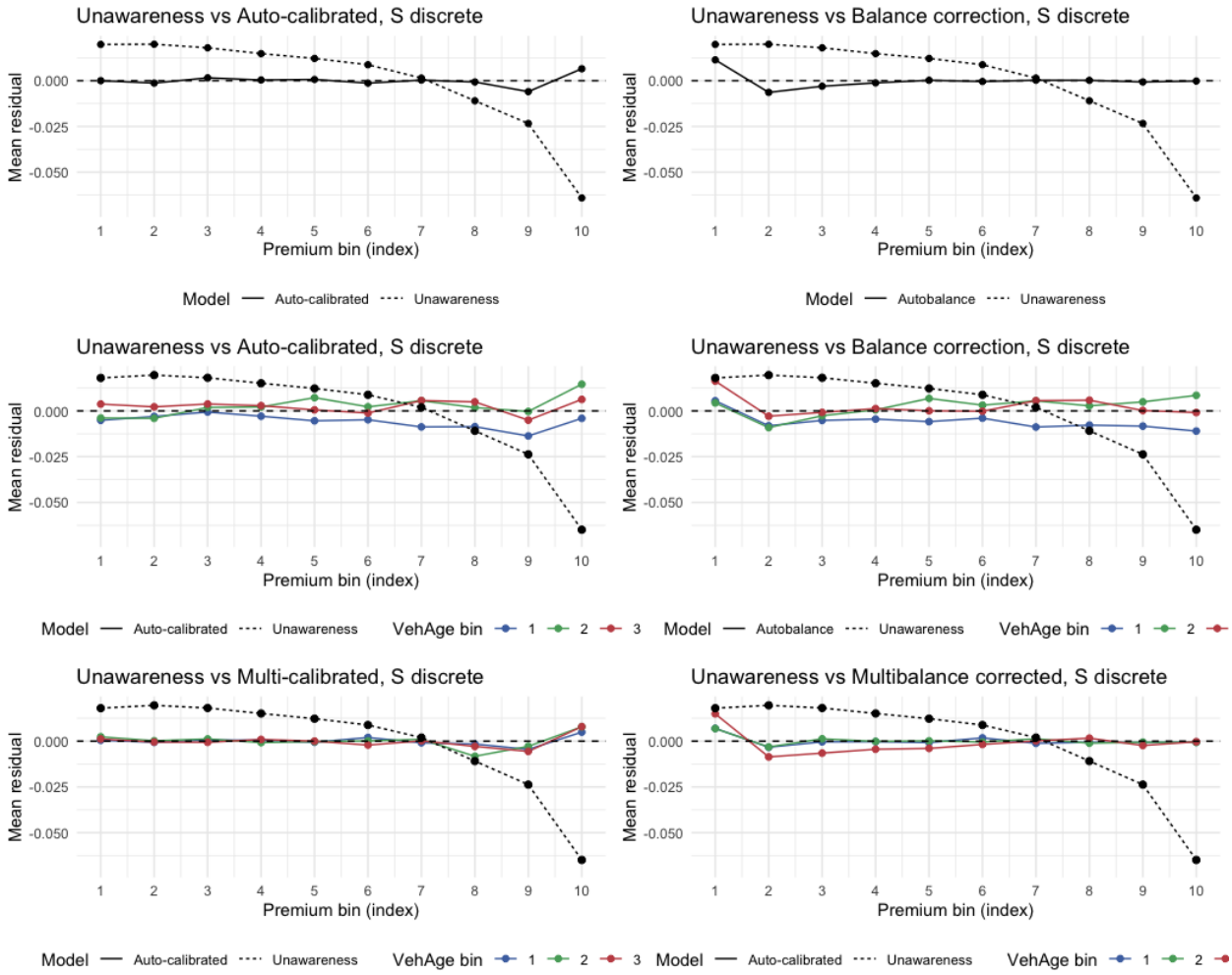


Figure 7.1: Mean residual pricing bias across premium bins for discrete vehicle-age groups. Top row: baseline vs. autocalibration. Middle row: autocalibration via iterative algorithm (left) and balance correction via isotonic regression (right). Bottom row: multicalibration (iterative) and multibalance correction (group-wise isotonic regression).

