AUTOCALIBRATION AND TWEEDIE-DOMINANCE FOR INSURANCE PRICING WITH MACHINE LEARNING

Denuit, M., Charpentier, A. and J. Trufin

LIDAM Discussion Paper ISBA/2021/13

UCLouvain

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

AUTOCALIBRATION AND TWEEDIE-DOMINANCE FOR INSURANCE PRICING WITH MACHINE LEARNING

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

Arthur Charpentier
Université du Québec à Montréal (UQAM)
Montreal, Quebec, Canada

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

March 4, 2021

Abstract

Boosting techniques and neural networks are particularly effective machine learning methods for insurance pricing. Often in practice, there are nevertheless endless debates about the choice of the right loss function to be used to train the machine learning model, as well as about the appropriate metric to assess the performances of competing models. Also, the sum of fitted values can depart from the observed totals to a large extent and this often confuses actuarial analysts. The lack of balance inherent to training models by minimizing deviance outside the familiar GLM with canonical link setting has been empirically documented in Wüthrich (2019, 2020) who attributes it to the early stopping rule in gradient descent methods for model fitting. The present paper aims to further study this phenomenon when learning proceeds by minimizing Tweedie deviance. It is shown that minimizing deviance involves a trade-off between the integral of weighted differences of lower partial moments and the bias measured on a specific scale. Autocalibration is then proposed as a remedy. This new method to correct for bias adds an extra local GLM step to the analysis. Theoretically, it is shown that it implements the autocalibration concept in pure premium calculation and ensures that balance also holds on a local scale, not only at portfolio level as with existing bias-correction techniques. The convex order appears to be the natural tool to compare competing models, putting a new light on the diagnostic graphs and associated metrics proposed by Denuit et al. (2019).

Keywords: Risk classification, Tweedie distribution family, Concentration curve, Bregman loss, Convex order.

1 Introduction and motivation

In the 1960s, North-American actuaries pioneered risk classification with the help of minimum bias methods, after Bailey and Simon (1960) and Bailey (1963). The central idea is that an acceptable set of premiums should reproduce the experience within sub-portfolios corresponding to each level of meaningful risk factors (like gender or age, for instance) and also the overall experience, i.e. be balanced for each level and in total (leading to the method of marginal totals, or MMT in short).

In the late 1980s, it turned out that for any GLM with canonical link and a score containing an intercept, there is an exact balance between fitted and observed aggregated responses over the whole data set and for any level of the categorical features. This formally related GLMs to minimum bias and MMT, as documented in Mildenhall (1999). This connection greatly facilitated the wide acceptance of GLMs in actuarial practice that has been particularly fast after the development of powerful computer tools.

Nowadays, actuaries resort to advanced statistical methods to be able to accurately assess policyholders' expected claim costs, or pure premium. Modern risk classification tools can deal with highly segmented problems resulting from the massive amount of information now available to insurers. Actuarial pricing models are generally calibrated so that a statistical goodness-of-fit measure is optimized on the training set whereas model tuning (like early-stopping decisions) is performed by cross-validation. Competing models are compared on a separate test or validation set. The objective function used for model training (often call loss function in machine learning) generally corresponds to deviance or log-likelihood, in a model accounting for the nature of data under consideration: typically, Poisson for claim counts, Gamma for average claim severities and compound Poisson sums with Gamma-distributed terms for claim totals, all belonging to the Tweedie family with power variance function. The metric used to assess model performances sometimes differs from the loss function used for training. Lift diagrams developed after Meyers and Cummings (2009) appear to be useful in that respect.

Actuarial risk classification remained bridged to statistical regression models and naturally followed their evolution from GLMs to GAMs, trees and random forests, gradient and statistical boosting, projection pursuit and neural networks, to name just a few. This evolution took place gradually, following the availability of computational resources in statistical software, without questioning the relevance of algorithms regarding minimum bias and interpretability of the objective function: the deviance established itself as the only objective function, even when the interpretability of the likelihood equations in terms of balance was lost. The underlying MMT caution to GLMs has been progressively forgotten and actuaries massively adopted every new tool extending the GLM machinery.

However, it turned out that tree-based boosting models and neural networks trained to minimize deviance generally end up with a significant under-estimation of total claims: the total balance is generally broken, even on the training data set. This seems to be an intrinsic result of using such an objective function even if no formal explanation has been provided so far beyond empirical findings. In this paper, we question the relevance of deviance as objective function, without the global balance constraint. It is shown that deviance may lead to dubious candidate premiums because the latter can deviate a lot from observed losses when totals are not kept unchanged. This is formally established with the help of

analytical results and demonstrated on empirical motor third-party liability insurance data.

A natural remedy consists in restoring global balance at each step of the iterative procedure used to optimize the loss function. This is easily implemented by revising the intercept (under canonical link function) or by constraining the optimization so that the observed total matches its fitted counterpart. This simple solution restores global balance but does not ensure that financial equilibrium also holds in meaningful sub-portfolios (remember that GLMs with canonical link not only imposes global balance, at portfolio level when the score comprises an intercept, but also at the level of risk classes determined by binary features).

For this reason, we propose a new strategy based on the concept of autocalibration (see, e.g., Kruger and Ziegel, 2020). This approach guarantees global balance as well as local equilibrium in the spirit of the original MMT. This simple and effective solution to the problem is implemented by adding an extra step implementing MMT within a local GLM analysis. Specifically, after the analysis has been performed with a method that does not necessarily respect marginal totals, a local constant GLM fit is achieved in order to restore the connection with MMT. Thus, as advocated by Wüthrich (2019, 2020), we also combine GLMs with advanced statistical learning tools but in a different way. Wüthrich (2019, 2020) used Neural Networks to produce new features in the last hidden layer, to be used in a GLM replacing the output layer. In this approach, Neural Networks allow actuaries to perform feature-engineering to feed the GLM score and marginal totals are respected by the use of GLMs in the last step. Wüthrich (2019, 2020) then explains how to interpret the working features generated by the last hidden layer of Neural Networks. This approach can also be related to the polishing procedure proposed by Zumel (2019) where random forests predictions are used to train a linear model in a second step. As pointed out by this author, the extra step should not be performed on the same data in order to avoid a potential source of overfitting (called nested-model bias). The local GLM approach proposed in this paper applies to any statistical learning model, not specifically to Neural Networks or random forests. It consists in using the candidate premium itself as a new feature in the last step. This ensures that the feature space reduces to the real line. By using a local constant, or intercept-only GLM, these predictors define optimal neighborhoods to perform local averaging of observed losses.

Lack of balance is sometimes imposed on purpose by the analyst, with the hope to stabilize the results. This is typically the case when the loss function comprises a penalty term (Lasso, Ridge of Elastic-Net, for instance), in additional to a goodness-of-fit measure. The local GLM step restoring balance advocated in the present paper is thus relevant in that setting, too.

The remainder of the paper is structured as follows. Section 2 provides the reader with formal definitions and notation used throughout the text. Section 3 recalls the form of loss functions that are consistent for the mean, and hence useful in pure premium determination. In Section 4, a mixture representation of Tweedie deviance is derived. Precisely, Tweedie deviance is decomposed into the sum of the integral of weighted differences of lower partial moments and the bias measured on a specific scale. This decomposition is used to understand the consequences of training the model by minimizing Tweedie deviance. Special attention is devoted to the Poisson regression case. The concept of Tweedie dominance is also introduced there, as a natural counterpart to Bregman dominance, or forecast dominance discussed in Kruger and Ziegel (2020). In Section 5, we propose a simple and powerful method to restore

balance at both global and local levels, based on the concept of autocalibration. Tweedie dominance between autocalibrated predictors reduces to the well-known convex order, or stop-loss order with equal means that has been proposed to compare predictors by Denuit et al. (2019) and Kruger and Ziegel (2020). This allows us to derive new results for the diagnostic tools and performance metrics proposed by Denuit et al. (2019). Precisely, it is shown that the concentration curve reduces to the Lorenz curve for autocalibrated predictors, so that the corresponding ABC metric is zero. A non-zero estimated ABC thus suggests that the predictor lacks of balance. This also provides a sound methodological argument for using Lorenz curves and lift diagrams to assess model performances in insurance pricing. From a practical point of view, actuaries could also replace Lorenz and lift curves with the simpler stop-loss transforms for comparing the predictors under consideration. A numerical study is provided in Section 6. The final Section 7 discusses the results and concludes.

2 Context, definition and notation

Let us now describe the notation used in this paper. We consider a response Y and a set of features $X_1, \ldots, X_p$ gathered in the vector $\boldsymbol{X} \in \mathcal{X}$ (classically, $\mathcal{X} \subset \mathbb{R}^p$). In this paper, the response is typically the number of claims reported to the insurance company by a given policyholder, the average claim severity or the total claim amount in relation with this contract. The dependence structure inside the random vector $(Y, X_1, \ldots, X_p)$ is exploited to extract the information contained in $\boldsymbol{X}$ about Y. In actuarial pricing, the aim is to evaluate the pure premium as accurately as possible. This means that the target is the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y (claim number or claim amount) given the available information $\boldsymbol{X}$. Henceforth, $\mu(\boldsymbol{X})$ is referred to as the true (pure) premium. Notice that in some applications, $\mu(\boldsymbol{X})$ only refers to one component of the pure premium. For instance, working in the frequency-severity decomposition of insurance losses, $\mu(\boldsymbol{X})$ can be either the expected number of insured events or the expected claim size, or severity.

Example 2.1. To illustrate various concepts introduced in this paper, we use simulated Poisson counts, conditional on a single covariate $X \sim \mathrm{Uniform}([0, 10])$, with mean $\mu(x) = 8 - x + 3(x-5)_+$, which is continuous and piecewise linear. Simulated data can be visualized on the left of Figure 2.1.

Example 2.2. In addition to the Poisson counts described in Example 2.1, we also use a second set of simulated Poisson counts, conditional on two independent covariates $X_1, X_2 \sim \mathrm{Uniform}([0, 10])$, with mean

$$\begin{aligned}\mu(x_1, x_2) \;=\; & 3\left(1 - \frac{x_1 - 5}{3}\right)^2 \exp\left(-\left(\frac{x_1 - 5}{3}\right)^2 - \left(\frac{x_2 - 5}{3} + 1\right)^2\right) \\ & -10\left(\frac{x_1 - 5}{15} - \left(\frac{x_1 - 5}{3}\right)^3 - \left(\frac{x_2 - 5}{3}\right)^5\right) \exp\left(-\left(\frac{x_1 - 5}{3}\right)^2 - \left(\frac{x_2 - 5}{3}\right)^2\right) \\ & -\frac{1}{3}\exp\left(-\left(\frac{x_1 - 5}{3} + 1\right)^2 - \left(\frac{x_2 - 5}{3}\right)^2\right) + 8\end{aligned}$$

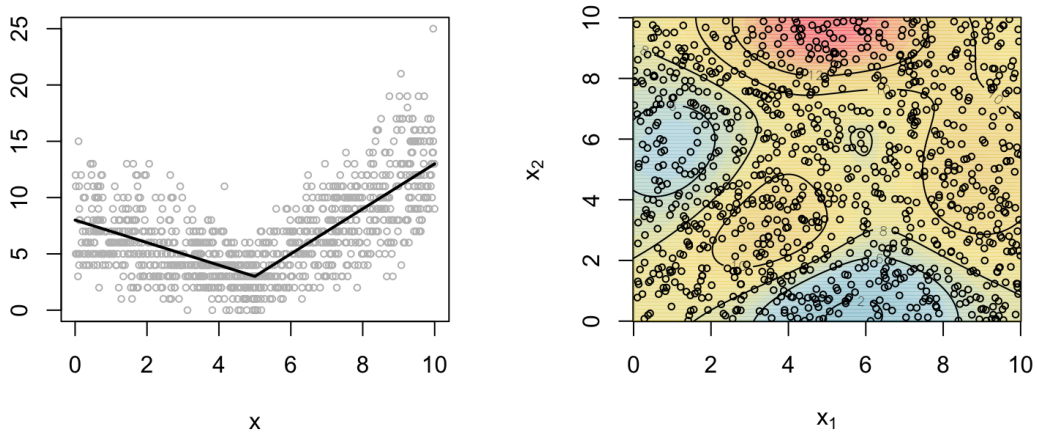

Figure 2.1: Generated data, with one and two covariates.

which is continuous and not separable. Simulated data are displayed in the right panel of Figure 2.1.

The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is unknown to the actuary, and may exhibit a complex behavior in $\boldsymbol{x}$. This is why this function is approximated by a (working, or actual) premium $\boldsymbol{x} \mapsto \pi(\boldsymbol{x})$ with a simpler structure. When the analyst is working in the frequency-severity decomposition of insurance losses, $\pi(\boldsymbol{x})$ targets the expected number of insured events or the mean claim severity, separately. Once fitted on the training data set using an appropriate learning procedure, this produces estimates $\widehat{\pi}(\boldsymbol{x})$ for $\mu(\boldsymbol{x})$, or fitted values.

The developments in this paper apply in any setting where a global balance is desirable, that is, where it is important that the sum of estimates does not deviate too much from the sum of actual observations at both the entire portfolio level and also more locally, in meaningful classes of policyholders. The reason is obvious: the sum of the pure premiums must match the claim total as accurately as possible so that the insurance company is able to indemnify all third-parties and beneficiaries in execution of the contracts, without excess nor deficit, by the very definition of pure premium (expense loadings and cost-of-capital charges are added into the calculation at a later stage, when moving to commercial premiums). This naturally translates into a global balance constraint: considering that the total claim figures are representative of next-year's experience, it is important that the sum of fitted premiums $\widehat{\pi}(\boldsymbol{x})$ match the sum of responses Y taken as proxy for the total premium income (i.e., the sum of $\mu(\boldsymbol{x})$), as closely as possible. But local equilibrium is also essential to guarantee a competitive pricing.

The merits of a given pricing tool can be assessed using the pair $\big(\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X})\big)$ so that we are back to the bivariate case even if there were thousands of features comprised in $\boldsymbol{X}$. What really matters is the correlation between $\widehat{\pi}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ but as $\mu(\boldsymbol{X})$ is unobserved the actuary can only use its noisy version Y to reveal the agreement of the true premium

$\mu(\boldsymbol{X})$ with its working counterpart $\widehat{\pi}(\boldsymbol{X})$. In insurance applications, $\widehat{\pi}(\boldsymbol{X})$ is supposed to be used as a premium so that correlation is important, but it is also essential that the sum of predictions $\widehat{\pi}(\boldsymbol{X})$ matches the sum of actual losses as closely as possible, as explained before. This is expressed by the global balance condition and its local version.

To ease the exposition, we assume that predictor $\widehat{\pi}(\boldsymbol{X})$ under consideration, as well as the conditional expectation $\mu(\boldsymbol{X})$ are continuous random variables admitting probability density functions. This is generally the case when there is at least one continuous feature contained in the available information $\boldsymbol{X}$ and the function $\widehat{\pi}$ is a continuously increasing function of a real score built from $\boldsymbol{X}$. However, this rules out predictions based on discrete features only, as well as piecewise constant predictors, e.g., a single tree. Indeed, then $\widehat{\pi}(\boldsymbol{X})$ takes only a limited number of values. As actuarial pricing is nowadays based on more sophisticated models (trees being combined into random forests, for instance), this continuity assumption does not really restrict the generality of the approach.

3 Consistent loss functions for pure premium

3.1 Bregman loss functions

Pure premiums corresponding to conditional expectations, they can be consistently estimated only if the expected loss is minimum for the mean response. This corresponds to the concept of consistent loss functions. Precisely, a loss function is consistent for the mean if no other quantity leads to a lower expected loss than the mean. Bregman loss functions are of the form

$$L(y,m) = \ell(y) - \ell(m) - \ell'(m)(y-m) \text{ for a convex function } \ell. \tag{3.1}$$

The loss function $L(\cdot,\cdot)$ in (3.1) is also referred to as q-loss function after Efron (1986). Bregman loss function measures the discrepancy between y and m on two distinct scales: on the ℓ-scale with the difference $\ell(y) - \ell(m)$ and on the original scale with the difference $y-m$. Both differences are then combined using the relative weight $\ell'(m)$. Sometimes, (3.1) is defined in terms of the concave function $-\ell$.

Savage (1971) showed, subject to weak regularity conditions, that for any loss function (3.1) and response Y,

$$\mathrm{E}[L(Y,\mathrm{E}[Y])] \leq \mathrm{E}[L(Y,m)] \text{ for any } m \text{ in the support of } Y.$$

The class of Bregman loss functions is then said to be consistent for the (conditional) mean functional. It is worth to stress that Savage (1971) refers to insurance to motivate his proposed approach, stating in Section 2 that "Money payable subject to a contingency, such as the accidential burning of a house or the outcome of a race, can be regarded as a commodity. Such commodities are explicitly dealt in by insurance companies and bookmakers". Savage (1971) then justifies the use of loss function (3.1) by the revelation of individual rate of substitution when presented a single, relatively simple, economic choice.

Bregman loss functions can take a wide variety of shapes: these loss functions can be asymmetric, with either under-predictions or over-predictions being more heavily penalized, and they can be strictly convex or have concave segments. Thus, restricting attention to loss

	Type	Name
$\xi < 0$	Continuous	-
$\xi = 0$	Continuous	Normal
$0 < \xi < 1$	Non existing	-
$\xi = 1$	Discrete	Poisson
$1 < \xi < 2$	Mixed, non-negative	Compound Poisson sum with Gamma-distributed terms
$\xi = 2$	Continuous, positive	Gamma
$2 < \xi < 3$	Continuous, positive	-
$\xi = 3$	Continuous, positive	Inverse Gaussian
$\xi > 3$	Continuous, positive	-

Table 3.1: Tweedie distributions and corresponding power parameters.

functions that generate the mean as the optimum value does not require imposing symmetry or other assumptions on the loss function. In fact, out of the infinite number of Bregman loss functions, only one is symmetric: the quadratic loss function $L(y, m) = (y - m)^2$ obtained with $\ell(m) = m^2$.

In actuarial studies targeting the pure premium, every Bregman loss function would be eligible. But a clever choice of loss function reflecting the nature of the response under consideration or its mean-variance relationship yields more accurate results. Notice that the term $\ell(y)$ is sometimes subtracted in (3.1). This does not modify the predictions and ensures that the loss function is defined over the whole range of the response. Let us also mention that there are classes of loss functions that are consistent with other indicators, such as quantiles or expectiles for instance. We refer the interested reader to Gneiting (2011) for a general overview. Loss functions are sometimes called scoring functions in the statistical literature but we do not adopt this terminology here in order to avoid any confusion with the score involved in $\pi(\boldsymbol{x})$.

3.2 Tweedie deviance

In this paper, we assume that the response obeys a probability distribution belonging to the Tweedie subclass of the Exponential Dispersion family. Precisely, this means that we assume that the logarithm of the probability mass function for a discrete response, or of the probability density function for a continuous response, is of the form $(y\theta - a(\theta))/\phi$ up to a constant term, for some known dispersion parameter ϕ and non-decreasing and convex cumulant function $a(\cdot)$ corresponding to a variance function of the form $V(\mu) = \mu^\xi$ for some power parameter ξ. For appropriate choice of ℓ in (3.1), we recover Tweedie deviance. Precisely, the function $\ell(m) = 2\big(m\theta_m - a(\theta_m)\big)$ with $a'(\theta_m) = m$ results in the deviance loss function

$$L(y, m) = 2\big(y(\theta_y - \theta_m) - a(\theta_y) + a(\theta_m)\big)$$

of the Exponential Dispersion distribution, where $a'(\theta_y) = y$.

Table 3.1 lists all Tweedie distributions. Negative values of ξ gives continuous distributions on the whole real axis. For $0 < \xi < 1$, there is no Exponential Dispersion distribution

with such variance function. Only the cases $\xi \geq 1$ are thus interesting for applications in insurance. In the remainder of this paper, we thus restrict our analysis to $\xi \geq 1$, and the corresponding deviance loss function is

$$L(y,m) = \begin{cases} 2\left(y \ln \frac{y}{m} - (y-m)\right) & \text{for } \xi = 1 \\ 2\left(-\ln \frac{y}{m} + \frac{y}{m} - 1\right) & \text{for } \xi = 2 \\ 2\left(\frac{y^{2-\xi}}{(1-\xi)(2-\xi)} - \frac{ym^{1-\xi}}{1-\xi} + \frac{m^{2-\xi}}{2-\xi}\right) & \text{for } \xi > 1 \text{ and } \xi \neq 2. \end{cases} \tag{3.2}$$

3.3 Comparison criterion

Let $\widehat{\pi}$ be the estimated mean response Y built from some training set (all formulas in this paper are meant given this training set). The respective performances of competing models can then be assessed on the basis of a validation set $\{(Y_i, \boldsymbol{X}_i),\ i = 1, 2, \ldots, n\}$, that has not been used to obtain $\widehat{\pi}$. Performances of $\widehat{\pi}$ are generally assessed with the help of out-of-(training) sample deviance, also called predictive deviance and given by

$$D_n(\xi, \widehat{\pi}) = \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, \widehat{\pi}(\boldsymbol{X}_i)\big).$$

If n is large enough then we can resort to the limiting value

$$D_n(\xi, \widehat{\pi}) \to \mathrm{E}\big[L\big(Y^{\text{new}}, \widehat{\pi}(\boldsymbol{X}^{\text{new}})\big)\big] \text{ as } n \to \infty,$$

where $(Y^{\text{new}}, \boldsymbol{X}^{\text{new}})$ is a new observation, independent of, and distributed as those $(Y_i, \boldsymbol{X}_i)$ contained in the training set. In this paper, we compare models on the basis of the large-sample version of the predictive deviance. This approach is meaningful in insurance applications where the analyst is typically in a data-rich situation.

Henceforth, we thus compare models on the basis of the predictive Tweedie deviance

$$\mathrm{E}\big[L(Y^{\text{new}}, \widehat{\pi}(\boldsymbol{X}^{\text{new}}))\big] = \begin{cases} 2\Big(\mathrm{E}\left[\widehat{\pi}(\boldsymbol{X}^{\text{new}}) - Y^{\text{new}} \ln \widehat{\pi}(\boldsymbol{X}^{\text{new}})\right] + \mathrm{E}\left[Y^{\text{new}} \ln Y^{\text{new}} - Y^{\text{new}}\right]\Big) & \text{for } \xi = 1 \\ 2\left(\mathrm{E}\left[\ln \widehat{\pi}(\boldsymbol{X}^{\text{new}}) + \frac{Y^{\text{new}}}{\widehat{\pi}(\boldsymbol{X}^{\text{new}})}\right] - \mathrm{E}\left[\ln Y^{\text{new}} - 1\right]\right) & \text{for } \xi = 2 \\ 2\left(\mathrm{E}\left[\frac{\widehat{\pi}(\boldsymbol{X}^{\text{new}})^{2-\xi}}{2-\xi} - \frac{Y^{\text{new}}\widehat{\pi}(\boldsymbol{X}^{\text{new}})^{1-\xi}}{1-\xi}\right] + \mathrm{E}\left[\frac{Y^{\text{new}\ 2-\xi}}{(1-\xi)(2-\xi)}\right]\right) & \text{for } \xi > 1 \text{ and } \xi \neq 2 \end{cases} \tag{3.3}$$

which depends on the power parameter ξ and the predictor $\widehat{\pi}$ under consideration. Notice that for any value of ξ, the last term in (3.3) does not depend on $\widehat{\pi}$ and is thus irrelevant for model comparison, so that in the remainder of the paper, we actually use

$$D(\xi, \widehat{\pi}) = \begin{cases} \mathrm{E}\left[\widehat{\pi}(\boldsymbol{X}^{\text{new}}) - Y^{\text{new}} \ln \widehat{\pi}(\boldsymbol{X}^{\text{new}})\right] & \text{for } \xi = 1 \\ \mathrm{E}\left[\ln \widehat{\pi}(\boldsymbol{X}^{\text{new}}) + \frac{Y^{\text{new}}}{\widehat{\pi}(\boldsymbol{X}^{\text{new}})}\right] & \text{for } \xi = 2 \\ \mathrm{E}\left[\frac{\widehat{\pi}(\boldsymbol{X}^{\text{new}})^{2-\xi}}{2-\xi} - \frac{Y^{\text{new}}\widehat{\pi}(\boldsymbol{X}^{\text{new}})^{1-\xi}}{1-\xi}\right] & \text{for } \xi > 1 \text{ and } \xi \neq 2 \end{cases} \tag{3.4}$$

to assess the performances of a given predictor $\widehat{\pi}$.

4 Performances assessment

4.1 Bregman dominance

Bregman dominance, also called forecast dominance is defined as dominance for every Bregman loss function (3.1). Precisely, $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if the inequality $\mathrm{E}[L(Y,\widehat{\pi}_2)] \leq \mathrm{E}[L(Y,\widehat{\pi}_1)]$ holds true for every loss function L of the form given in (3.1). We refer the interested reader to Kruger and Ziegel (2020) and the references therein for an extensive presentation of this concept. Bregman dominance is thus a particular stochastic order relation used to compare the performances of two estimators $\widehat{\pi}_1$ and $\widehat{\pi}_2$ for the conditional means. We refer the reader to Shaked and Shanthikumar (2007) for a general presentation of stochastic order relations and to Denuit et al. (2005) for applications to insurance.

It is well known that any convex function $y \mapsto \ell(y)$ can be represented as a mixture of stop-loss functions $y \mapsto (y-t)_+$. Under weak regularity conditions, Ehm et al. (2016) established that every loss function (3.1) admits a mixture representation of the form

$$L(y,m) = \int L_t(y,m)\mathrm{d}H(t)$$

where H is a non-negative measure, and

$$L_t(y,m) = (y-t)_+ - (m-t)_+ - (y-m)\mathrm{I}[m>t].$$

Thus, every scoring function consistent for the mean can be written as a weighted average over elementary or extremal loss functions L_t. As an important consequence, an estimate that is preferable in terms of each extremal L_t is preferable in terms of any Bregman loss function and is thus superior in terms of Bregman dominance.

Based on the mixture representation derived by Ehm et al. (2016), Kruger and Ziegel (2020) established in their Theorem 2.1 that $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if, and only if, the inequality

$$\mathrm{E}[(\widehat{\pi}_2 - t)_+] + \mathrm{E}\big[(Y-\widehat{\pi}_2)\mathrm{I}[\widehat{\pi}_2 > t]\big] \geq \mathrm{E}[(\widehat{\pi}_1 - t)_+] + \mathrm{E}\big[(Y-\widehat{\pi}_1)\mathrm{I}[\widehat{\pi}_1 > t]\big] \tag{4.1}$$

holds for all t. Both sides of the latter inequality is the expectation of the elementary loss function $L_t(y,m)$ once the stop-loss premiums $\mathrm{E}[(Y-t)_+]$ appearing on both sides have been canceled. It is interesting to notice that (4.1) comprises one term involving the predictor, only (its stop-loss transform with respect to the threshold t) while the joint distribution of the pair $(Y,\widehat{\pi}_2)$ only enters the second term. This echoes to the optimality conditions introduced in Denuit et al. (2019) which also combined both aspects: marginal distribution of the predictor and joint distribution with the response.

4.2 Tweedie dominance

In insurance pricing, it seems natural to restrict Bregman loss functions to Tweedie deviances (3.4). This leads to the definition of Tweedie dominance: $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in

terms of Tweedie dominance if the inequality $D(\xi, \widehat{\pi}_2) \leq D(\xi, \widehat{\pi}_1)$ holds true for every power parameter $\xi \geq 1$.

The next result provides the actuary with a sufficient condition for a model to outperform a competitor in terms of Tweedie dominance.

Proposition 4.1. *Define*

$$\psi_\xi(\pi) = \begin{cases} \ln \pi \text{ for } \xi = 2 \\ \\ \frac{\pi^{2-\xi}}{2-\xi} \text{ else.} \end{cases} \tag{4.2}$$

Then, $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Tweedie dominance if

$$\mathrm{E}\left[\psi_\xi(\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}))\right] \geq \mathrm{E}\left[\psi_\xi(\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}))\right] \text{ for all } \xi \geq 1 \tag{4.3}$$

and

$$\mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) \leq t\right]\big] \geq \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) \leq t\right]\big] \quad \text{for all } t \geq 0. \tag{4.4}$$

Proof. For $\xi = 1$, $\widehat{\pi}_2$ is superior to $\widehat{\pi}_1$ if

$$\begin{aligned} & \mathrm{E}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}})\right] - \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_1(\boldsymbol{X}^{\text{new}})\right] \geq \mathrm{E}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}})\right] - \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_2(\boldsymbol{X}^{\text{new}})\right] \\ \Leftrightarrow \quad & \mathrm{E}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}})\right] - \mathrm{E}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}})\right] + \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_2(\boldsymbol{X}^{\text{new}})\right] - \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_1(\boldsymbol{X}^{\text{new}})\right] \geq 0. \end{aligned} \tag{4.5}$$

Since the identity

$$\begin{aligned} \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}(\boldsymbol{X}^{\text{new}})\right] &= \int_0^\infty \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\ln \widehat{\pi}(\boldsymbol{X}^{\text{new}}) > t\right]\big] \mathrm{d}t - \int_{-\infty}^0 \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\ln \widehat{\pi}(\boldsymbol{X}^{\text{new}}) \leq t\right]\big] \mathrm{d}t \\ &= \int_1^\infty \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\text{new}}) > s\right]\big] \frac{1}{s} \mathrm{d}s - \int_0^1 \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\text{new}}) \leq s\right]\big] \frac{1}{s} \mathrm{d}s \end{aligned}$$

holds true for any predictor $\widehat{\pi}$, we have

$$\begin{aligned} & \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_2(\boldsymbol{X}^{\text{new}})\right] - \mathrm{E}\left[Y^{\text{new}} \ln \widehat{\pi}_1(\boldsymbol{X}^{\text{new}})\right] \\ = & \int_1^\infty \big(\mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) > s\right]\big] - \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) > s\right]\big]\big) \frac{1}{s} \mathrm{d}s \\ & - \int_0^1 \big(\mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) \leq s\right]\big] - \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) \leq s\right]\big]\big) \frac{1}{s} \mathrm{d}s \\ = & \int_1^\infty \big(\mathrm{E}\big[Y^{\text{new}} \left(1 - \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) \leq s\right]\right)\big] - \mathrm{E}\big[Y^{\text{new}} \left(1 - \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) \leq s\right]\right)\big]\big) \frac{1}{s} \mathrm{d}s \\ & + \int_0^1 \big(\mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) \leq s\right]\big] - \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) \leq s\right]\big]\big) \frac{1}{s} \mathrm{d}s \\ = & \int_0^\infty \big(\mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\text{new}}) \leq s\right]\big] - \mathrm{E}\big[Y^{\text{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\text{new}}) \leq s\right]\big]\big) \frac{1}{s} \mathrm{d}s \\ \geq & \ 0 \end{aligned}$$

by (4.4). Hence, both conditions (4.3) and (4.4) ensure inequality (4.5).

Turning to the case $\xi = 2$, $\widehat{\pi}_2$ is superior to $\widehat{\pi}_1$ if

$$\begin{aligned} & \mathrm{E}\left[\ln \widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})\right] + \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})}\right] \geq \mathrm{E}\left[\ln \widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})\right] + \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})}\right] \\ \Leftrightarrow \quad & \mathrm{E}\left[\ln \widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})\right] - \mathrm{E}\left[\ln \widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})\right] + \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})}\right] - \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})}\right] \geq 0. \end{aligned} \tag{4.6}$$

Hence, it suffices to notice that (4.4) implies $\mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})}\right] - \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})}\right] \geq 0$ since the identity

$$\begin{aligned} \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})}\right] &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\frac{1}{\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})} \geq t\right]\right] \mathrm{d}t \\ &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}}) \leq \frac{1}{t}\right]\right] \mathrm{d}t \\ &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}}) \leq s\right]\right] \frac{1}{s^2} \mathrm{d}s. \end{aligned}$$

is valid for every predictor $\widehat{\pi}$.

Finally, in the remaining cases, i.e. $\xi > 1$ and $\xi \neq 2$, $\widehat{\pi}_2$ is superior to $\widehat{\pi}_1$ if

$$\begin{aligned} & \mathrm{E}\left[\frac{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})^{2-\xi}}{2-\xi}\right] + \frac{1}{\xi-1}\mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] \geq \mathrm{E}\left[\frac{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})^{2-\xi}}{2-\xi}\right] + \frac{1}{\xi-1}\mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] \\ \Leftrightarrow \quad & \mathrm{E}\left[\frac{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})^{2-\xi}}{2-\xi}\right] - \mathrm{E}\left[\frac{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})^{2-\xi}}{2-\xi}\right] + \frac{1}{\xi-1}\left(\mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] - \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right]\right) \geq 0. \end{aligned} \tag{4.7}$$

Whatever the predictor $\widehat{\pi}$, we can write

$$\begin{aligned} \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})^{1-\xi} \geq t\right]\right] \mathrm{d}t \\ &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}}) \leq \frac{1}{t^{\frac{1}{\xi-1}}}\right]\right] \mathrm{d}t \\ &= \int_0^\infty \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}}) \leq s\right]\right] \frac{\xi-1}{s^\xi} \mathrm{d}s, \end{aligned}$$

The announced result then follows from

$$\begin{aligned} & \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] - \mathrm{E}\left[\frac{Y^{\mathrm{new}}}{\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})^{\xi-1}}\right] \\ &= \int_0^\infty \left(\mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}}) \leq s\right]\right] - \mathrm{E}\left[Y^{\mathrm{new}} \mathrm{I}\left[\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}}) \leq s\right]\right]\right) \frac{\xi-1}{s^\xi} \mathrm{d}s. \end{aligned}$$

This ends the proof. □

Notice that from the proof of Proposition 4.1, one can easily deduce a mixture representation of the predictive deviance (3.4) used for model comparison, that is,

$$D(\xi,\widehat{\pi})=\int_0^\infty\left(\frac{1}{t^{\xi-1}}\mathrm{P}[\widehat{\pi}\leq t]-\frac{1}{t^{\xi}}\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}\leq t]\big]\right)\mathrm{d}t. \tag{4.8}$$

This representation of $D(\xi,\widehat{\pi})$ enables to relate Bregman dominance to Tweedie dominance. Recall that if $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance then the expected loss for every loss function L of the form given in (3.1) is smaller for $\widehat{\pi}_2$, which is in particular the case for predictive Tweedie deviances. From

$$\begin{aligned}\mathrm{E}\big[(\widehat{\pi}_1-t)\mathrm{I}[\widehat{\pi}_1>t]\big]+\mathrm{E}\big[(Y-\widehat{\pi}_1)\mathrm{I}[\widehat{\pi}_1>t]\big] &= \mathrm{E}\big[(Y-t)\mathrm{I}[\widehat{\pi}_1>t]\big]\\ &= \mathrm{E}[Y]-t-\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_1\leq t]\big]+t\mathrm{P}[\widehat{\pi}_1\leq t],\end{aligned}$$

we can see that inequality (4.1) can be rewritten as

$$\mathrm{E}[Y]-t-\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_2\leq t]\big]+t\mathrm{P}[\widehat{\pi}_2\leq t]\geq\mathrm{E}[Y]-t-\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_1\leq t]\big]+t\mathrm{P}[\widehat{\pi}_1\leq t]$$

or equivalently as

$$t\mathrm{P}[\widehat{\pi}_2\leq t]-\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_2\leq t]\big]\geq t\mathrm{P}[\widehat{\pi}_1\leq t]-\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_1\leq t]\big]. \tag{4.9}$$

Then, dividing both sides of (4.9) by t^{ξ} leads to

$$\frac{1}{t^{\xi-1}}\mathrm{P}[\widehat{\pi}_2\leq t]-\frac{1}{t^{\xi}}\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_2\leq t]\big]\geq\frac{1}{t^{\xi-1}}\mathrm{P}[\widehat{\pi}_1\leq t]-\frac{1}{t^{\xi}}\mathrm{E}\big[Y\mathrm{I}[\widehat{\pi}_1\leq t]\big], \tag{4.10}$$

so that (4.8) yields $D(\xi,\widehat{\pi}_2)\leq D(\xi,\widehat{\pi}_1)$, as expected.

Instead of Tweedie dominance, the actuary could select a specific parameter ξ, only. This is typically the case with $\xi=1$ (Poisson regression for counts), $\xi=2$ or $\xi=3$ (Gamma or Inverse Gaussian regression for claim severities). In this case, condition (4.3) in Proposition 4.1 is imposed only for that specific valued of ξ.

The main ingredient of the proof of Proposition 4.1 is the integral of the difference of functions in (4.4) weighted by

$$\frac{\xi-1+\mathrm{I}[\xi=1]}{s^{\xi}}=\frac{\xi-1+\mathrm{I}[\xi=1]}{V(s)}$$

where $V(\cdot)$ is the variance function associated with the response distribution. Weights thus appear to be inversely proportional to the variability. Under the Tweedie variance function, $V(s)$ is a power of s so that smaller weights are assigned to the differences at larger values of the response.

Proposition 4.1 shows that improving a predictor $\widehat{\pi}$ can be achieved by

(i) increasing the overall bias measured on a modified scale induced by the auxiliary function ψ_ξ defined in (4.2). This may act against the conservation of observed totals.