Table 7.2 reports out-of-sample performance on the test set using Poisson deviance and Gini coefficient. The baseline model exhibits by far the highest deviance, confirming that miscalibration in the premium score materially affects predictive accuracy. Although the baseline GAM is estimated by maximum likelihood in the covariate space, it does not guarantee that the induced premium satisfies autocalibration. Systematic distortions across premium levels therefore translate into likelihood losses when aggregated over the portfolio.

Autocalibration yields a substantial reduction in deviance relative to the baseline, showing that correcting premium-level distortions already improves adequacy. Introducing conditional calibration with respect to vehicle age further reduces deviance: the multicalibrated premium achieves the lowest deviance overall, improving upon the autocalibrated specification. In comparison, the non-iterative balance and multibalance corrections achieve slightly higher deviance values. Overall, the iterative procedures (autocalibration and multicalibration) outperform their isotonic counterparts in terms of deviance.

In terms of discriminatory power, all calibrated models substantially improve upon the baseline. The Gini coefficients of the calibrated models are relatively close, with autocalibration and balance correction attaining the highest values.

Table 7.2: Out-of-sample performance comparison on the test set.

Model	Deviance	Gini Coefficient
Baseline	36 310.68	0.5783
Autocalibrated	33 882.79	**0.7237**
Balance correction	33 912.20	0.7221
Multicalibrated	**33 779.73**	0.6876
Multibalance correction	33 808.72	0.6938

7.2 Continuous sensitive feature

We now consider the case where the sensitive feature S is continuous. In this setting, multicalibration enforces $\mathrm{E}[Y|\pi(\boldsymbol{X}) = p,\ S = s] = p$, for (p, s) in the support of $(\pi(\boldsymbol{X}), S)$. In analogy with the categorical case, we present two empirical approaches. The first one implements multicalibration through a direct estimation of the conditional mean surface in the joint (π, S) space. The second one enforces multicalibration through a regularized iterative bias-correction procedure. We work with the same data structure as before.

7.2.1 Direct empirical multibalance correction via bivariate local GLM

When the sensitive feature S is continuous, we estimate the conditional mean surface

$$m(p, s) = \mathrm{E}[Y|\pi(\boldsymbol{X}) = p, S = s]\,,$$

through a local GLM in the joint (π, S) space.

To estimate $m(p, s)$ at a given (p, s), we assign to each observation (N_i, π_i, S_i, w_i) a weight

$$\nu_i(p, s) = \nu\left(\frac{\pi_i - p}{h_\pi}, \frac{S_i - s}{h_S}\right),$$

where $\nu(\cdot,\cdot)$ is a symmetric weight function supported on a bounded set and h_π and $h_S > 0$ are bandwidth parameters controlling locality in each direction. At (p, s), we fit an intercept-only GLM adapted to the nature of the response, using exposure weights w_i. Under the canonical link function g, we obtain the local likelihood equation

$$\sum_{i=1}^{n} \nu_i(p,s)\, w_i\, Y_i = \sum_{i=1}^{n} \nu_i(p,s)\, w_i\, \mu(p,s),$$

whose solution yields

$$\widehat{m}(p,s) = \frac{\sum_{i=1}^{n} \nu_i(p,s)\, w_i\, Y_i}{\sum_{i=1}^{n} \nu_i(p,s)\, w_i}.$$

The response variable is again the observed claim frequency $Y = N/w$, and estimation is performed using an exposure-weighted local GLM. A natural plug-in estimator replaces $\pi(\boldsymbol{X}_i)$ with $\widehat{m}\big(\pi(\boldsymbol{X}_i), S_i\big)$. While this estimator targets conditional balance with respect to (π, S), it does not generally preserve marginal autocalibration with respect to π. To restore marginal consistency, we therefore introduce a centering step. First, we estimate the marginal conditional mean

$$\widehat{m}_0(p) = \mathrm{E}[Y|\pi(\boldsymbol{X}) = p]$$

using a one-dimensional local GLM in the premium score. We then define

$$\delta(p,s) = \widehat{m}(p,s) - \widehat{m}_0(p),$$

and estimate its conditional expectation given p via a one-dimensional local regression. Denoting this estimate by $\widehat{\mathrm{E}}[\delta(p,S)|\pi = p]$, we construct the centered multibalance correction

$$\widehat{\pi}_{\mathrm{mbc}}(\boldsymbol{X}_i) = \widehat{m}_0\big(\pi(\boldsymbol{X}_i)\big) + \delta\big(\pi(\boldsymbol{X}_i), S_i\big) - \widehat{\mathrm{E}}\big[\delta\big(\pi(\boldsymbol{X}_i), S\big)|\pi = \pi(\boldsymbol{X}_i)\big]\,.$$

This centering guarantees that the average correction at each premium level coincides with the marginal balance adjustment while redistributing residual bias across values of S. By construction, $\widehat{\pi}_{\mathrm{mbc}}(\boldsymbol{X}) \approx \mathrm{E}[Y|\pi(\boldsymbol{X}), S]$, while preserving approximate autocalibration in the premium dimension.

This procedure constitutes the continuous analogue of the group-wise multibalance correction described in Section 7.1.1. While it enforces conditional balance in a single smoothing step, its performance depends on the local density of observations in the joint (π, S) space. In regions with limited exposure, two-dimensional smoothing may become unstable, which motivates the regulariaed iterative alternative introduced next.

7.2.2 Regularized empirical enforcement of multicalibration via iterative bias correction

We introduce a regularized empirical procedure that enforces multicalibration sequentially. Rather than estimating the full conditional mean surface in one step, we iteratively remove residual bias in the joint (π, S) space, combining one-dimensional and two-dimensional smoothers with exposure-driven shrinkage to stabilize local corrections.

Local residual biases. Let $\pi^{(0)}(\boldsymbol{X}) = \pi(\boldsymbol{X})$ denote the baseline premium. For iterations $j = 0, 1, 2, \ldots, J$, define the residual

$$R_i^{(j)} = Y_i - \pi^{(j)}(\boldsymbol{X}_i).$$

We estimate two smooth conditional mean functions at each iteration:

$$\widehat{b}^{(j)}(p) = \mathrm{E}\big[Y - \pi^{(j)}(\boldsymbol{X})|\pi^{(j)}(\boldsymbol{X}) = p\big]\,,$$

$$\widehat{b}^{(j)}(p,s) = \mathrm{E}\big[Y - \pi^{(j)}(\boldsymbol{X})|\pi^{(j)}(\boldsymbol{X}) = p,\ S = s\big]\,.$$

The first term corresponds to the marginal premium-level bias and plays the role of the bin-level bias in the discrete algorithm, thereby enforcing autocalibration with respect to the premium. The second captures systematic deviations that remain conditional on both the premium level and the grouping variable. The one-dimensional and two-dimensional smooth regressions are performed in the current premium space $\pi^{(j)}(\boldsymbol{X})$.

Exposure-driven shrinkage. To stabilize local bias estimates in regions with limited data, we introduce credibility weights in the joint (π, S) space. Prior to the iterative procedure, we use a k-nearest-neighbor estimator to approximate the local density of observations around each point (p, s) and convert it into a local effective exposure $w_{\mathrm{loc}}(p, s)$. We define the credibility weight as

$$z(p,s) = \frac{w_{\mathrm{loc}}(p,s)}{w_{\mathrm{loc}}(p,s) + c},$$

where $c > 0$ is a tuning parameter controlling the amount of shrinkage.

In contrast with the categorical case, where credibility weights depend on iteration-specific cell exposures, the quantity $w_{\mathrm{loc}}(p, s)$ reflects intrinsic data availability in the continuous (π, S) space. We therefore compute the weights $z(p, s)$ once and keep them fixed throughout the iterations. This choice ensures that shrinkage captures structural sparsity of the data rather than fluctuations induced by successive premium updates and avoids feedback between correction and regularization.