(ii) increasing the dependence between $\widehat{\pi}$ and the response, so to decrease

$$\begin{aligned}\mathrm{E}\Big[Y^{\mathrm{new}}\mathrm{I}\big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\big]\Big] &= \mathrm{Cov}\Big[Y^{\mathrm{new}},\mathrm{I}\big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\big]\Big]+\mathrm{E}\Big[Y^{\mathrm{new}}\Big]\mathrm{P}\Big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\Big]\\ &= \mathrm{E}\Big[Y^{\mathrm{new}}\Big|\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\Big]\mathrm{P}\Big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\Big].\end{aligned}$$

The latter lower partial moment essentially depends on the correlation structure of the pair $\big(Y^{\mathrm{new}},\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\big)$. Notice that the quantities appearing in the decomposition above are closely related to expectation dependence as defined by Wright (1987).

In an insurance ratemaking context, the lower partial moment can be interpreted as best-profile premium income

$$\begin{aligned}\mathrm{E}\Big[Y^{\mathrm{new}}\mathrm{I}\big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\big]\Big] &= \mathrm{E}\Big[\mu(\boldsymbol{X}^{\mathrm{new}})\mathrm{I}\big[\widehat{\pi}(\boldsymbol{X}^{\mathrm{new}})\leq t\big]\Big]\\ &\approx \frac{1}{n}\sum_{i|\widehat{\pi}(\boldsymbol{X}_i)\leq t}\mu(\boldsymbol{X}_i).\end{aligned}$$

Here, $\sum_{i|\widehat{\pi}(\boldsymbol{X}_i)\leq t}\mu(\boldsymbol{X}_i)$ is the true premium income for the sub-portfolio formed by gathering all policyholders with predicted premium at most equal to t. These policyholders exhibit the best risk profiles according to the candidate premium $\widehat{\pi}$.

Of course, there is a trade-off between these two goals when computing $D(\xi,\widehat{\pi})$. If the model is very flexible then it can produce a predictor that correlates a lot with the response and the lower partial moment can be decreased to a large extent compared to models imposing a rigid form for the predictor (like GLMs). More bias can then be allowed.

Remark 4.2. The discussion in the present paper holds the training set as fixed. This is different from the discussion devoted to model stability, or bias-variance trade-off. Indeed, variance is meant as a property varying the training set, and the same for bias. Expectations and variances are thus taken with respect to varying training sets when bias-variance trade-off is discussed whereas here, the training set is kept fixed and we consider that a large enough validation set is available. We believe that this approach is more in line with insurance practice.

4.3 Application to Poisson regression

Poisson deviance is by far the most widely used one in insurance applications. It applies for instance to claim counts in property and casualty insurance, death counts in life insurance, and numbers of transitions in health insurance. We assume that we deal with a response Y obeying the Poisson distribution. A vector $\boldsymbol{X}$ of features is available to predict the mean response $\mu(\boldsymbol{X})$. To ease the presentation, we assume unit exposures.

Considering (4.2)-(4.3) with $\xi=1$, the bias is thus measured on the response scale. Condition (4.3) then favors $\widehat{\pi}_2$ if

$$\mathrm{E}[\mu(\boldsymbol{X}^{\mathrm{new}})]-\mathrm{E}[\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})]\geq\mathrm{E}[\mu(\boldsymbol{X}^{\mathrm{new}})]-\mathrm{E}[\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})]$$

or, equivalently,

$$\mathrm{E}[Y^{\mathrm{new}}]-\mathrm{E}[\widehat{\pi}_2(\boldsymbol{X}^{\mathrm{new}})]\geq\mathrm{E}[Y^{\mathrm{new}}]-\mathrm{E}[\widehat{\pi}_1(\boldsymbol{X}^{\mathrm{new}})].$$

A larger bias is thus advantageous. As a consequence, adopting Poisson deviance as loss function outside GLMs may create a total premium income gap

$$\begin{aligned} \mathrm{E}\Big[Y^{\text{new}}\Big] - \mathrm{E}\Big[\widehat{\pi}(\boldsymbol{X}^{\text{new}})\Big] &= \mathrm{E}\Big[\mu(\boldsymbol{X}^{\text{new}})\Big] - \mathrm{E}\Big[\widehat{\pi}(\boldsymbol{X}^{\text{new}})\Big] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\mu(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{\pi}(\boldsymbol{X}_i) \end{aligned}$$

where $\sum_{i=1}^{n}\mu(\boldsymbol{X}_i)$ is the true premium income for the validation set whereas $\sum_{i=1}^{n}\widehat{\pi}(\boldsymbol{X}_i)$ is the one obtained by adopting predictor $\widehat{\pi}$ for premium calculation. This gap may become quite large with highly flexible models such as Neural Networks or boosting.

5 Restoring balance at global and local scales

In order to give properties of a predictor $\widehat{\pi}$, let $F_{\widehat{\pi}}^{-1}$ be the quantile function (or Value-at-Risk) of $\widehat{\pi}$, defined as the generalized inverse of its distribution function $F_{\widehat{\pi}}$, that is,

$$F_{\widehat{\pi}}^{-1}(\alpha) = \inf\{t|F_{\widehat{\pi}}(t) \geq \alpha\} \text{ for a probability level } \alpha.$$

One can also consider quantile based region as sets of $\mathcal{X}$, associated with $\widehat{\pi}$, defined either as

$$\mathcal{X}_{\widehat{\pi},\underline{\alpha\%}} = \left\{\boldsymbol{x} \in \mathcal{X} : \widehat{\pi}(\boldsymbol{x}) \leq F_{\widehat{\pi}}^{-1}(\alpha)\right\}$$

or

$$\mathcal{X}_{\widehat{\pi},\overline{\alpha\%}} = \left\{\boldsymbol{x} \in \mathcal{X} : \widehat{\pi}(\boldsymbol{x}) > F_{\widehat{\pi}}^{-1}(\alpha)\right\}$$

Heuristically, $\mathcal{X}_{\widehat{\pi},\underline{\alpha\%}}$ is the set of insured characteristics related to the smallest risk, with proportion $\alpha\%$, while $\mathcal{X}_{\widehat{\pi},\overline{\alpha\%}}$ is related to the largest risks, with proportion $1-\alpha\%$ (where small and large risks are related to predictor $\widehat{\pi}$). Of course, those sets can also be defined for the true premium μ.

On Figure 5.1, we can visualize set $\mathcal{X}_{\mu,\underline{20\%}}$ and $\mathcal{X}_{\mu,\overline{80\%}}$ of bottom-20% and top-20% true risks, according to μ, in the univariate case described in Example 2.1 and the bivariate case described in Example 2.2. On Figure 5.2, sets $\mathcal{X}_{\widehat{\pi},\underline{20\%}}$ and $\mathcal{X}_{\widehat{\pi},\overline{80\%}}$ can be visualized, with two GAM models, with a low degree of freedom on the left, and a high degree of freedom on the right, for the univariate case described in Example 2.1. On Figure 5.3, sets $\mathcal{X}_{\widehat{\pi},\underline{20\%}}$ and $\mathcal{X}_{\widehat{\pi},\overline{80\%}}$ can be visualized (respectively on the left and on the right), with an additive GAM model on top ($\widehat{\pi}(x_1,x_2) = \widehat{s}_1(x_1)+\widehat{s}_2(x_2)$), and a bivariate spline based Poisson model below ($\widehat{\pi}(x_1,x_2) = \widehat{s}(x_1,x_2)$), for the bivariate case described in Example 2.2. Plain areas are $\mathcal{X}_{\widehat{\pi},\alpha}$ sets, while dashed ones are $\mathcal{X}_{\mu,\alpha}$ sets.

5.1 Autocalibration

Recall that a predictor $\widehat{\pi}$ is said to be autocalibrated if $\widehat{\pi}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})]$. We refer the reader to Kruger and Ziegel (2020) for a general presentation of this concept. By Jensen inequality, autocalibration thus ensures that

$$\mathrm{E}[g(\widehat{\pi}(\boldsymbol{X}))] \leq \mathrm{E}[g(Y)] \text{ for every convex function } g,$$

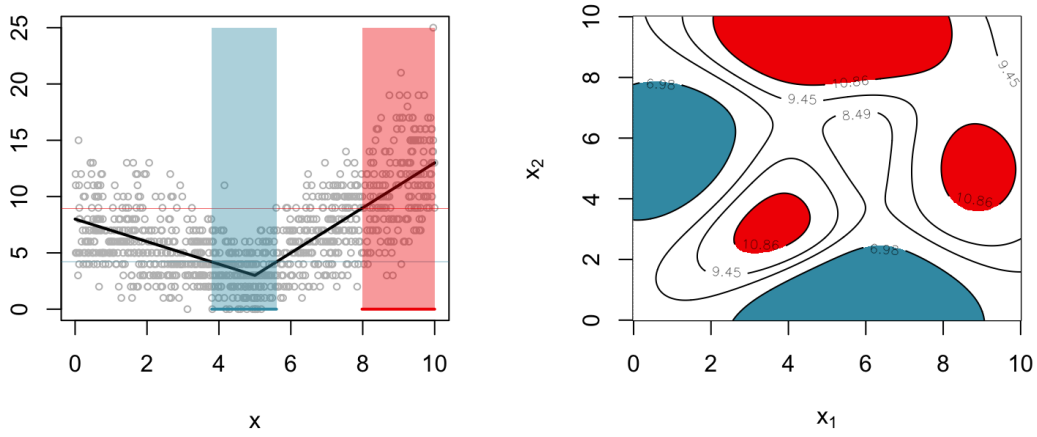

Figure 5.1: Theoretical quantiles for μ, $\mathcal{X}_{\mu,\underline{20\%}}$ and $\mathcal{X}_{\mu,\overline{80\%}}$ for the Poisson counts described in Example 2.1 in the left panel and in Example 2.2 in the right panel.

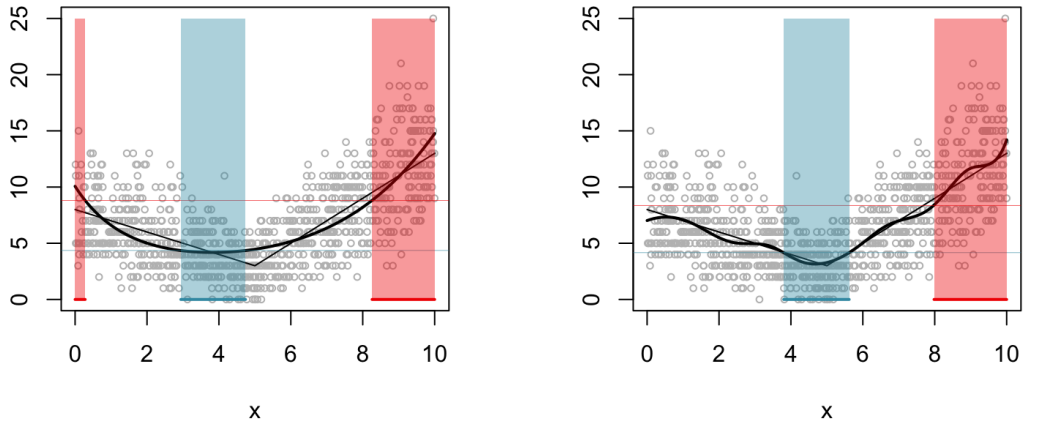

Figure 5.2: Model based quantiles for simulated Poisson counts described in Example 2.1, according to $\widehat{\pi}_{\text{low-splines}}$ (Poisson regression with splines and low degree of freedom, on the left) and $\widehat{\pi}_{\text{high-splines}}$ (Poisson regression with splines and high degree of freedom, on the right), $\mathcal{X}_{\widehat{\pi},\underline{20\%}}$ and $\mathcal{X}_{\widehat{\pi},\overline{80\%}}$.

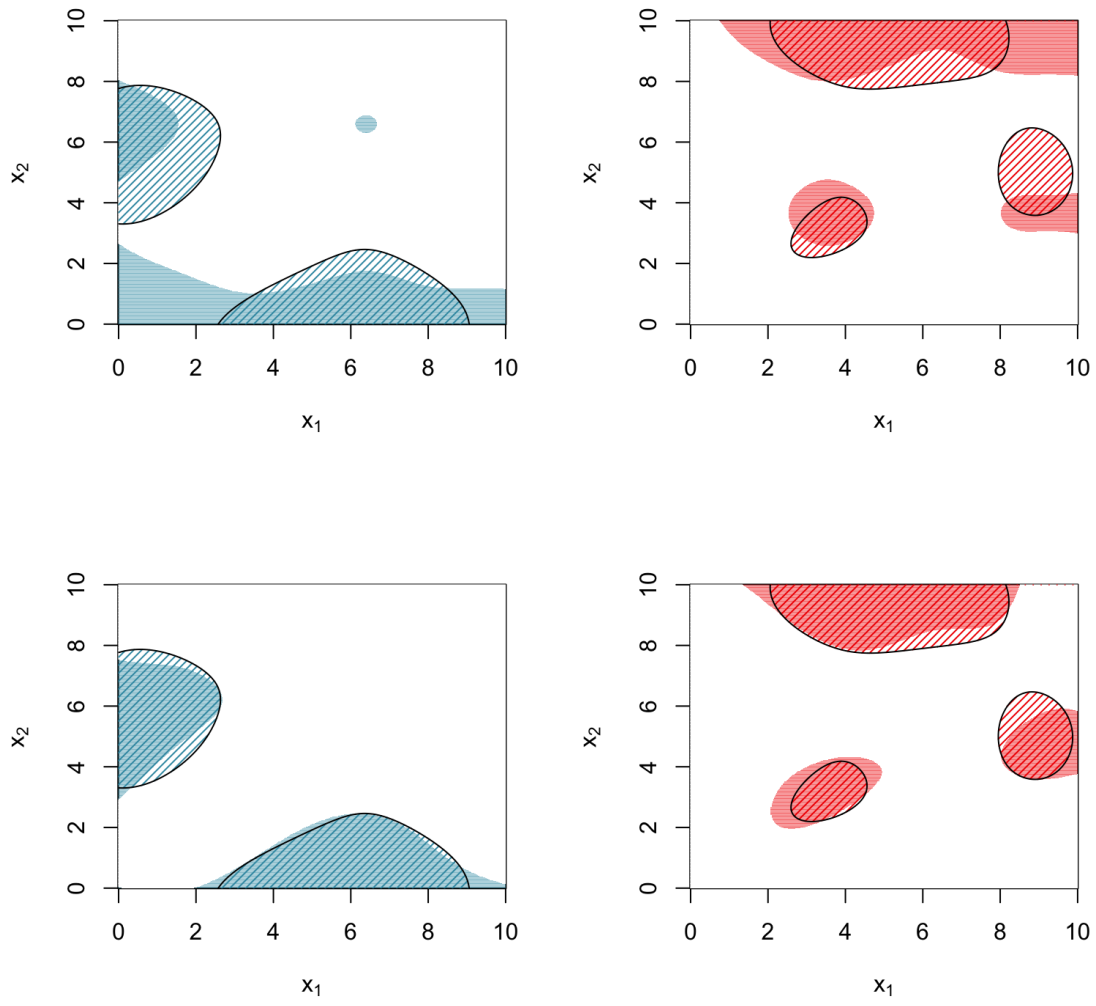

Figure 5.3: Model based quantiles for simulated Poisson counts described in Example 2.2, according to $\widehat{\pi}_{\text{add-splines}}$ (Poisson regression with additive splines, on top) and $\widehat{\pi}_{\text{biv-splines}}$ (Poisson regression with bivariate splines, below), in plain colors, $\mathcal{X}_{\widehat{\pi},\underline{20\%}}$ and $\mathcal{X}_{\widehat{\pi},\overline{80\%}}$. Dashed regions are theoretical quantiles.

or equivalently, that

$$\mathrm{E}[\widehat{\pi}(\boldsymbol{X})] = \mathrm{E}[Y] \text{ and } \mathrm{E}[(\widehat{\pi}(\boldsymbol{X}) - t)_+] \leq \mathrm{E}[(Y - t)_+] \text{ for all } t \geq 0.$$

These inequalities correspond to the convex order between $\widehat{\pi}(\boldsymbol{X})$ and Y. Thus, autocalibration implies that the predictor is less variable than the response, in the sense of the convex order.

Notice that Bregman dominance reduces to the convex order for autocalibrated predictors, as pointed out by Kruger and Ziegel (2020). This is easily deduced from (4.1) since

$$\mathrm{E}\big[(Y - \widehat{\pi})\mathrm{I}[\widehat{\pi} > t]\big] = \mathrm{E}\big[(Y - \mathrm{E}[Y|\widehat{\pi}])\mathrm{I}[\widehat{\pi} > t]\big] = 0$$

for every autocalibrated predictor.

Under mild technical requirement, a simple way to restore global balance consists in switching from $\widehat{\pi}$ to its balance-corrected version $\widehat{\pi}_{\mathrm{BC}}$ defined as

$$\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})]$$

that averages to $\mathrm{E}[Y]$, as shown in the next result.

Property 5.1. *If $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ is continuously increasing then the balance-corrected version $\widehat{\pi}_{\mathrm{BC}}$ of the candidate premium $\widehat{\pi}$ satisfies the autocalibration property.*

Proof. If $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ is continuously increasing, that is, if $\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X})$ is continuously increasing in $\widehat{\pi}(\boldsymbol{X})$, then Lemma 2.2 in Shaked et al. (2012) allows us to write

$$\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})] = \mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X})] = \widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X})$$

so that the resulting $\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X})$ is indeed autocalibrated. This ends the proof. □

In insurance applications, autocalibration induces local balance and imposes financial equilibrium not only at portfolio level but also in any sufficiently large sub-portfolio. This concept thus appears to be particularly appealing in a ratemaking context.