We define the credibility-weighted conditional correction pointwise as

$$\delta^{(j)}(p,s) = z(p,s)\Big(\widehat{b}^{(j)}(p,s) - \widehat{b}^{(j)}(p)\Big).$$

As in Section 7.2.1, we apply a centering step to preserve autocalibration with respect to the premium level. We therefore define

$$\widetilde{b}^{(j)}(p,s) = \widehat{b}^{(j)}(p) + \delta^{(j)}(p,s) - \mathrm{E}\big[\delta^{(j)}(p,S)|\pi^{(j)}(\boldsymbol{X}) = p\big]\,.$$

We compute the centering term via a one-dimensional local regression of $\delta^{(j)}(\pi^{(j)}(\boldsymbol{X}), S)$ on $\pi^{(j)}(\boldsymbol{X})$.

$$\widetilde{b}^{(j)}(p,s) = \widehat{b}^{(j)}(p) + \delta^{(j)}(p,s) - \mathrm{E}\big[\delta^{(j)}(p,S)|\pi^{(j)}(\boldsymbol{X}) = p\big]\,.$$

We estimate the conditional expectation in the last term using the same one-dimensional local smoothing step in the premium score. This guarantees that the correction integrates to $\widehat{b}^{(j)}(p)$ at each premium level, thereby maintaining marginal autocalibration while redistributing residual bias across values of the continuous grouping variable S.

Update step. For a step size $\eta \in (0, 1]$, we update the premium additively as

$$\pi^{(j+1)}(\boldsymbol{X}_i) = \pi^{(j)}(\boldsymbol{X}_i) + \eta\, \widetilde{b}^{(j)}\big(\pi^{(j)}(\boldsymbol{X}_i), S_i\big).$$

The step-size parameter η controls the speed of adjustment, and we choose it to ensure stable convergence of the algorithm.

Stopping criterion. At initialization, we construct fixed grids for the premium and the grouping variable using quantile-based partitions of $\pi^{(0)}(\boldsymbol{X})$ and S. We use these grids exclusively to evaluate convergence. For each cell (k, ℓ) of this fixed grid, we compute the exposure-weighted averages $\overline{\pi}_{k\ell}^{(j)}$ and $\overline{b}_{k\ell}^{(j)}$ of the current premium and correction. We stop the procedure when

$$\max_{k,\ell} \frac{\left|\eta\, \overline{b}_{k\ell}^{(j)}\right|}{\overline{\pi}_{k\ell}^{(j)}} \leq \delta,$$

for a user-specified tolerance parameter $\delta > 0$.

Interpretation. Upon completion of the algorithm, we obtain a premium $\pi^{(J)}(\cdot)$ for which no systematic residual bias remains at the chosen resolution in the joint (π, S) space.

On the role of credibility weights. As in the categorical case, we interpret the link with classical Bühlmann–Straub credibility in an algebraic rather than structural sense. We do not postulate any latent parameter Θ, and the procedure remains purely cross-sectional and algorithmic.

In the continuous setting, we view

$$b(p, s) = \mathrm{E}[Y - \pi(\boldsymbol{X}) | \pi(\boldsymbol{X}) = p,\ S = s]$$

as a latent systematic bias surface attached to the current premium. Observations located near (p, s) provide repeated information for estimating this conditional mean residual. When local effective exposure $w_{\mathrm{loc}}(p, s)$ is small, these estimates become unstable, which motivates shrinkage toward the marginal bias.

The credibility weight $z(p, s) = \frac{w_{\mathrm{loc}}(p,s)}{w_{\mathrm{loc}}(p,s)+c}$ therefore acts as an exposure-driven regularization device, reducing variance in sparse regions while allowing more granular adjustments where sufficient data are available. We thus use credibility as a stabilization mechanism rather than as the consequence of a hierarchical stochastic model, and the multicalibration procedure remains a post-processing algorithm enforcing conditional balance.

7.2.3 Numerical illustrations

We now illustrate the continuous multicalibration procedures using the same MTPL frequency dataset and data split as in Section 7.1.3. Here, we consider vehicle age as a continuous grouping variable and set

$$S = \texttt{VehAge}.$$

Vehicle age is not used as a rating variable but only as a calibration dimension. We obtain baseline premiums $\pi(\cdot)$ from the same Poisson GAM with log-link and exposure offset described previously. These premiums exhibit systematic residual distortions across premium levels and conditional on vehicle age.