Heuristically, function $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ is the average value of points $\mu(\widehat{\pi}^{-1}(s))$, where $\widehat{\pi}^{-1}(s)$ simply denotes the set of $\boldsymbol{x}$ fulfilling this condition. To visualize such a function, consider the univariate case discussed in Example 2.1. Consider first the case where $\widehat{\pi}$ is obtained from a standard Poisson GLM regression with log-link, so that $\widehat{\pi}$ is strictly monotonic, and therefore convertible. On Figure 5.4, on the left $\widehat{\pi}$ is the strong curve. The horizontal line is at level s. There, $x = \widehat{\pi}^{-1}(s)$ is the axis of the two points, the one below being the one of interest, $\mu(\widehat{\pi}^{-1}(s))$ that we wish to compute. With several values s, we get a collection of points $\mu(\widehat{\pi}^{-1}(s))$ that are on the right of Figure 5.4. Here, $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s]$ is not monotonic, but $\widehat{\pi}$ is clearly underfitting. On Figure 5.5, two GAM models are considered, and $\widehat{\pi}$ is no-longuer monotonic. On top, $\widehat{\pi}$ is a convex function, and $\widehat{\pi}^{-1}(s)$ is generally a collection of two points. Function $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s]$ is then the average of those two points, since in this example, we assume that X is uniformly distributed (otherwise, weights should be proportional to the density). This gives the black strong line that we can observe on the right, with a simple GAM model, with a low degree of freedom, on top, and a GAM model

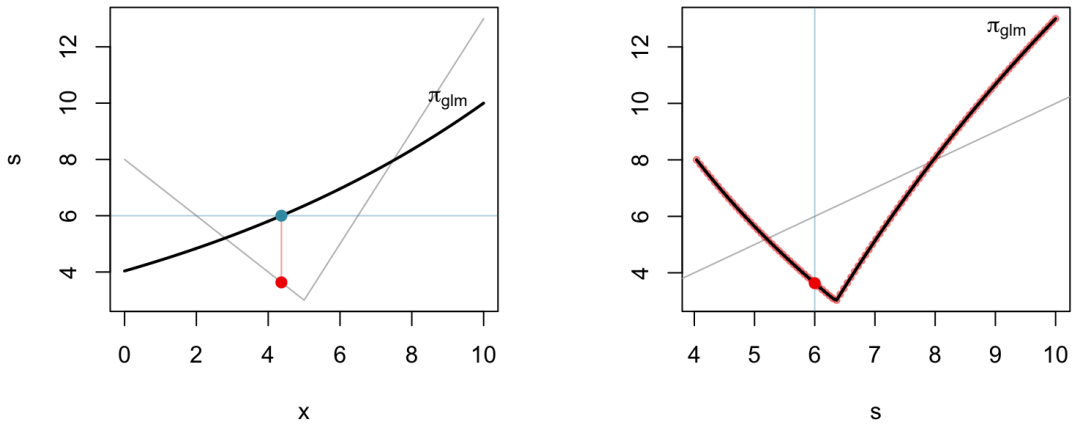

Figure 5.4: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s] = \mu(\widehat{\pi}^{-1}(s))$ for a simple Poisson regression (glm), where $\widehat{\pi}$ is monotonic, on the right. On the left, we can visualize μ and $\widehat{\pi}$, for all x's. s is on the y-axis, $(\widehat{\pi}^{-1}(s), s)$ is the blue point, $(\widehat{\pi}^{-1}(s), \mu(\widehat{\pi}^{-1}(s))$ is the red point. Points $s \mapsto \mu(\widehat{\pi}^{-1}(s))$ can be visualized on the right side.

with more degrees of freedom, below. Note that on those two examples, $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s]$ is an increasing function.

In those examples, we computed the true function $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s]$, since we used simulated data, with true premium μ, and a known distribution for X. In practice, it is not possible, but one can approximate that function using a local regression on the pairs $(\widehat{\pi}(x_i), y_i)$. This has been performed on Figure 5.6, with the GLM on the left (the true function is the thin line in the back) and the GAM model on the right. Note that here also, if $\widehat{\pi}$ does not underfit, the approximation of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s]$ is an increasing function.

In the bivariate case described in Example 2.2, the set of $\boldsymbol{X}$'s such that $\widehat{\pi}(\boldsymbol{X}) = s$ is now a collection of curves if $\widehat{\pi}$ is continuous, as we can visualize on Figure 5.7, with the same model, but two levels s, on top and below. Here $\widehat{\pi}$ is the standard additive GAM model. On the right, we can visualize the distribution of the random variable $\mu(\widehat{\pi}^{-1}(s))$, that is also Y given $\widehat{\pi}(\boldsymbol{X}) = s$. The expected value is represented by the vertical lines. Varying s, we obtain the curve on the left of Figure 5.8, which is $s \mapsto \mathrm{E}[\mu(\boldsymbol{X})|\widehat{\pi}(\boldsymbol{X}) = s]$. The two points are the two values of Figure 5.7. Note here that the function is not increasing, for low values of s, but μ was absolutely not an additive function, so again, underfitting seems to compromise the increasing property of $s \mapsto \mathrm{E}[\mu(\boldsymbol{X})|\widehat{\pi}(\boldsymbol{X}) = s]$ (that we will have with bivariate splines). The approximated version (without the knowledge of μ) is drawn on the right.

On Figure 5.9, π is now the bivariate spline based Poisson regression model. On Figure 5.10, we can visualize the theoretical function $s \mapsto \mathrm{E}[\mu(\boldsymbol{X})|\widehat{\pi}(\boldsymbol{X}) = s]$ on the left, and its approximated version on the right. Increasingness is clearly visible there.

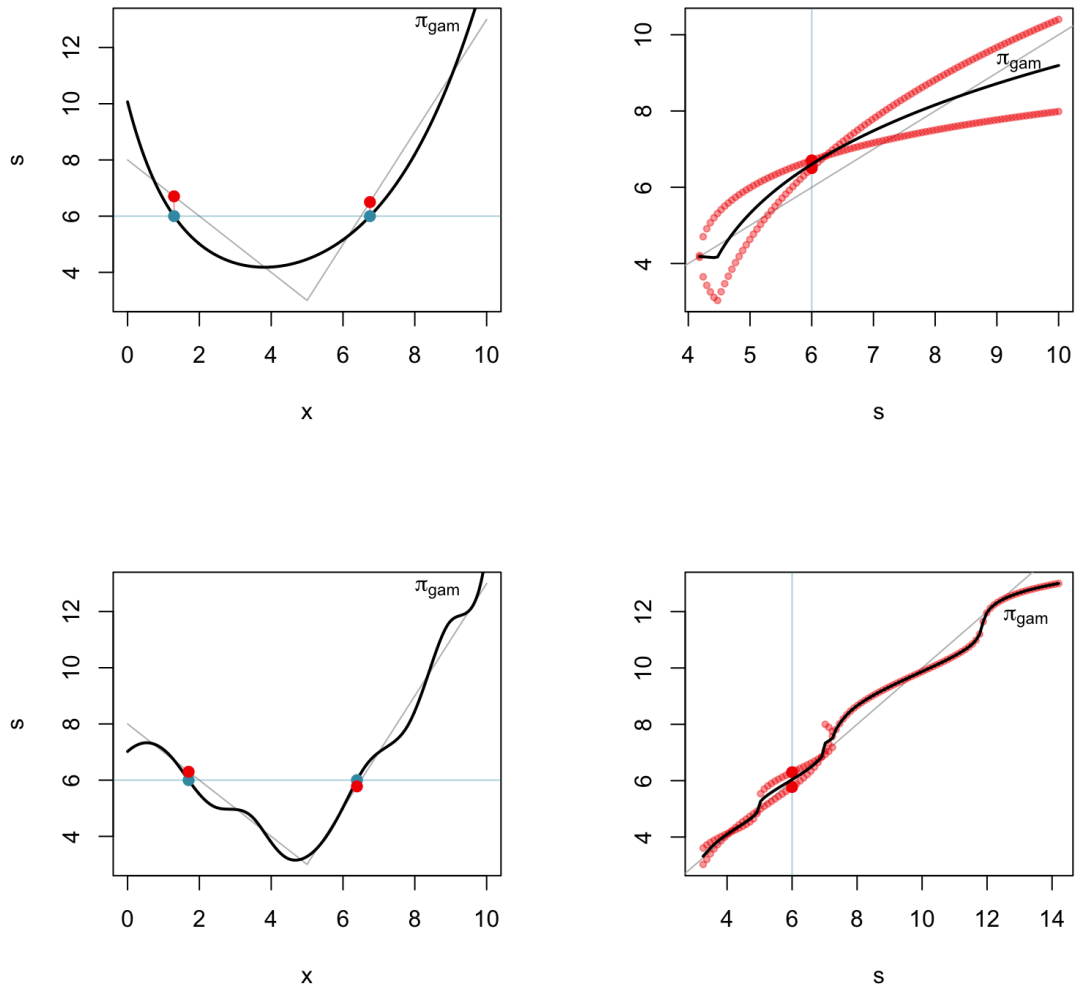

Figure 5.5: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s] = \mathrm{E}[\mu(\widehat{\pi}^{-1}(s))]$ for two Poisson regression, with splines (gam, with more degrees of freedom below) on the right. On the left, we can visualize μ and $\widehat{\pi}$, for all x's. s is on the y-axis, $(\widehat{\pi}^{-1}(s), s)$ are the blue points. Since $\widehat{\pi}$ is no longer monotonic there can be several x's such that $\widehat{\pi}(x) = s$. $(\widehat{\pi}^{-1}(s), \mu(\widehat{\pi}^{-1}(s))$ are the red points. The average of points $\mu(\widehat{\pi}^{-1}(s))$'s is function $s \mapsto \mathrm{E}[\mu(\widehat{\pi}^{-1}(s))]$ that can be visualized on the right side, with the plain black line.

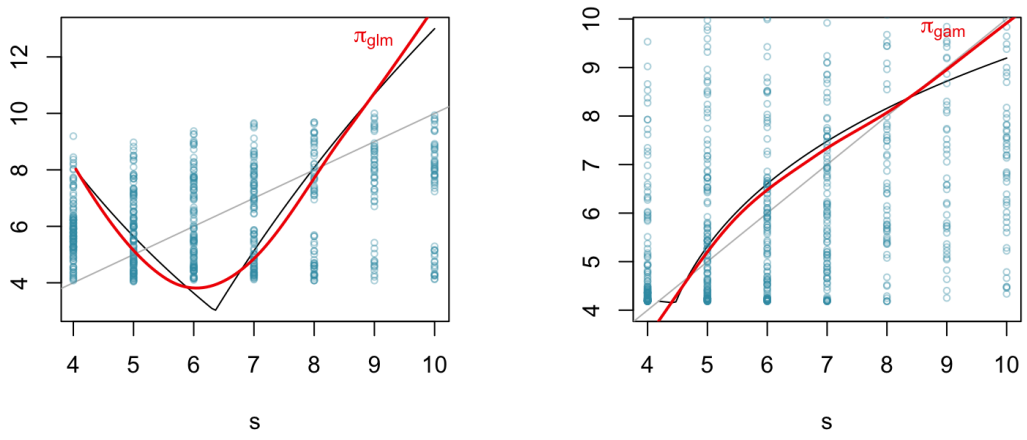

Figure 5.6: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(X) = s] = \mathrm{E}[\mu(\widehat{\pi}^{-1}(s))]$ for the Poisson GLM (on the left) and GAM (on the right), with the plain thin dark line. The approximated version of that function is drawn, from the local regression on the pairs $(\widehat{\pi}(x_i), y_i)$, in red.

5.2 Autocalibrating a given predictor

We can restore global balance, or unbias the predictor by reconciling the predicted and observed total on the training set. In this way, we recover the global balance property imposed after the seminal work by Bailey and Simon (1960). However, this does not ensure that balance holds locally. Indeed, global balance is only one of the GLM likelihood equations under canonical link, corresponding to the intercept. In order to extend the other likelihood equations imposed in the GLM setting to general machine learning procedures, we need to mimic the way local GLM proceeds for fitting, by defining meaningful neighborhoods for statistical learning.

To this end, we work under canonical link function. The only feature entering the analysis is the logarithm of the predictor, that is, $\ln \widehat{\pi}(\boldsymbol{x}_i)$ so that the local GLM is thus fitted to $(Y_i, e_i, \widehat{\pi}(\boldsymbol{x}_i))$ for some relevant exposure e_i. An intuitively acceptable solution would consist in imposing marginal constrains on local neighborhoods defined by mean of $\widehat{\pi}$. This allows for some local transfers of claims and premiums from neighboring policyholders and so implements local balance conditions in sub-portfolios corresponding to these neighborhoods. This is in essence the local GLM approach (see Loader, 1999, for a detailed account) that allows the actuary to maintain the relationship with MMT.

In order to obtain an autocalibrated version $\widehat{\pi}_{\mathrm{BC}}$ of $\widehat{\pi}$, let us consider a specific risk profile $\boldsymbol{x}$. A weight $\nu_i(\widehat{\pi}(\boldsymbol{x}))$ is assigned to each $(Y_i, e_i, \widehat{\pi}(\boldsymbol{x}_i))$, $i = 1, \ldots, n$, computed from some weight function $\nu(\cdot)$ chosen to be continuous, symmetric, peaked at 0 and defined on $[-1, 1]$. These weights depend on the relative distance of $\widehat{\pi}(\boldsymbol{x}_i)$ with respect to $\widehat{\pi}(\boldsymbol{x})$. A common choice for $\nu(\cdot)$ is the tricube weight function but several alternatives are available, including rectangular (or uniform), Gaussian or Epanechnikov, for instance. Here, $\nu_i(\widehat{\pi}(\boldsymbol{x}))$ is larger for policyholders i such that $\widehat{\pi}(\boldsymbol{x}_i)$ is close to $\widehat{\pi}(\boldsymbol{x})$ and decreases when $\widehat{\pi}(\boldsymbol{x}_i)$ gets far away

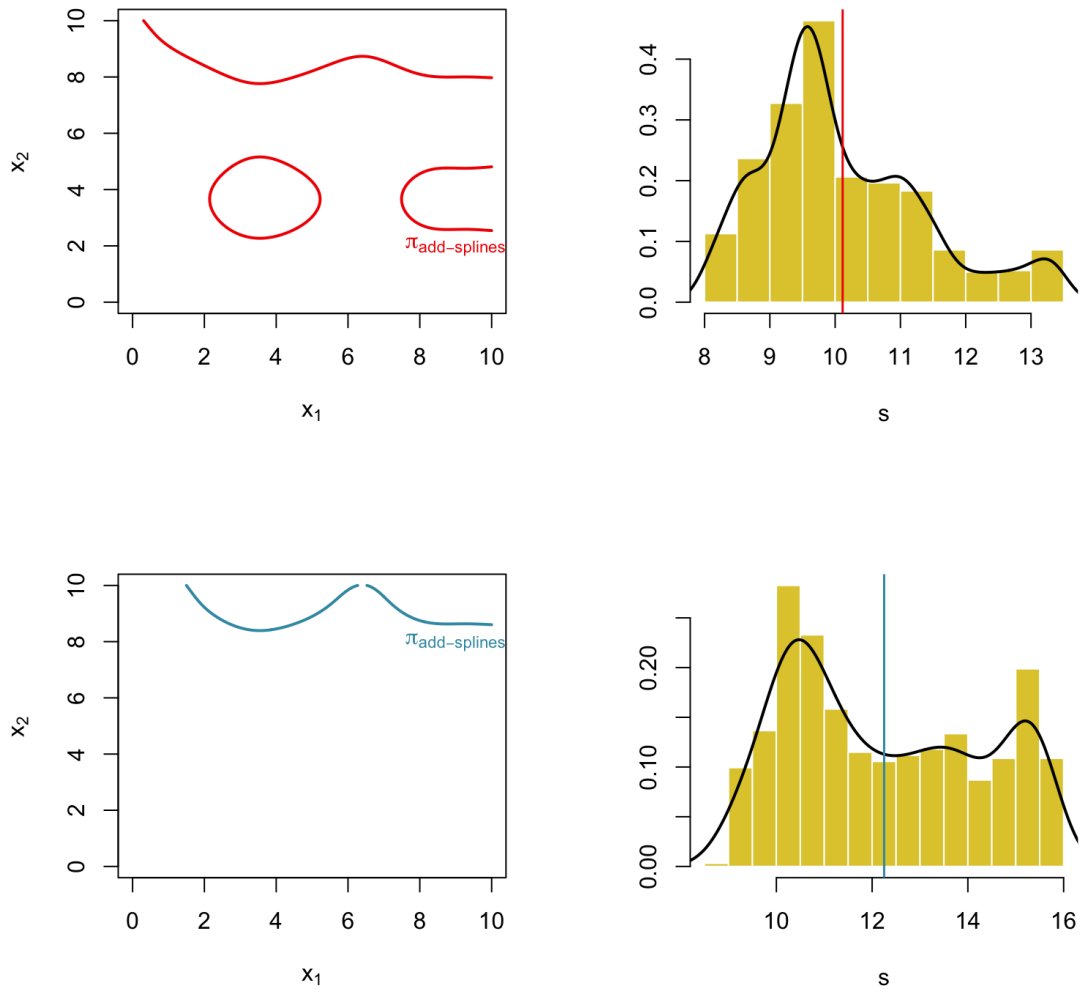

Figure 5.7: On the left, sets of points $\boldsymbol{x}$ such that $\widehat{\pi}(\boldsymbol{x}) = s$ for $s = 10$ on top and $s = 11.5$ below, when $\widehat{\pi}$ is the prediction from the classical GAM model, with additive components $s_1(x_1) + s_2(x_2)$. On the right, density of variable $\mu(\boldsymbol{X})$ conditional on $\widehat{\pi}(\boldsymbol{X}) = s$. Vertical lines are $\mathrm{E}[\mu(\boldsymbol{X})|\widehat{\pi}(\boldsymbol{X}) = s]$ for the two values of s.

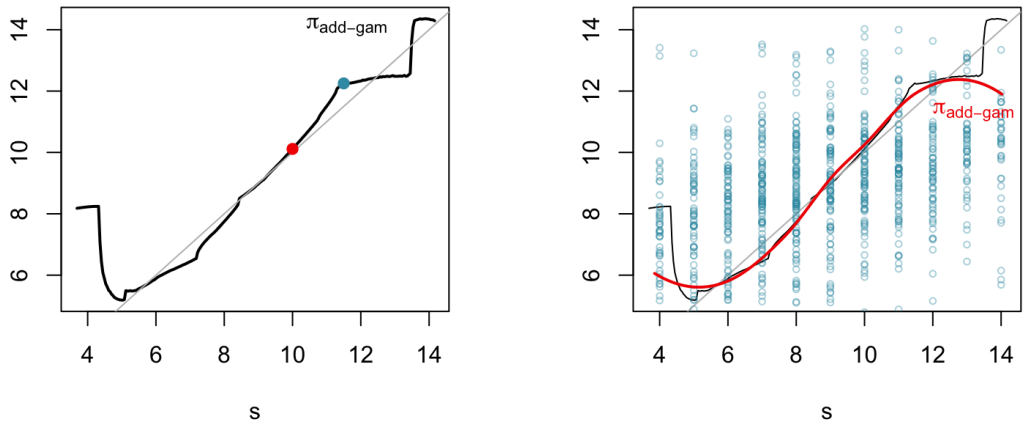

Figure 5.8: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s] = \mathrm{E}[\mu(\widehat{\pi}^{-1}(s))]$ for the bivariate spline Poisson model, with the theoretical curve on the left. Points red and blue correspond to the levels discussed on Figure 5.9. The approximated version of that function is drawn, from the local regression on the pairs $(\widehat{\pi}(\boldsymbol{x}_i), y_i)$, in red.

from $\widehat{\pi}(\boldsymbol{x})$.

The local GLM likelihood equation

$$\sum_{i=1}^{n} \nu_i(\widehat{\pi}(\boldsymbol{x}))y_i = \sum_{i=1}^{n} \nu_i(\widehat{\pi}(\boldsymbol{x}))e_i\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{x})$$

thus matches MMT constraints: smoothing is ensured by transferring part of the experience at neighboring $\widehat{\pi}$ values to obtain $\widehat{\pi}_{\mathrm{BC}}$. A local constant, or intercept-only GLM thus provides the appropriate fitting procedure when balance must be respected. Opting for a rectangular weight function complies with MMT: the weights ν_i are constant within the smoothing window and zero otherwise so that the sums reduce to observations comprised within this window and the uniform weights factor out.