Starting from $\pi(\cdot)$, we consider five premium constructions. The first is the unawareness premium itself. The second corresponds to autocalibration obtained as a limiting case of the iterative algorithm when the credibility parameter c is large, enforcing autocalibration through one-dimensional local regression in the premium score. As a direct non-iterative benchmark targeting the same objective, we also consider balance correction defined by

$$\pi_{\mathrm{bc}}(\boldsymbol{X}) = \mathrm{E}[Y|\pi(\boldsymbol{X})],$$

estimated via a one-dimensional local GLM.

We then compare two approaches enforcing conditional balance with respect to the continuous variable $S = \texttt{VehAge}$. We obtain full multicalibration through the iterative procedure of Section 7.2.2. At each iteration, we estimate the marginal bias $\widehat{b}^{(j)}(p)$ and the joint bias surface $\widehat{b}^{(j)}(p, s)$ via local GLMs in the current premium space, apply exposure-driven shrinkage, and centre the correction as described in Section 7.2.1. The local GLMs are estimated using the `locfit` function in `R` with neighborhood parameter $\alpha = 0.5$, representing the nearest neighbor fraction. We use local linear fits, step size $\eta = 0.2$, and stopping threshold $\delta = 0.01$.

As a direct non-iterative alternative, we implement multibalance correction via the centered bivariate local GLM of Section 7.2.1. Specifically, we estimate the conditional mean surface $m(p, s) = \mathrm{E}[Y|\pi(\boldsymbol{X}) = p,\ S = s]$ using a two-dimensional local Poisson regression in the joint (π, S) space with the same smoothing parameters. We then apply the centering step to preserve marginal autocalibration.

We evaluate residual bias on the validation set using ten premium bins constructed from $\pi(\cdot)$ and for each value of `VehAge`. For visual clarity, these curves are summarised by a shaded envelope corresponding to the minimum and maximum residual bias across vehicle ages within each premium bin. Figure 7.2 displays the residual structure. The baseline premiums show a clear monotone distortion across premium bins, with substantial underpricing in the highest bins. After autocalibration (iterative or direct balance correction), the global residual curve becomes essentially flat, confirming that both approaches enforce autocalibration, as shown on the top panel. However, dispersion across vehicle age remains visible within each premium bin, indicating persistent conditional bias with respect to S, as illustrated by the wider shaded areas in the middle panel of Figure 7.2 for both approaches.

Applying full multicalibration or multibalance correction substantially reduces this dispersion. The shaded regions in the bottom row narrow considerably relative to autocalibration, demonstrating that both methods effectively remove conditional distortions along the continuous vehicle-age dimension. The iterative multicalibration procedure yields the smoothest residual patterns, reflecting the stabilizing effect of shrinkage and repeated bias updates. The direct multibalance correction achieves comparable alignment in a single step, with minor local variations attributable to finite-sample smoothing in two dimensions.

Table 7.3 reports out-of-sample performance on the test set using Poisson deviance and Gini coefficient. The pattern is again consistent with the categorical case discussed in Sec-