Our main message here is thus that a local, intercept-only GLM with rectangular weights implements local balance, or MMT in sub-portfolios gathering policyholders with about the same predicted value. The rectangular weight function involved in the statistical procedure optimally transfers part of claim experience between neighboring policyholders. This approach implements smoothness from a statistical point of view while remaining fully transparent and understandable. Indeed, local averaging can just be seen as an application of the mutuality principle at the heart of insurance.

5.3 Link with Lorenz and concentration curves

Denuit et al. (2019) suggested to use concentration and Lorenz curves to evaluate the performances of a candidate predictor or to compare two competing alternatives. As defined

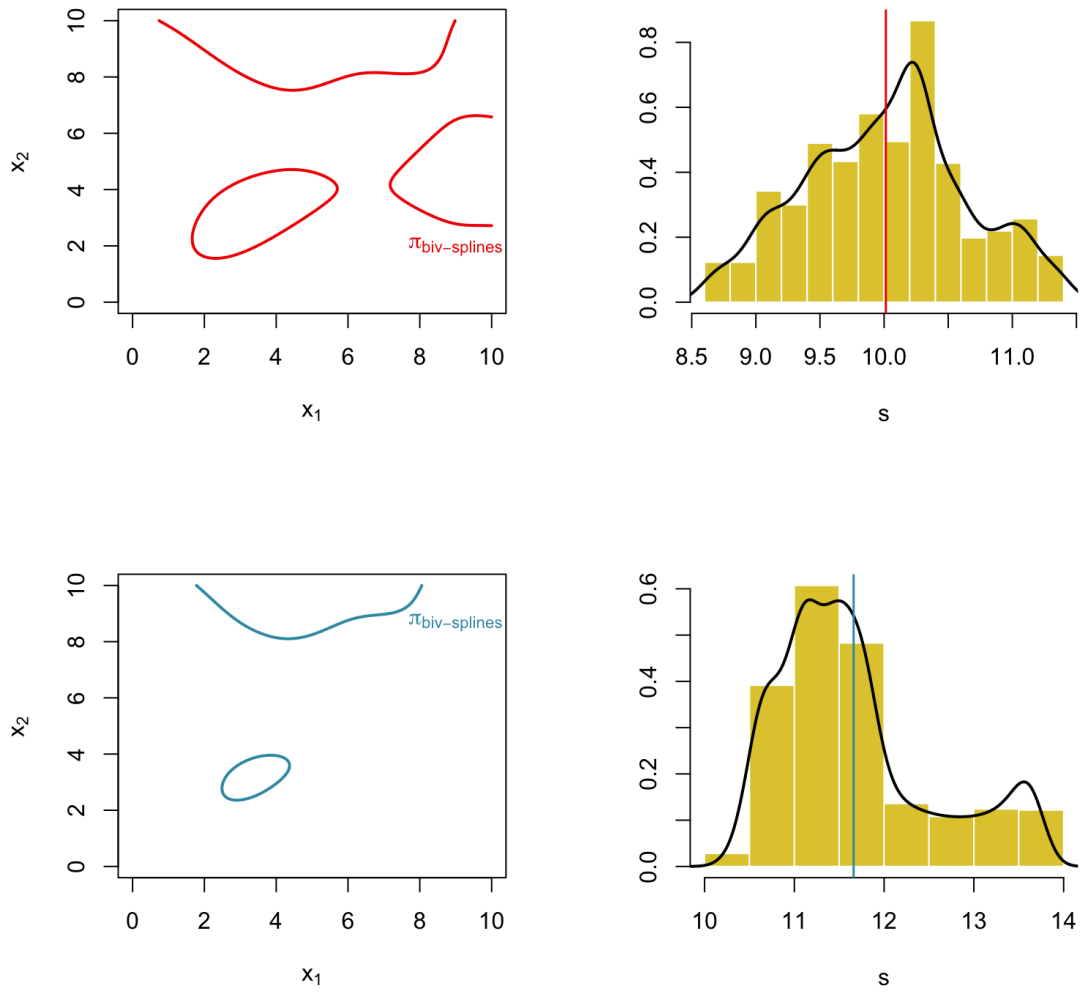

Figure 5.9: On the left, sets of points $\boldsymbol{x}$ such that $\widehat{\pi}(\boldsymbol{x}) = s$ for $s = 10$ on top and $s = 11.5$ below, when $\widehat{\pi}$ is the prediction from the Poisson regression with bivariate spline $s(x_1, x_2)$. On the right, density of variable $\mu(\boldsymbol{X})$ conditional on $\widehat{\pi}(\boldsymbol{X}) = s$. Vertical lines are $\mathrm{E}[\mu(\boldsymbol{X})|\widehat{\pi}(\boldsymbol{X}) = s]$ for the two values of s.

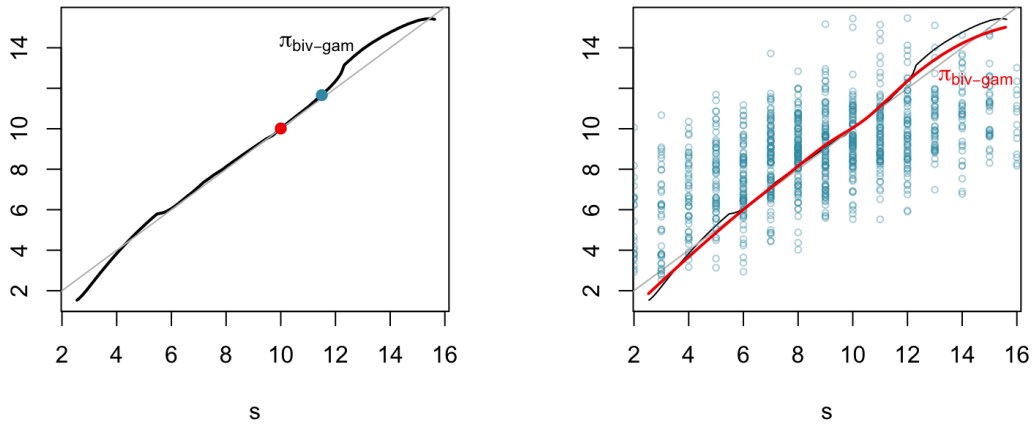

Figure 5.10: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s] = \mathrm{E}[\mu(\widehat{\pi}^{-1}(s))]$ for the bivariate spline Poisson model, with the theoretical curve on the left. Points red and blue correspond to the levels discussed on Figure 5.9. The approximated version of that function is drawn, from the local regression on the pairs $(\pi(\boldsymbol{x}_i), y_i)$, in red.

previously, $F_{\widehat{\pi}}^{-1}$ be the quantile function (or Value-at-Risk) of $\widehat{\pi}$. On Figure 5.11, we can visualize the evolution of $F_{\widehat{\pi}}$ (and actually also F_{μ}) in the univariate case discussed in Example 2.1, and the bivariate case of Example 2.2.

The concentration curve of the true premium $\mu(\boldsymbol{X})$ with respect to the working premium $\widehat{\pi}$ based on the information contained in the vector $\boldsymbol{X}$ is defined there as

$$\alpha \mapsto \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha] = \frac{\mathrm{E}\big[\mu(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\big]}{\mathrm{E}[\mu(\boldsymbol{X})]}.$$

We see that $\mathrm{CC}\big[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha\big]$ represents the percentage of the total premium income corresponding to the sub-portfolio $\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)$, i.e. to the $\alpha\%$ of contracts with the smallest premium $\widehat{\pi}$. For an exhaustive review of the properties of a concentration curve we refer the interested reader to the book by Yitzhaki and Schechtman (2013). The Lorenz curve LC associated with the predictor $\widehat{\pi}(\boldsymbol{X})$ is defined as

$$\begin{aligned} \alpha \mapsto \mathrm{LC}[\widehat{\pi}(\boldsymbol{X}); \alpha] &= \mathrm{CC}[\widehat{\pi}(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha] \\ &= \frac{\mathrm{E}\big[\widehat{\pi}(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\big]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]}. \end{aligned}$$

The next result shows that Lorenz and concentration curves coincide for autocalibrated predictors.

Property 5.2. *If $\widehat{\pi}$ is autocalibrated then*

$$\mathrm{LC}[\widehat{\pi}(\boldsymbol{X}); \alpha] = \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha]$$

for every probability level α.

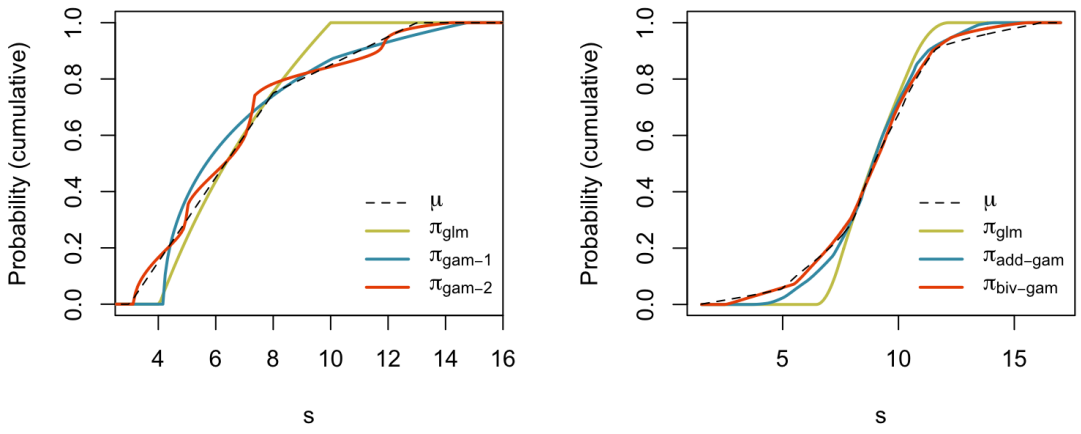

Figure 5.11: Evolution of $F_{\widehat{\pi}}$, the distribution function of $\widehat{\pi}$, i.e. $s \mapsto \mathrm{P}[\widehat{\pi}(\boldsymbol{X}) \leq s]$, in the univariate case of Example 2.1 on the left, and the bivariate case of Example 2.2 on the right.

Proof. It suffices to write

$$\begin{aligned}
\mathrm{LC}[\widehat{\pi}(\boldsymbol{X}); \alpha] &= \frac{\mathrm{E}\left[\widehat{\pi}(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})]\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}[Y\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]|\widehat{\pi}(\boldsymbol{X})]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[Y\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}[Y\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]|\boldsymbol{X}]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mathrm{E}[Y|\boldsymbol{X}]\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mu(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]} \\
&= \frac{\mathrm{E}\left[\mu(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\right]}{\mathrm{E}[\mu(\boldsymbol{X})]} \\
&= \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha],
\end{aligned}$$

which ends the proof. □

The Area Between the Curves, or ABC metric is defined after Denuit et al. (2019) as the area between the two curves CC and LC. It is interesting to notice that $\mathrm{ABC}[\widehat{\pi}(\boldsymbol{X})] = 0$ for an

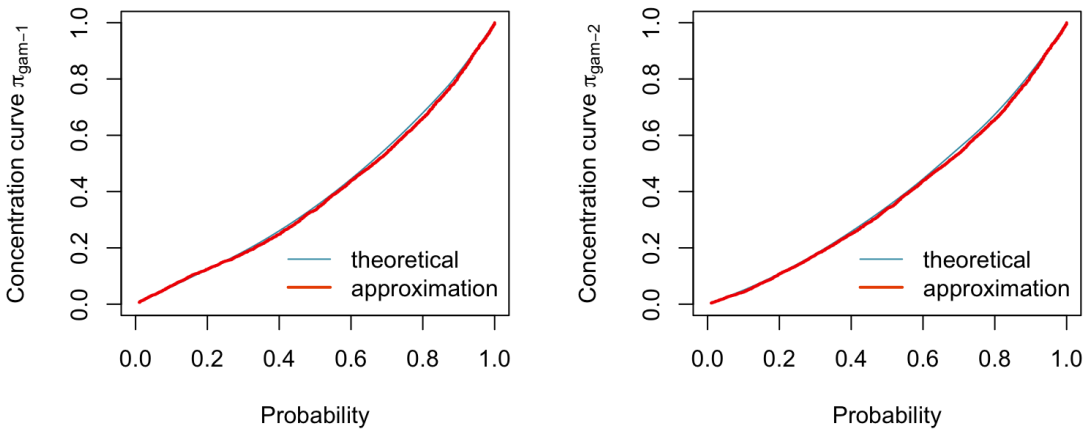

Figure 5.12: Evolution of $\mathrm{CC}(\widehat{\pi}, \mu; \alpha)$ and $\widetilde{\mathrm{CC}}(\widehat{\pi}, \mu; \alpha)$ in the univariate case.

autocalibrated predictor. A nonzero estimated ABC thus suggests that the predictor under consideration is not autocalibrated. Initially proposed as an indicator of the performance of a given predictor, it turns out that ABC is more useful for detecting the need to implement autocalibration.

Functions $\alpha \mapsto \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha]$ can be visualized on the cases of Examples 2.1 and 2.2. On Figure 5.12 we can visualize the two GAM models in the univariate case of Example 2.1. The theoretical line is the plot of CC, while the approximated one, denote $\widetilde{\mathrm{CC}}$ is

$$\alpha \mapsto \frac{1}{n\overline{y}} \sum_{i=1}^{n} y_i \mathrm{I}[\widehat{\pi}(\boldsymbol{x}_i) \leq q_\alpha)]$$

where q_α is the empirical quantile of level α of $\widehat{\pi}(\boldsymbol{x}_i)$'s. Note that computations of integrals for the theoretical lines were performed, here using quasi-Monte Carlo techniques.

On Figure 5.13 we can visualize not only the two GAM models of Example 2.1, in the univariate case, but also the GLM. On the right, we can visualize the derivative of that concentration curve. The analogous in the bivariate case of Example 2.2 can be visualized on Figure 5.14.

5.4 Bregman dominance under autocalibration

Let $\widehat{\pi}_1$ and $\widehat{\pi}_2$ be two autocalibrated predictors. Then, $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if, and only if,

$$\begin{aligned}
& \widehat{\pi}_2 \text{ dominates } \widehat{\pi}_1 \text{ in convex order} \\
\Leftrightarrow \quad & \mathrm{LC}[\widehat{\pi}_2(\boldsymbol{X}); \alpha] \leq \mathrm{LC}[\widehat{\pi}_1(\boldsymbol{X}); \alpha] \text{ for all probability levels } \alpha \\
\Leftrightarrow \quad & \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_2(\boldsymbol{X}); \alpha] \leq \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_1(\boldsymbol{X}); \alpha] \text{ for all probability levels } \alpha
\end{aligned}$$

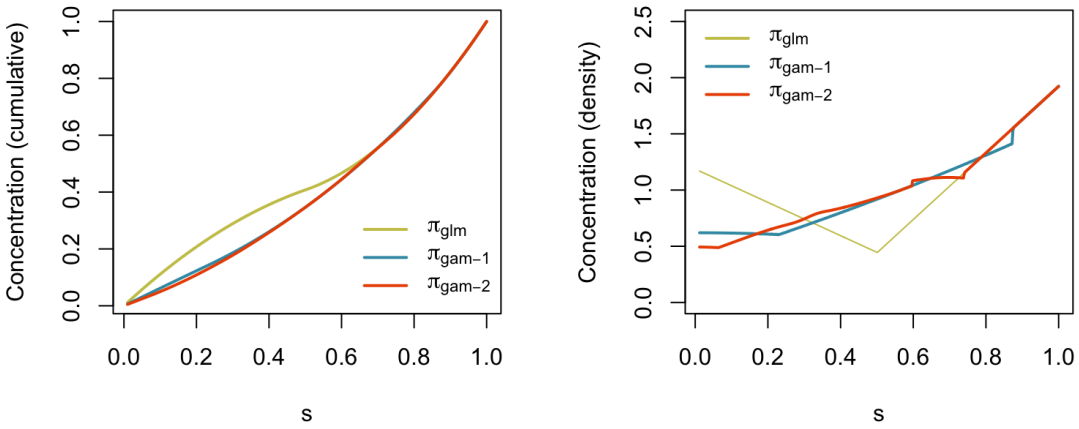

Figure 5.13: Evolution of $\mathrm{CC}(\widehat{\pi}, \mu; \alpha)$ and $\mathrm{CC}'(\widehat{\pi}, \mu; \alpha)$ is on the right, in the univariate case.

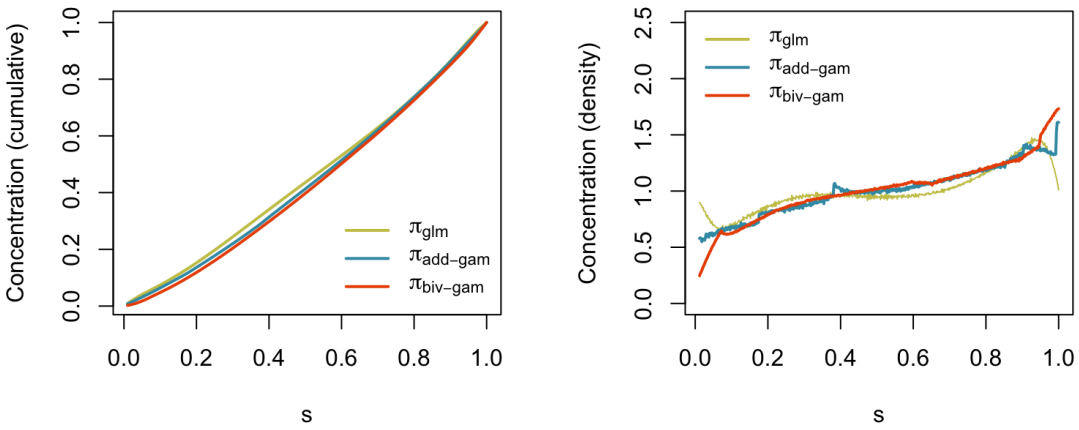

Figure 5.14: Evolution of $\mathrm{CC}(\widehat{\pi}, \mu; \alpha)$ and $\mathrm{CC}'(\widehat{\pi}, \mu; \alpha)$ is on the right, in the bivariate case.

since Lorenz and concentration curves coincide for autocalibrated predictors. It is interesting to notice that, for autocalibrated predictors, $\widehat{\pi}_2$ outperforms $\widehat{\pi}_1$ in terms of Bregman dominance if, and only if, $\widehat{\pi}_2$ is more discriminatory than $\widehat{\pi}_1$ in the sense defined in Denuit et al. (2019).