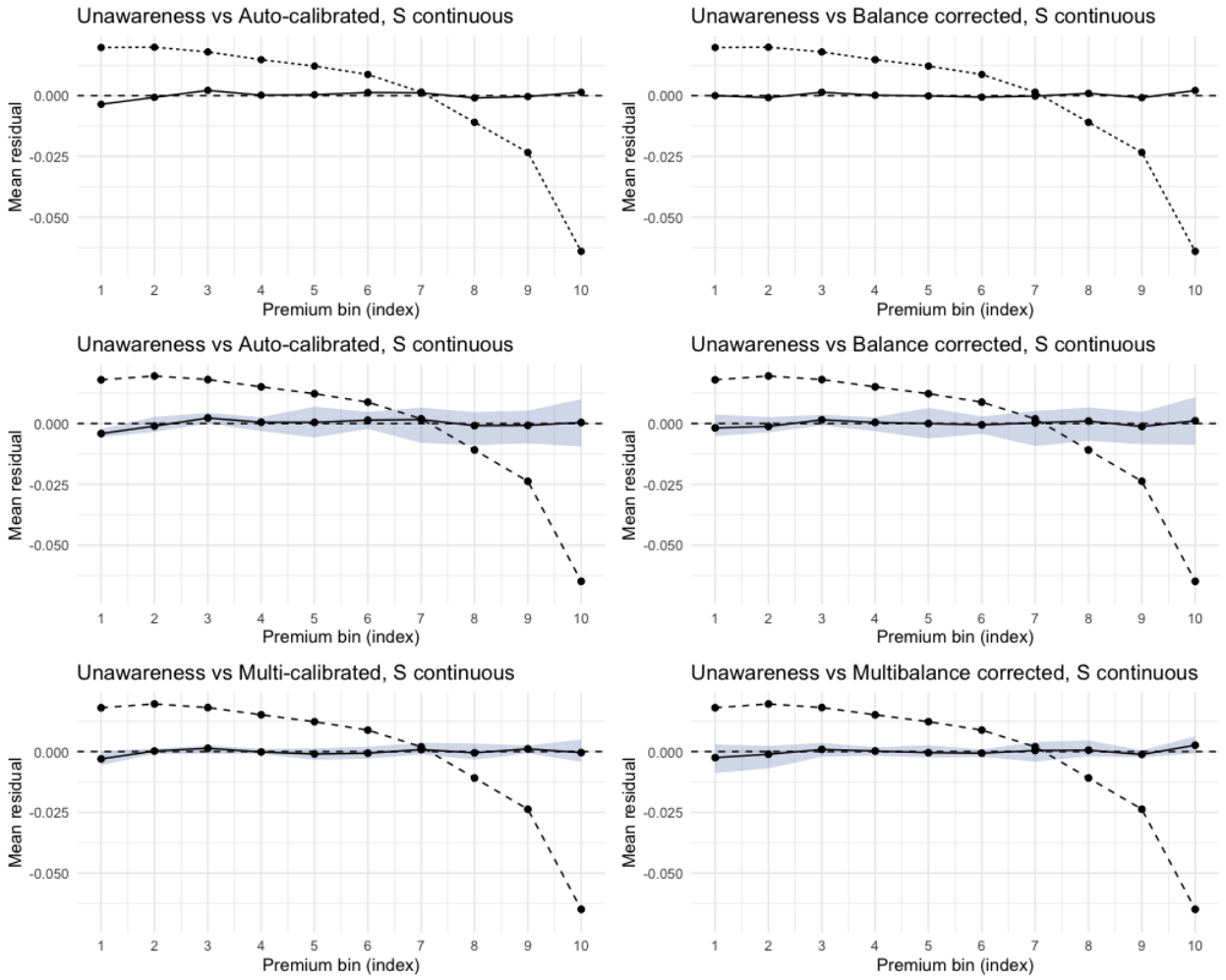


Figure 7.2: Mean residual pricing bias across premium bins for continuous vehicle-age. Top row: baseline vs. autocalibration. Middle row: autocalibration via iterative algorithm (left) and balance correction via isotonic regression (right). Bottom row: multicalibration (iterative) and multibalance correction (groupwise isotonic regression).

tion 7.1.3. The baseline model exhibits the highest deviance, reflecting score-level miscalibration. Autocalibration substantially reduces deviance, confirming that correcting premium-level distortions improves model adequacy.

Introducing conditional calibration with respect to continuous vehicle age yields further improvements. The multicalibrated premium achieves the lowest deviance overall, improving upon the autocalibrated specification. The non-iterative balance and multibalance corrections attain very similar but slightly higher deviance values. Overall, the iterative procedures (autocalibration and multicalibration) provide the best adequacy performance in the continuous setting.

In terms of discriminatory power, all calibrated models markedly improve upon the baseline. The Gini coefficients of the calibrated specifications are relatively close. Multibalance correction attains the highest Gini value, followed closely by balance correction and autocal-

ibration, while multicalibration yields a slightly lower value.

Table 7.3: Out-of-sample performance comparison on the test set.

Model	Deviance	Gini Coefficient
Baseline	36 310.68	0.5783
Autocalibrated	33 793.14	0.6941
Balance correction	33 799.32	0.6973
Multicalibrated	**33 781.86**	0.6904
Multibalance correction	33 814.61	**0.7023**

8 Discussion

This paper implemented multicalibration for insurance premiums. This notion is particularly appealing because it ensures financial equilibrium and guarantees that the expected return of the insurance operation is identical across groups defined on the basis of the sensitive feature. Several methods for implementing multicalibration have been proposed and illustrated on a motor insurance portfolio. Interestingly, multicalibration improves the performances of the candidate premium whereas imposing fairness generally deteriorates accuracy.