Compared to the preceding sections, thresholds are now replaced with quantiles (and thus differ with the predictor under consideration), in line with what is done in practice when lift diagrams are used. There is thus a link between concentration curves used at validation stage (Denuit et al., 2019) and the adoption of deviance as objective function for calibration. Notice also that for two predictors $\widehat{\pi}_1$ and $\widehat{\pi}_2$ such that $s \mapsto \mathrm{E}[Y|\widehat{\pi}_k(\boldsymbol{X}) = s]$ is continuously increasing for $k \in \{1, 2\}$, $\widehat{\pi}_{2,\mathrm{BC}}$ outperforms $\widehat{\pi}_{1,\mathrm{BC}}$ in terms of Bregman dominance if, and only if, $\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_{2,\mathrm{BC}}(\boldsymbol{X}); \alpha] \leq \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_{1,\mathrm{BC}}(\boldsymbol{X}); \alpha]$ for all α, or equivalently, if, and only if, $\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_2(\boldsymbol{X}); \alpha] \leq \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_1(\boldsymbol{X}); \alpha]$ for all α. Therefore, one prefers $\widehat{\pi}_2$ over $\widehat{\pi}_1$ if $\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_2(\boldsymbol{X}); \alpha] \leq \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}_1(\boldsymbol{X}); \alpha]$ for all α, so that $\widehat{\pi}_{2,\mathrm{BC}}$ outperforms $\widehat{\pi}_{1,\mathrm{BC}}$ in terms of Bregman dominance.

As noticed in Denuit et al. (2019), two predictors might well be incomparable because their respective concentration curves intersect. In such a case, we can base the comparison on the integral of the concentration curve defined as

$$\mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X})] = \int_0^1 \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{\pi}(\boldsymbol{X}); \alpha] \mathrm{d}\alpha.$$

For two predictors $\widehat{\pi}_1$ and $\widehat{\pi}_2$ such that $s \mapsto \mathrm{E}[Y|\widehat{\pi}_k(\boldsymbol{X}) = s]$ is continuously increasing for $k \in \{1, 2\}$ and $\mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{\pi}_2(\boldsymbol{X})] \leq \mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{\pi}_1(\boldsymbol{X})]$, one thus have

$$\mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{\pi}_{2,\mathrm{BC}}(\boldsymbol{X})] \leq \mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{\pi}_{1,\mathrm{BC}}(\boldsymbol{X})]$$

and

$$\mathrm{ABC}[\widehat{\pi}_{2,\mathrm{BC}}(\boldsymbol{X})] = \mathrm{ABC}[\widehat{\pi}_{1,\mathrm{BC}}(\boldsymbol{X})] = 0,$$

so that $\widehat{\pi}_{2,\mathrm{BC}}(\boldsymbol{X})$ is preferred over $\widehat{\pi}_{1,\mathrm{BC}}(\boldsymbol{X})$ according to ICC (and ABC) metrics.

6 Application on real data

In order to illustrate the techniques discussed in the previous section, consider the `freMTPL2freq` dataset, from the `CASDataset` R package of Charpentier (2014). The variable of interest is here is the annual claim frequency, and the classical model will be a Poisson one, on variable `ClaimNb`, with exposure `Exposure` and various explanatory variables (continuous and categorical).

6.1 Construction of three models

Three models will be considered : a generalized linear model $\widehat{\pi}^{\mathsf{glm}}$, a generalized additive model $\widehat{\pi}^{\mathsf{gam}}$ where continuous features are transformed nonlinearly using spline functions, a boosting algorithm $\widehat{\pi}^{\mathsf{bst}}$. For the last models, we used the `h2o` package, with `h2o.gbm`. In order to model our counting variable, we set `distribution = "poisson"` and `offset_column = "Exposure"`. For the boosting inference procedure, we used `ntrees = 30` (the impact

of the number of iterations will be discussed later on) and `nfolds = 5`[1]. From the initial dataset, with 678,013 rows, 70% are randomly chosen for our training dataset (474,609 rows) that is used to construct $\widehat{\pi}$, and the remaining 30% (203,404 rows) are used as a validation dataset, to draw the following graphs.

On Figure 6.1, we can visualise the distribution of predictions $\widehat{\pi}(\boldsymbol{x}_i)$. The average value $\overline{\pi}$ for the four models, on the training dataset, is given on the left of Table 6.1 (additional information is given, namely quantiles). As a benchmark, if we run a simple Poisson regression of the intercept only, we obtain $\widehat{\beta}_0 = -2.2911$, corresponding to a baseline prediction $\overline{\pi} = 0.10115$.

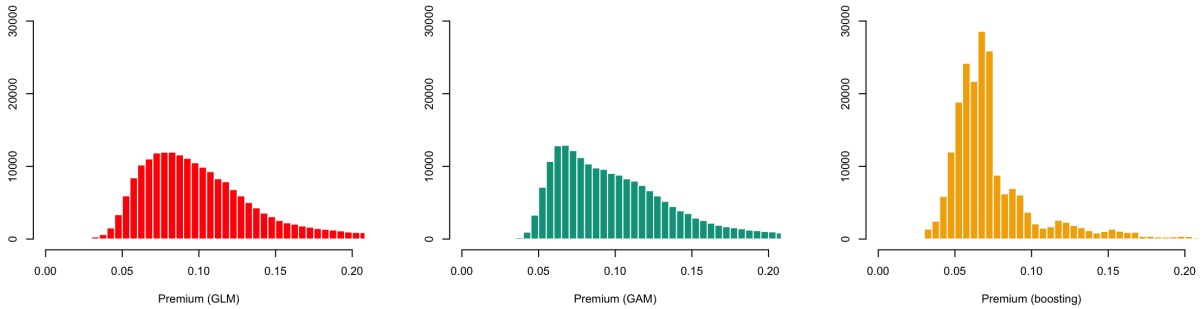

Figure 6.1: Histogram of $\{\widehat{\pi}(\boldsymbol{x}_1), \cdots, \widehat{\pi}(\boldsymbol{x}_n)\}$ on the validation dataset. In this section, ■ corresponds to the standard Poisson regression (GLM), ■ corresponds to the smooth additive regression (GAM), and ■ is the boosting model.

	$\widehat{\pi}^{\text{glm}}$	$\widehat{\pi}^{\text{gam}}$	$\widehat{\pi}^{\text{bst}}$
average $\overline{\pi}$	0.1087	0.1092	0.0820
10% quantile	0.0605	0.0598	0.0498
90% quantile	0.1682	0.1713	0.1244

	$\widehat{\pi}^{\text{glm}}_{BC}$	$\widehat{\pi}^{\text{gam}}_{BC}$	$\widehat{\pi}^{\text{bst}}_{BC}$
average $\overline{\pi}$	0.1063	0.1072	0.1037
10% quantile	0.0606	0.0545	0.0500
90% quantile	0.1706	0.1821	0.1747

Table 6.1: Summary statistics on $\{\widehat{\pi}(\boldsymbol{x}_1), \cdots, \widehat{\pi}(\boldsymbol{x}_n)\}$, on the validation dataset (assuming an exposure of 1 to provide annualized predictions), for π on the left, and the corrected version π_{BC} on the right.

On Figure 6.2, we can visualize the evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ when s varies on the range of possible premiums (here $s \in [0, 0.2]$). For the Generalized Linear Model $\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s] \sim s$, while $\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s] > s$ for the boosting model. The same information can be visualized on Figure 6.3, where the x-axis is not the premium, but the quantile version, i.e. $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}}(u)]$. The positive bias we observe on $\widehat{\pi}^{\text{bst}}$ means that this model *underestimates* the true price of the risk. As shown on Table 6.1, the bias correction will help to correct.

[1]The codes used in this section can be found on the github repository
`https://github.com/freakonometrics/autocalibration`

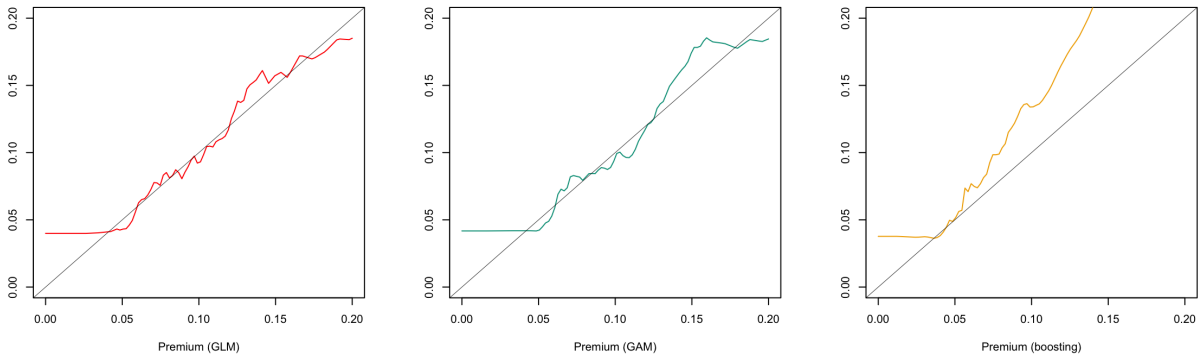

Figure 6.2: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ (the thin straight line corresponds to $\mu = \widehat{\pi}$).

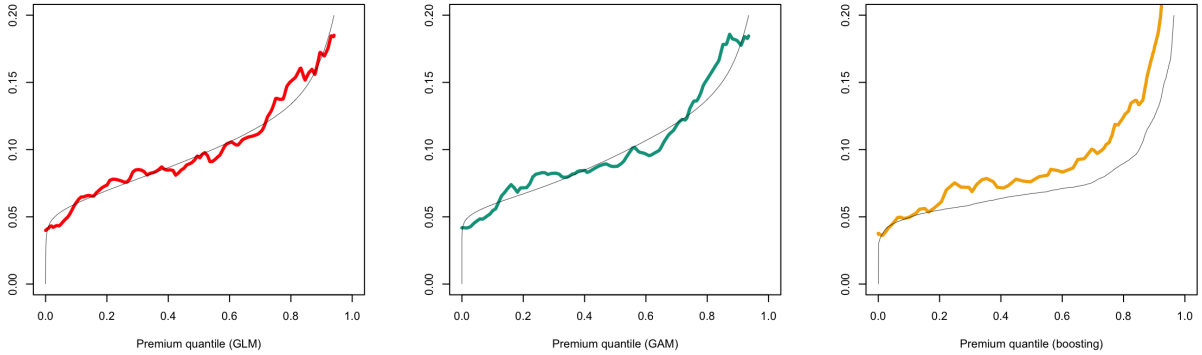

Figure 6.3: Evolution of $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]$ (the thin straight line corresponds to $\mu = \widehat{\pi}$).

On Figure 6.4, we can visualize the evolution of the difference between $\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]$ and $F_{\widehat{\pi}}^{-1}(u)$, when u varies. This is interpreted as the additive local bias on model $\widehat{\pi}$. A multiplicative version can be visualized on Figure 6.5, with $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]/F_{\widehat{\pi}}^{-1}(u)$, which is the (local) multiplier used to go from $\widehat{\pi}$ to $\widehat{\pi}_{BC}$.

6.2 From $\widehat{\pi}$ to $\widehat{\pi}_{\mathrm{BC}}$

On Figure 6.6, we can compare π and $\widehat{\pi}_{\mathrm{BC}}$: on top, we have the QQ plot of $\widehat{\pi}_{\mathrm{BC}}$ against $\widehat{\pi}$, which corresponds to the plot $u \mapsto \{F_{\widehat{\pi}}^{-1}(u), F_{\widehat{\pi}_{\mathrm{BC}}}^{-1}(u)\}$, and on the bottom, a scatterplot of $\{\widehat{\pi}(\boldsymbol{x}_i, \widehat{\pi}_{\mathrm{BC}}(\boldsymbol{x}_i\}$ (for a subset of the validation database). The distribution of $\widehat{\pi}_{\mathrm{BC}}$ can be visualized on Figure 6.7.

On Figure 6.8, we can visualize the evolution of $u \mapsto \mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = F_{\widehat{\pi}_{\mathrm{BC}}}^{-1}(u)]$, on the unit interval $[0, 1]$. On Figure 6.9, we can visualize the evolution of the difference between $\mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = F_{\widehat{\pi}_{\mathrm{BC}}}^{-1}(u)]$ and $F_{\widehat{\pi}}^{-1}(u)$, when u varies, which is the additive local bias on model $\widehat{\pi}$. A multiplicative version can be visualized on Figure 6.10, with $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}_{\mathrm{BC}}}^{-1}(u)]/F_{\widehat{\pi}_{\mathrm{BC}}}^{-1}(u)$. On the boosting version, we see clearly the impact of the correction, and

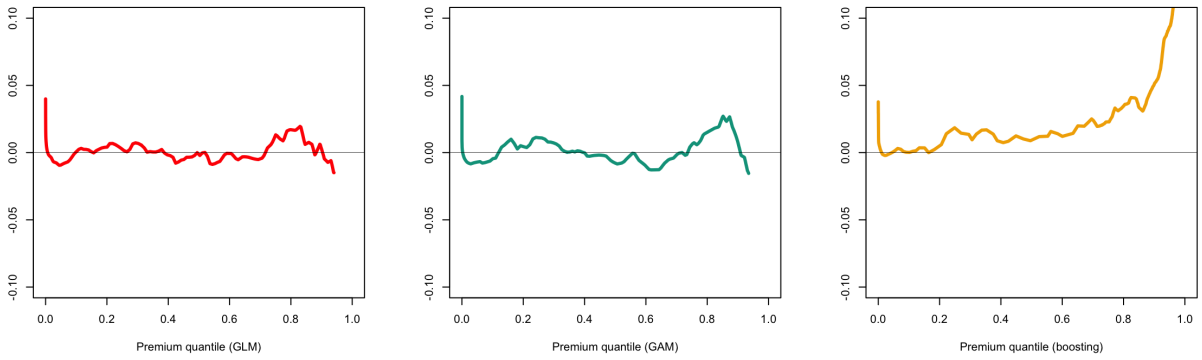

Figure 6.4: Evolution of $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)] - F_{\widehat{\pi}}^{-1}(u)$ (the thin horizontal straight line $y = 0$ corresponds to $\mu = \widehat{\pi}$).

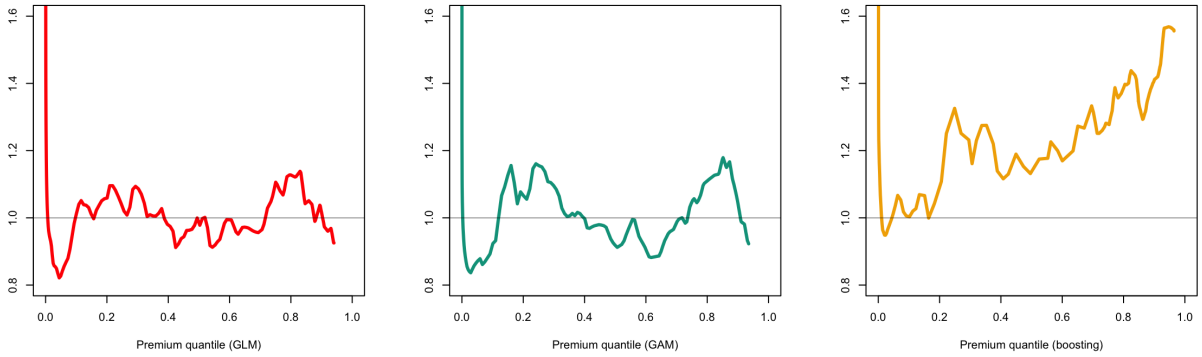

Figure 6.5: Evolution of $u \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]/F_{\widehat{\pi}}^{-1}(u)$ (the thin horizontal straight line $y = 1$ corresponds to $\mu = \widehat{\pi}$).

that $\widehat{\pi}_{\mathrm{BC}}^{\mathsf{bst}}$ is now locally unbiased.

6.3 Partial Dependence Plots

In order to compare the prices induced by this correction, we can compute partial dependence plots, i.e. for a variable of interest x_k out of $\boldsymbol{x}$, consider $x \mapsto \mathrm{E}[\widehat{\pi}(x, \boldsymbol{X}_{-k})]$ where $\widehat{\pi}(x, \boldsymbol{x}_{-k})$ corresponds to vector $\boldsymbol{x}$ where x is substituted to x_k. The empirical version is

$$\mathrm{pdp}_k(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}(x_{1,i}, \cdots, x_{k-1,i}, x, x_{k+1,i}, \cdots, x_{p,i}).$$

On Figure 6.11, we can see the partial dependence plot for the age the driver, `DrivAge`, while on Figure 6.12, we can see the partial dependence plot for the population density in the city the driver lives in, `Density`.

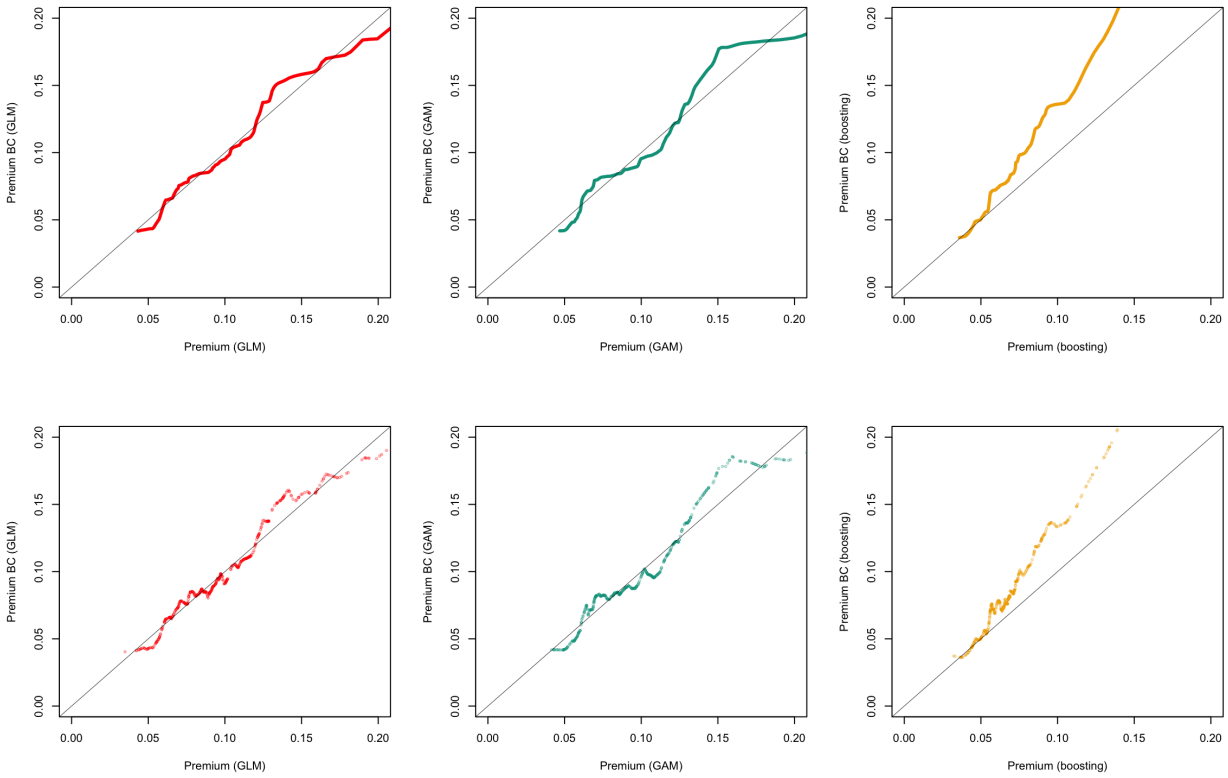

Figure 6.6: QQ plot of $\widehat{\pi}_{\text{BC}}$ against $\widehat{\pi}$ (plain line), on top, and scatterplot a subset of points $\{\widehat{\pi}(\boldsymbol{x}_i), \widehat{\pi}_{\text{BC}}(\boldsymbol{x}_i)\}$ from the validation dataset, below.