Acknowledgements

The first Author gratefully acknowledges funding from the FWO and F.R.S.-FNRS under the Excellence of Science (EOS) programme, project ASTeRISK (40007517). The third Author gratefully acknowledges the support received from the Research Chair ACTIONS under the aegis of the Risk Foundation, an initiative by BNP Paribas Cardif and the French Institute of Actuaries.

References

- Barocas, S., Hardt, M., Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org.

- Baumann, J., Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. Philosophy & Technology 36, article # 45.

- Charpentier, A. (2024). Insurance, Biases, Discrimination and Fairness. Springer.

- Ciatto, N., Verelst, H., Trufin, J., Denuit, M. (2023). Does autocalibration improve goodness of lift? European Actuarial Journal 13, 479-486.

- Denuit, M., Charpentier, A., Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing with machine learning. Insurance: Mathematics and Economics 101, 485-497.

- Denuit, M., Dhaene, J., Goovaerts, M.J., Kaas, R. (2005). Actuarial Theory for Dependent Risks: Measures, Orders and Models. Wiley, New York.

- Denuit, M., Sznajder, D., Trufin, J. (2019). Model selection based on Lorenz and concentration curves, Gini indices and convex order. Insurance: Mathematics and Economics 89, 128-139.

- Denuit, M., Trufin, J. (2023). Model selection with Pearson's correlation, concentration and Lorenz curves under autocalibration. European Actuarial Journal 13, 871-878.

- Denuit, M., Trufin, J. (2024). Convex and Lorenz orders under balance correction in nonlife insurance pricing: Review and new developments. Insurance: Mathematics and Economics 118, 123-128.

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. Proceedings Innovations in Theoretical Computer Science Conference (ITCS 2012), pp. 214–226.

- Fissler, T., Lorentzen, C., Mayer, M. (2022). Model comparison and calibration assessment: User guide for consistent scoring functions in machine learning and actuarial practice. arXiv preprint arXiv:2202.12780.

- Frees E. W., Huang F. (2023). The discriminating (pricing) actuary. North American Actuarial Journal 27, 2-24.

- Fröhlich, C., Williamson, R.C. (2024). Insights from insurance for fair machine learning. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 407-421.

- Hebert-Johnson, U., Kim, M., Reingold, O., Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. Proceedings of International Conference on Machine Learning, pp. 1939-1948.

- Krüger, F., Ziegel, J.F. (2021). Generic conditions for forecast dominance. Journal of Business & Economic Statistics 39, 972-983.

- Kusner, M. J., Loftus, J., Russell, C., Silva, R. (2017). Counterfactual Fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., (Eds.), Advances in Neural Information Processing Systems, vol. 30.

- Lindholm, M., Lindskog, F., Palmquist, J. (2023). Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. Scandinavian Actuarial Journal 2023, 946-973.

- Lindholm, M., Richman, R., Tsanakas A., Wüthrich, M.V. (2022). Discrimination-free insurance pricing. Astin Bulletin 52, 55-89.

- Lundborg, A. R. (2022). Modern methods for variable significance testing. PhD thesis, Apollo - University of Cambridge Repository.

- Noll, A., Salzmann, R., Wüthrich, M. (2018). Case study: French motor third-party liability claims. Available at SSRN https://ssrn.com/abstract=3164764.

- Scantamburlo, T., Baumann, J., Heitz, C. (2025). On prediction-modelers and decision-makers: Why fairness requires more than a fair prediction model. AI & Society 40, 353-369.

- Wüthrich, M. V. (2023). Model selection with Gini indices under auto-calibration. European Actuarial Journal 13, 469-477.

- Wüthrich, M. V., Ziegel, J. (2024). Isotonic recalibration under a low signal-to-noise ratio. Scandinavian Actuarial Journal 2024(3), 279-299

- Zhang, Y., Huang, L., Yang, Y., Shao, X. (2025). Testing conditional mean independence using generative neural networks. arXiv preprint arXiv:2501.17345.