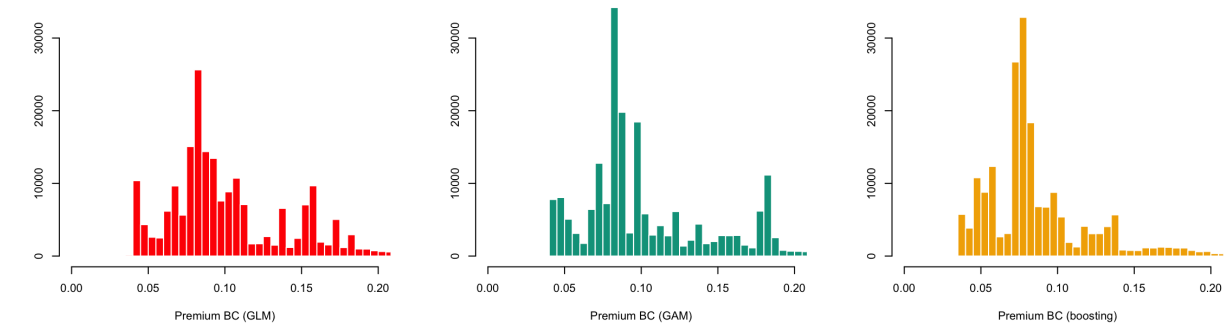

Figure 6.7: Histogram of $\{\widehat{\pi}_{\text{BC}}(\boldsymbol{x}_1), \cdots, \widehat{\pi}_{\text{BC}}(\boldsymbol{x}_n)\}$ on the validation dataset.

6.4 Smoothing parameter α and off-sample validation

The specification of bandwidth $h(s)$ is discussed in Loader (1999, Section 2.2.1). For a nearest neighbor bandwidth, compute all distances $|s - s_i|$ between the fitting point s and the data points s_i, we choose $h(x)$ to be the kth smallest distance, where $k = [n\alpha]$. In the `R` function `locfit`, when `alpha` is given as a single number, it represents a nearest neighbor

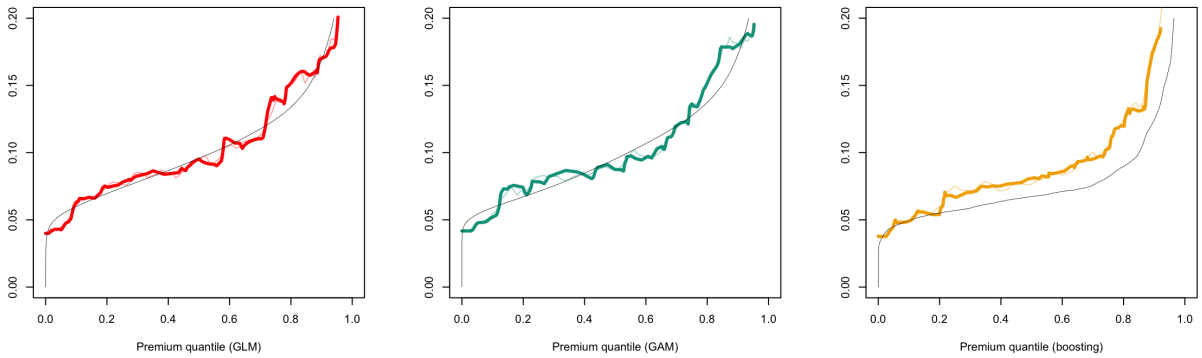

Figure 6.8: Evolution of $s \mapsto \mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}_{\mathrm{BC}}}(u)]$ (the thin colored line is the former version of $\widehat{\pi}$).

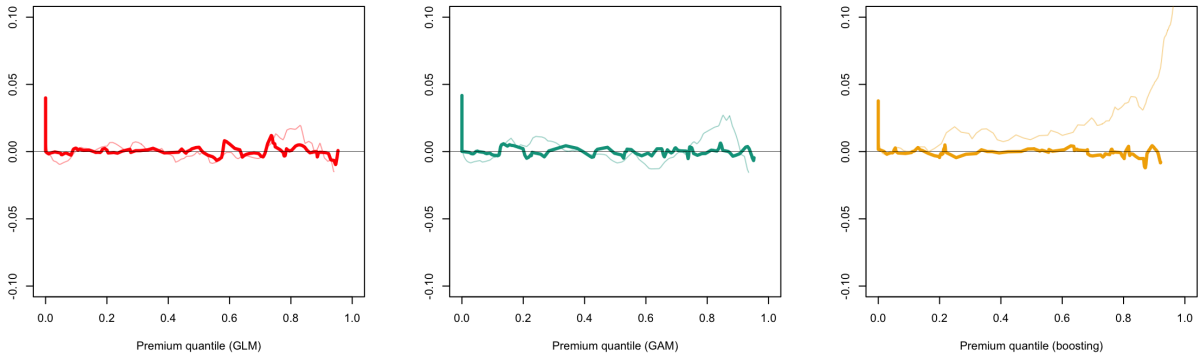

Figure 6.9: Evolution of $u \mapsto \mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}_{\mathrm{BC}}}(u)] - F^{-1}_{\widehat{\pi}_{\mathrm{BC}}}(u)$ (the thin colored line is the former version of $\widehat{\pi}$).

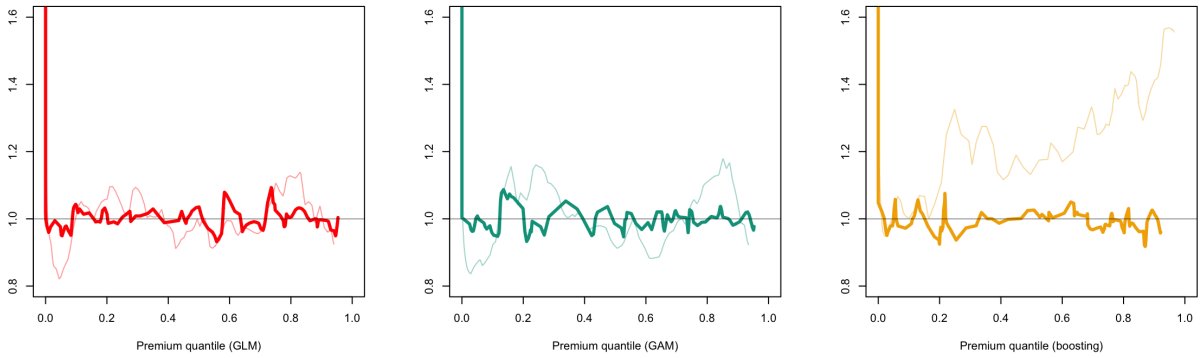

Figure 6.10: Evolution of the multiplicative correction, $u \mapsto \mathrm{E}[Y|\widehat{\pi}_{\mathrm{BC}}(\boldsymbol{X}) = F^{-1}_{\widehat{\pi}_{\mathrm{BC}}}(u)]/F^{-1}_{\widehat{\pi}_{\mathrm{BC}}}(u)$ (the thin colored line is the former version of $\widehat{\pi}$).

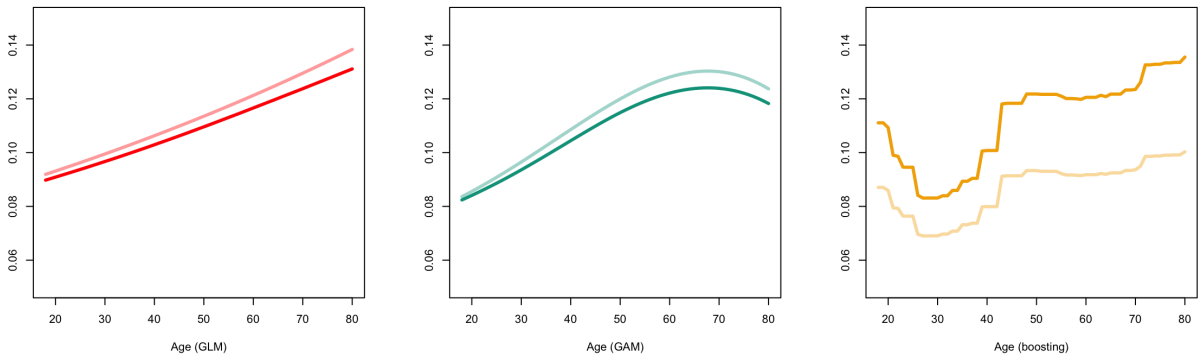

Figure 6.11: Partial Dependence Plot of the age of the driver, `DrivAge`, for the three models, with $\widehat{\pi}_{\text{BC}}$ (strong line) and $\widehat{\pi}$ (thin line).

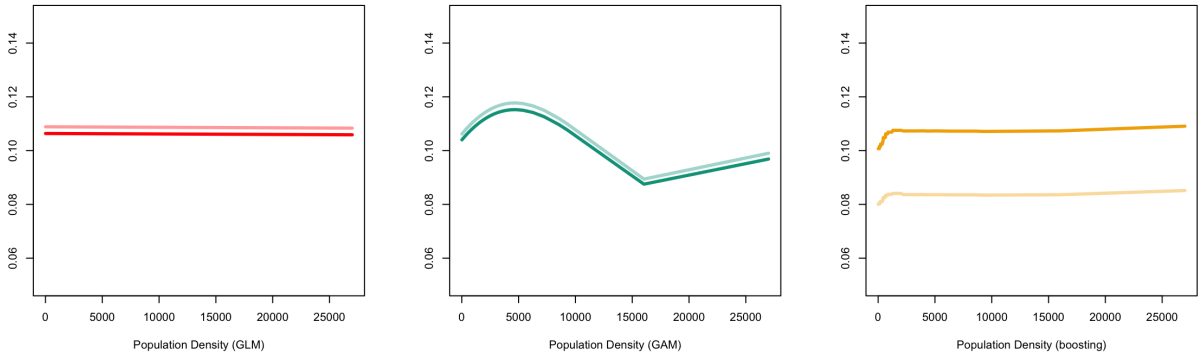

Figure 6.12: Partial Dependence Plot of the population density in the city the driver lives in, `Density`, for the three models, with $\widehat{\pi}_{\text{BC}}$ (strong line) and $\widehat{\pi}$ (thin line).

fraction (the default smoothing parameter is $\alpha = 70\%$. But a second component can be added, $\alpha = (\alpha_0, \alpha_1)$. That second component represents a constant bandwidth, and $h(s)$ will be computed as follows: as previously, $k = [n\alpha_0]$, and if $d_{(i)}$ represents the ordered statistics of $d_i = |s - s_i|$, $h(s) = \max\{d_{(k)}, \alpha_1\}$. The default value in R is `alpha=c(0.7,0)`.

As we can see on Table 6.1, the boosting algorithm was globally underestimating the average premium. Using a local regression can actually lower the bias. On Figure 6.13, we used 60% of the original dataset to train our three models (training dataset), then the local regression was performed on 20% of the dataset (smoothing dataset) and finally various quantities (bias and empirical loss) were computed on the remaining 20% (validation dataset). On the left, we can visualize the evolution of the bias $\frac{1}{n}\sum_{i=1}^{n}(\widehat{\pi}^{\mathsf{bst}}_{\text{BC}}(\boldsymbol{x}_i) - y_i)$ as a function of α. Here, smoothing has no impact on the overall bias of GLM and GAM (here the two models have a (small) positive bias on the validation dataset). Smoothing has an impact on the correction of $\widehat{\pi}^{\mathsf{bst}}$. We can observe that a small α can lead to a much smaller bias (possibly null). Note that we used $\alpha = 5\%$ for previous graphs of this section. In the middle of Figure 6.13,

we can visualize the evolution of the empirical Poisson loss $\frac{1}{n}\sum_{i=1}^{n}(\widehat{\pi}_{\text{BC}}^{\text{bst}}(\boldsymbol{x}_i) - y_i \ln \widehat{\pi}_{\text{BC}}^{\text{bst}}(\boldsymbol{x}_i))$. Note that again, a small α leads to a smaller loss. It is also interesting to see that the similar performances are obtained in case of overfitting. Thus, autocalibration also corrects for overfitting by locally averaging the initial predictors.

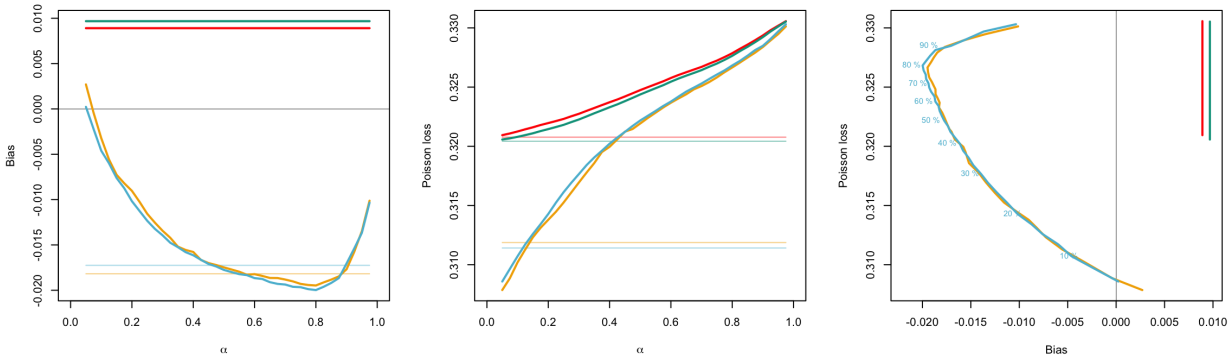

Figure 6.13: Evolution of the bias as a function a α, from 2.5% to 97.5%, on the left, of the Poisson loss in the center, and joint scatterplot of bias and empirical loss on the right. Horizontal light lines correspond to original estimates $\widehat{\pi}$ while strong curves are $\widehat{\pi}_{\text{BC}}$ for various smoothing parameter α. As previously ■ corresponds to the standard Poisson regression (GLM), ■ corresponds to the smooth additive regression (GAM), while ■ is the boosting model with 30 trees (as previously) and ■ is the boosting model with 1000 trees (overfitting).

To discuss further the impact of α, on Figures 6.14 to 6.16, we have the impact of various values for α (considered as a single value), with a 2% nearest neighbor fraction on Figure 6.14, 20% on Figure 6.15, and 80% on Figure 6.16. Note that $\alpha = 5\%$ was used in the previous section. The various graphs are those described in the previous section, but only to illustrate the construction of the construction of $\widehat{\pi}_{\text{BC}}^{\text{gbm}}$ from $\widehat{\pi}^{\text{gbm}}$.

On the top of Figures 6.14 to 6.16, we can visualize on the left $u \mapsto \text{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\pi}^{-1}(u)]$. Then on the right, we can visualize the evolution of the difference between $\text{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]$ and $F_{\widehat{\pi}}^{-1}(u)$, when u varies, that is interpreted as the additive local bias on model $\widehat{\pi}$, and the multiplicative version is on the right, with $u \mapsto \text{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]/F_{\widehat{\pi}}^{-1}(u)$, which is the (local) multiplier used to go from $\widehat{\pi}$ to $\widehat{\pi}_{\text{BC}}$. Then below, we can compare $\widehat{\pi}$ and $\widehat{\pi}_{\text{BC}}$: with the QQ plot of $\widehat{\pi}_{\text{BC}}$ against π on the left – i.e. the plot $u \mapsto \{F_{\widehat{\pi}}^{-1}(u), F_{\widehat{\pi}_{\text{BC}}}^{-1}(u)\}$ – and in the center, a scatterplot of the pairs $(\widehat{\pi}(\boldsymbol{x}_i), \widehat{\pi}_{\text{BC}}(\boldsymbol{x}_i))$. The empirical distribution of $\widehat{\pi}_{\text{BC}}(\boldsymbol{x}_i)$ is then given on the right, with an histogram. Then, on the third row, we can visualize $s \mapsto \text{E}[Y|\widehat{\pi}_{\text{BC}}(\boldsymbol{X}) = s]$ on the left and $u \mapsto \text{E}[Y|\widehat{\pi}_{\text{BC}}(\boldsymbol{X}) = F_{\widehat{\pi}_{\text{BC}}}^{-1}(u)]$ in the center (as previously, the thin line in the background is the equivalent with the original $\widehat{\pi}$). Then on the right, we can visualize the evolution of the difference between $\text{E}[Y|\widehat{\pi}(\boldsymbol{X}) = F_{\widehat{\pi}}^{-1}(u)]$ and $F_{\widehat{\pi}}^{-1}(u)$, the additive bias, while the multiplicative bias is on the left of the last row, on the bottom. If the correction worked well, we should have (almost) horizontal line for those two graphs, $y = 0$ for the additive bias, and $y = 1$ for the multiplicative one. And

finally, we can visualize partial dependence plots, with the partial dependence plot for the age the driver, `DrivAge`, in the center of that last row, and the partial dependence plot for the population density in the city the driver lives in, `Density`, on the right.

6.5 Comparing models

As we have seen when moving from $\widehat{\pi}$ to $\widehat{\pi}_{\text{BC}}$ the distribution of premiums changed substantially (see Figure 6.1 and 6.7) and the transformation was not (perfectly) monotonic (see Figure 6.6). Nevertheless, the rank correlation between $\widehat{\pi}$ and $\widehat{\pi}_{\text{BC}}$ is rather high (0.99 for the three models).

Finally, on Figure 6.18 we can visualize the concentration curves of the three models, against $\widehat{\pi}_{\text{BC}}^{\mathsf{gam}}$.

7 Discussion

The main message of this paper is as follows. Advanced learning models are able to produce scores that better correlate with the response, as well as with the true premium compared to classical GLMs. This comes from the additional freedom obtained by letting scores to depend in a flexible way of available features, not only linearly. But breaking the overall balance is the price to pay for this higher correlation. Because no constraint on the replication of the observed total, or global balance is imposed, machine learning tools are also able to substantially increase overall bias.

To prevent this to occur, the balance-corrected version of any predictor can be obtained by local GLM, recognizing the nature of the response Y: local Poisson GLM for claim counts, local Gamma GLM for average claim severities and compound Poisson sums with Gamma-distributed terms for claim totals. The canonical link function is adopted so that maximum-likelihood estimates replicate observed totals. An intercept-only GLM is fitted locally, with rectangular weight function, on a set of observations close to the one of interest with proximity assessed with the help of the predictor to be corrected.

In this approach, the supervised learning model is used to produce a real-valued signal $\widehat{\pi}(\boldsymbol{X})$, reducing the high-dimensional feature space to the real line. In fact, the score of the model is enough to compute $\widehat{\pi}_{\text{BC}}$ so that the learning model is essentially used to assess proximity among individuals, before performing local averaging. In that respect, the proposed approach shares some similarities with k-NN, or k nearest-neighbors except that here, proximity is assessed with the help of the real-valued $\widehat{\pi}$ produced by the learning model. And the proposed extra autocalibration step also counteracts possible overfitting by locally averaging the initial predictor.

The relationship established with consistent loss function and Bregman dominance also appears to be particularly instructive. EU General Data Protection Regulation (GDPR) gives consumers the right to ask for an explanation of an algorithmic decision concerning them. In an insurance context, this means that any provider should be able to explain in an understandable way the reasons for not granting coverage or why the premium charged is so expensive. Model transparency has been studied so far in terms of features interpretability, with the help of importance measures or partial dependence plots mainly. However, the role

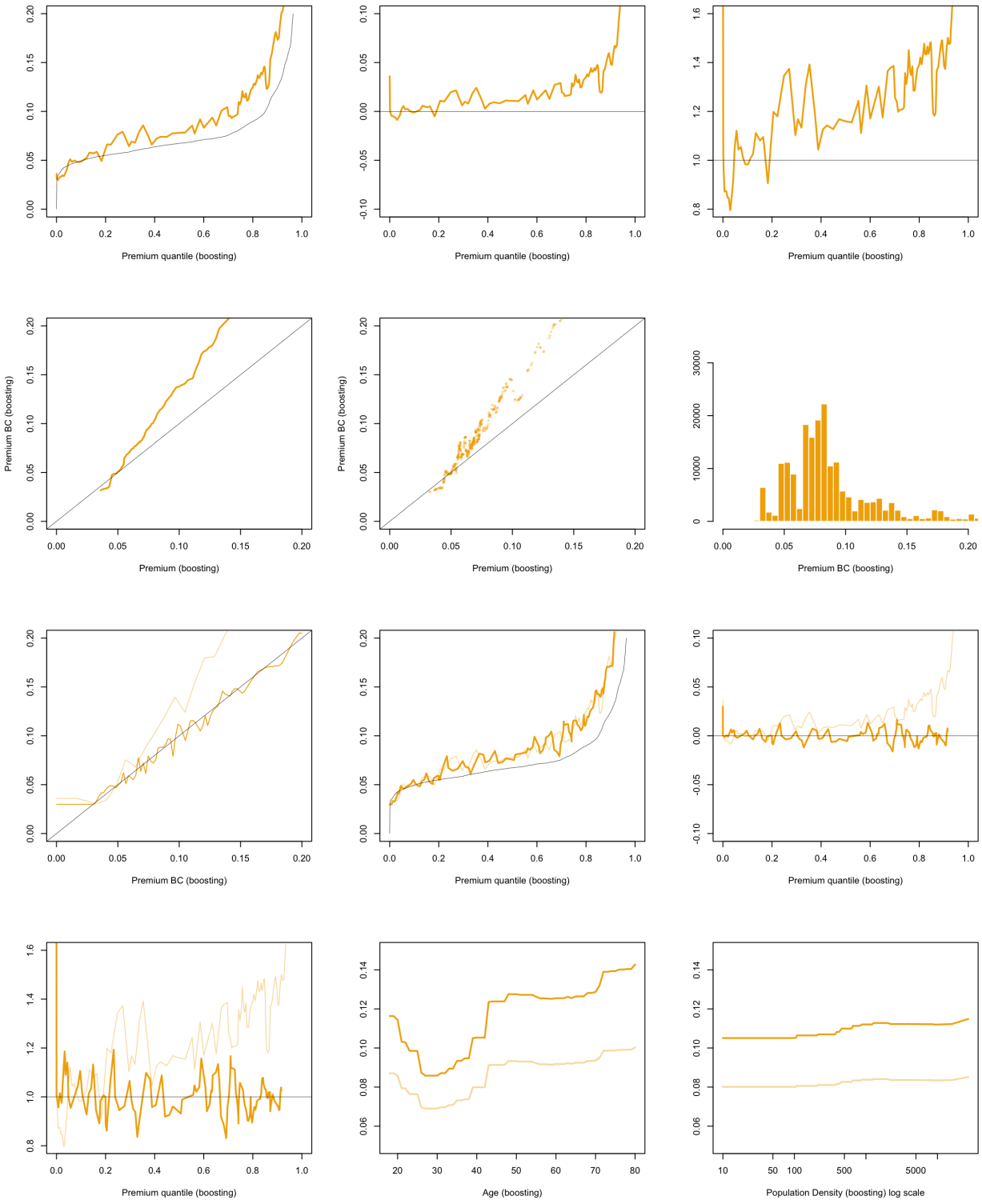

Figure 6.14: Various graphs of $\widehat{\pi}^{\mathsf{bst}}$ and $\widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$ when $\alpha = 2\%$ for $\widehat{\pi}^{\mathsf{gbm}}$

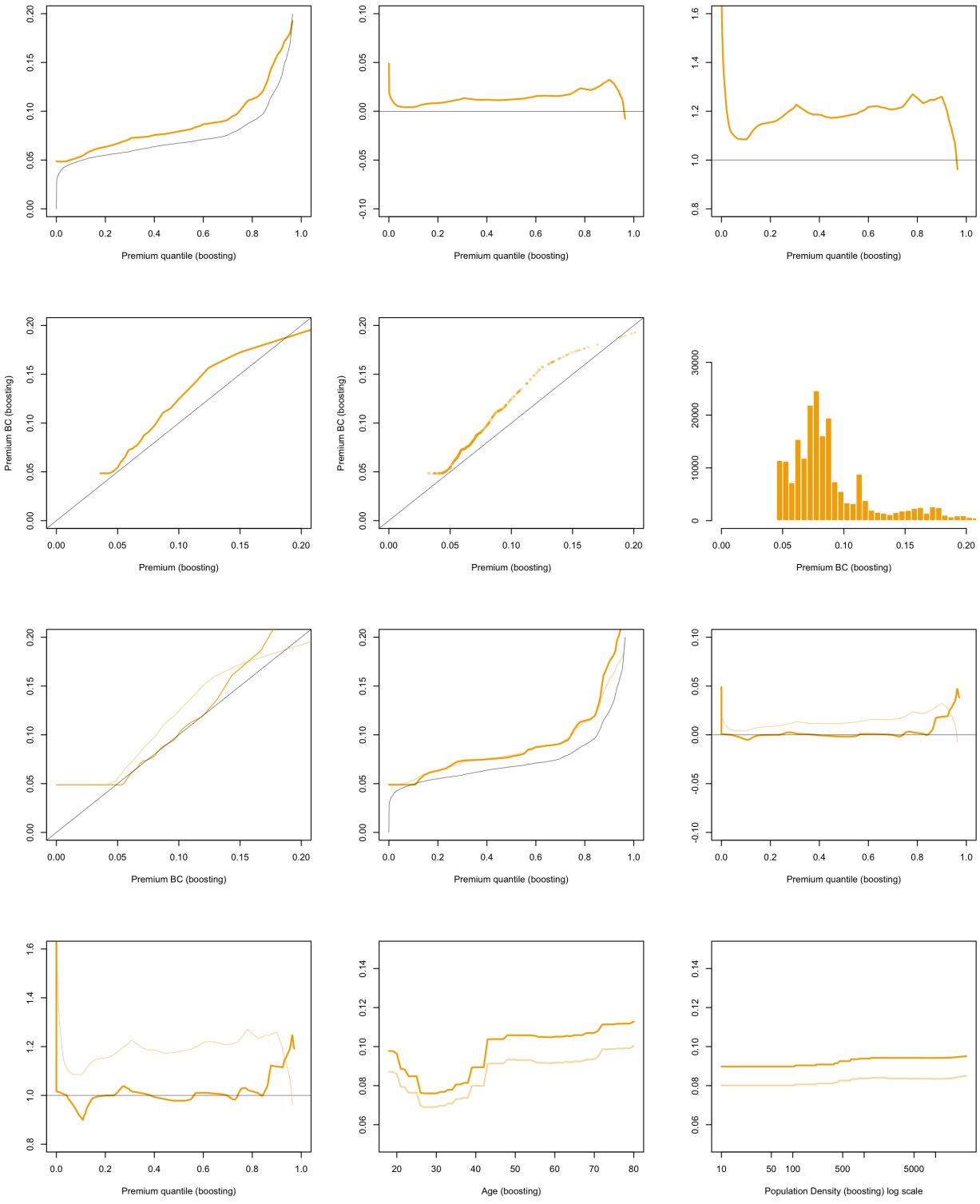

Figure 6.15: Various graphs of $\widehat{\pi}^{\mathsf{bst}}$ and $\widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$ when $\alpha = 20\%$ for $\widehat{\pi}^{\mathsf{gbm}}$

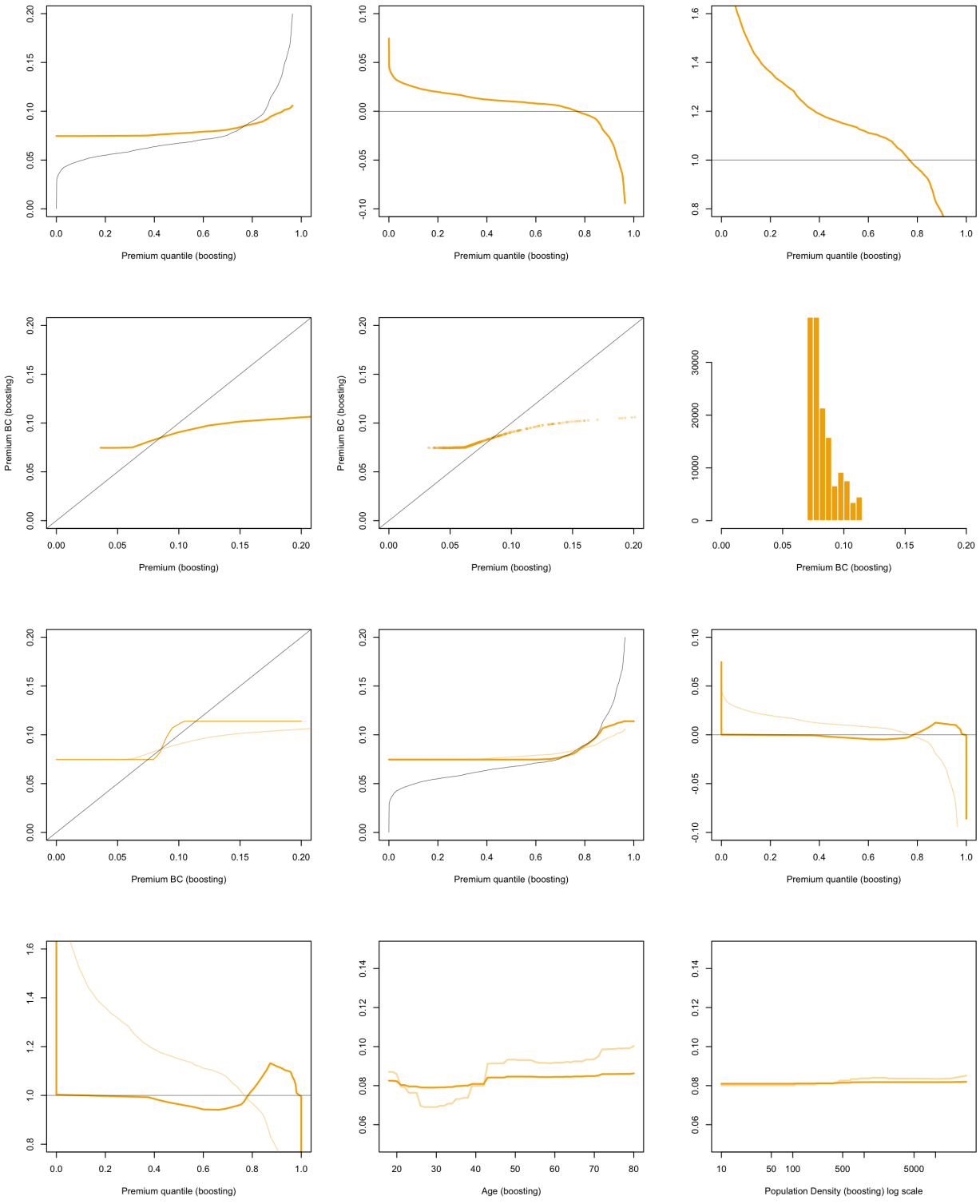

Figure 6.16: Various graphs of $\widehat{\pi}^{\mathsf{bst}}$ and $\widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$ when $\alpha = 80\%$ for $\widehat{\pi}^{\mathsf{gbm}}$

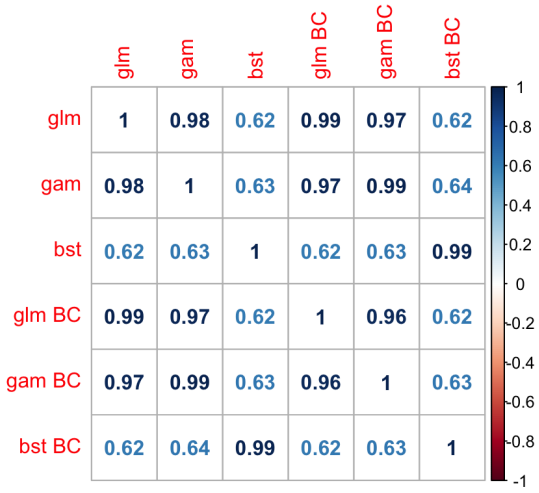

Figure 6.17: Spearman's rank correlation between $\widehat{\pi}^{\mathsf{glm}}$, $\widehat{\pi}^{\mathsf{gam}}$, $\widehat{\pi}^{\mathsf{bst}}$ and their autocalibrated counterparts, $\widehat{\pi}^{\mathsf{glm}}_{\mathrm{BC}}$, $\widehat{\pi}^{\mathsf{gam}}_{\mathrm{BC}}$ and $\widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$, on the validation dataset

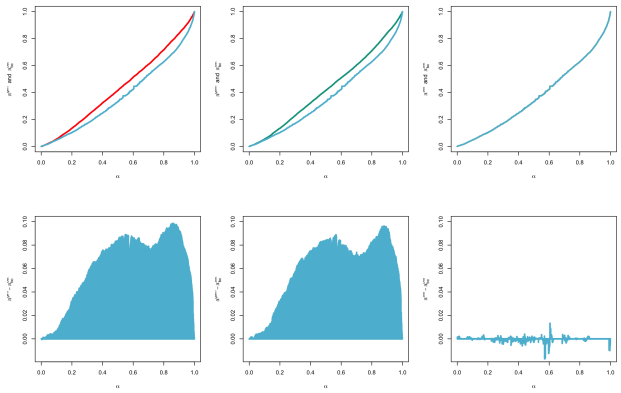

Figure 6.18: Concentration curves, with $\widehat{\pi}^{\mathsf{glm}}$, $\widehat{\pi}^{\mathsf{gam}}$, $\widehat{\pi}^{\mathsf{bst}}$, against $\widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$, from the left to the right on top and $\widehat{\pi}^{\mathsf{glm}} - \widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$, $\widehat{\pi}^{\mathsf{gam}} - \widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$ and $\widehat{\pi}^{\mathsf{bst}} - \widehat{\pi}^{\mathsf{bst}}_{\mathrm{BC}}$ from the left to the right, on the bottom.

of the objective function remains largely ignored and we believe that this component is also fundamental for transparency: end-users should be comfortable with the way the model is trained, which obviously impacts on the role granted to features to explain the response of interest and hence the premium. With minimum bias, the insurer can explain that the amount of premium charged has been computed in order to compensate insurance risks within meaningful sub-portfolios, maintaining an overall balance. The proposed balance correction mechanism implements the very same idea, using machine learning tools to define meaningful neighborhoods where collected premiums must match claims to be compensated. This is in accordance with the fundamental mutuality principle at the heart of insurance. But whether

it is transparent to explain policyholders that their premium minimizes Tweedie deviance is certainly debatable. The consistency of Bregman loss functions for pure premiums offers a convenient justification in that respect. By adopting such loss functions, actuaries can guarantee the policyholders, the consumers' organizations and the regulatory authorities that their approach ensures that the machine learning algorithm searches for the correct pure premium.

References

- Bailey, R.A. (1963). Insurance rates with minimum bias. Proceedings of the Casualty Actuarial Society 50, 4-11.

- Bailey, R.A., Simon, L.J. (1960). Two studies on automobile insurance ratemaking. ASTIN Bulletin 1, 192-217.

- Charpentier, A. (2014). Computational Actuarial Science with R. Chapman and Hall / CRC.

- 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.

- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81, 461–470.

- Ehm, W., Gneiting, T., Jordan, A., Kruger, F. (2016) Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. Journal of the Royal Statistical Society – Series B 78, 505-562.

- Gneiting, T. (2011). Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.

- Kruger, F., Ziegel, J.F. (2020). Generic conditions for forecast dominance. Journal of Business & Economic Statistics, in press.

- Loader, C. (1999). Local Regression and Likelihood. Springer, New York.

- Meyers, G., Cummings, A. D. (2009). "Goodness of Fit" vs. "Goodness of Lift". Actuarial Review 36, 16-17.

- Mildenhall, S.J. (1999). A systematic relationship between minimum bias and generalized linear models. Proceedings of the Casualty Actuarial Society 86, 393-487.

- Savage, L.J. (1971). Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783-810.

- Shaked, M., Shanthikumar, J.G. (2007). Stochastic Orders. Springer, New York.

- Shaked, M., Sordo, M. A., Suarez-Llorens, A. (2012). Global dependence stochastic orders. Methodology and Computing in Applied Probability 14, 617-648.

- Wright, R. (1987). Expectation dependence of random variables, with an application in portfolio theory. Theory and Decision 22, 111-124.

- Wüthrich, M.V. (2019). From Generalized Linear Models to Neural Networks, and back. Available at SSRN 3491790.

- Wüthrich, M.V. (2020). Bias regularization in neural network models for general insurance pricing. European Actuarial Journal 10, 179-202.

- Yitzhaki, S., Schechtman, E. (2013). The Gini Methodology: A Primer on Statistical Methodology. Springer.

- Zumel, N. (2019). An ad-hoc method for calibrating uncalibrated models. Win-Vector Blog, WordPress.