TESTING FOR AUTO-CALIBRATION WITH LORENZ AND CONCENTRATION CURVES

Michel Denuit, Julie Huyghe, Julien Trufin, Thomas Verdebout

REPRINT | 2024 / 15

UCLouvain

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

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

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

Insurance Mathematics and Economics 117 (2024) 130–139

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Testing for auto-calibration with Lorenz and Concentration curves

Michel Denuit [a], Julie Huyghe [b], Julien Trufin [b,*], Thomas Verdebout [c]

[a] Institute of Statistics, Biostatistics and Actuarial Science - ISBA, Louvain Institute of Data Analysis and Modeling - LIDAM, UCLouvain, Louvain-la-Neuve, Belgium
[b] Department of Mathematics, Université Libre de Bruxelles (ULB), Brussels, Belgium
[c] ECARES and Department of Mathematics, Université Libre de Bruxelles (ULB), Brussels, Belgium

ARTICLE INFO

JEL classification:
C12
G22

Keywords:
Concentration curve
Lorenz curve
Integrated Concentration Curve (ICC)
Area Between the Curves (ABC)
Gini coefficient
Auto-calibrated estimators

ABSTRACT

Dominance relations and diagnostic tools based on Lorenz and Concentration curves in order to compare competing estimators of the regression function have recently been proposed. This approach turns out to be equivalent to forecast dominance when the estimators under consideration are auto-calibrated. A new characterization of auto-calibration is established, based on the graphs of Lorenz and Concentration curves. This result is exploited to propose an effective testing procedure for auto-calibration. A simulation study is conducted to evaluate its performances and its relevance for practice is demonstrated on an insurance data set.

1. Introduction and motivation

The estimation of the mean function has attracted a lot of interest in the literature. This task is called regression or supervised learning, and consists in estimating the conditional expectation of a response (or dependent variable) given a set of features (or explanatory variables). The book by Hastie et al. (2009) offers a nice overview of this topic. Comparing the performances of different candidate estimators is thus important in that respect, beyond classical goodness-of-fit criteria.

Mean estimators are obtained by minimizing a loss function on a (training) data set. A loss function is said to be consistent for the mean if no other quantity leads to a lower expected loss than the mean. We refer the reader to Gneiting (2011) for a general exposition. Savage (1971) established that Bregman loss functions are consistent for the mean. Ehm et al. (2016) then defined forecast dominance as dominance with respect to all loss functions that are consistent for the parameter under consideration (thus with respect to all Bregman loss functions for the mean parameter).

Building on Frees et al. (2011), Denuit et al. (2019) suggested to compare the relative merits of estimators with the help of the position of their respective Lorenz and Concentration curves. Denuit et al. (2021b) related this approach to expectation dependence (Kowalczyk and Pleszczynska, 1977; Wright, 1987) and proposed an effective testing procedure for this dependence concept that has been successfully applied to model selection.

Krüger and Ziegel (2021) introduced the important concept of auto-calibration ensuring that estimators can be used without bias correction. Denuit et al. (2021a) demonstrated the usefulness of this concept in insurance studies and described a practical way to auto-calibrate a candidate estimator, referred to as balance correction. An alternative approach has been proposed by Wüthrich and Ziegel (2023), referred to as isotonic recalibration.

Auto-calibration is particularly useful in applications where it is desirable that sums of estimates match sums of observations as closely as possible, at both global and local scales. Consider insurance pricing, for instance. Auto-calibration ensures in this case that the amounts of pure premiums collected from policyholders subject to the same insurance rates match on average the corresponding claim totals. Subsidizing transfers among categories of policyholders are thus avoided in this way. This is the essence of likelihood equations in Generalized Linear Models with canonical link functions and binary features. Auto-calibration is more demanding in that this balance property has to hold in any neighborhood defined by the candidate estimator. The developments in this paper apply in any setting where balance is desirable, that is, where it is important that sums of estimates do not deviate too much from the sum of actual observations at both global and local scales.

* Corresponding author.
E-mail address: julien.trufin@ulb.ac.be (J. Trufin).

https://doi.org/10.1016/j.insmatheco.2024.04.003
Received May 2023; Received in revised form April 2024; Accepted 19 April 2024
Available online 6 May 2024

Auto-calibration also eases the comparison of competing estimators. Under auto-calibration, the comparison methods proposed by Denuit et al. (2019) and Krüger and Ziegel (2021) both reduce to the well-known convex order that has been extensively used in applied probability to compare variability of two random variables beyond the simple comparison of their respective standard deviations (Shaked and Shanthikumar, 2007). It has been shown by Denuit and Trufin (2022) that balance correction implementing auto-calibration improves the performances of the initial estimator as measured by out-of-sample Bregman divergence.

The objective of the present paper is to provide analysts with an effective testing procedure for auto-calibration. As mentioned in Wüthrich and Ziegel (2023), regression models are often not auto-calibrated. There are methods for obtaining auto-calibration through a recalibration step, such as the ones described in Denuit et al. (2021a), in Lindholm et al. (2023) and in Wüthrich and Ziegel (2023), as already mentioned. These methods aim to restore auto-calibration from a given regression model. While these methods are very useful for restoring the auto-calibration property, they are not part of the calibration procedure and come into play at a later stage through a recalibration step. The effective testing procedure for auto-calibration proposed in this paper can be used by the analyst to assess the relevance of applying this recalibration step to any given regression model. Moreover, if an extra-step is used to restore the auto-calibration property, our testing procedure can be useful to check that the recalibration step has been properly implemented. We refer the reader to Henriksen (2023) for a practitioner's point of view about auto-calibration and its relevance for insurance pricing.

From a methodological point of view, the present work decomposes into the following steps. First, we show that Lorenz and Concentration curves coincide under auto-calibration up to a multiplicative constant. Then, we design a testing procedure for the equality of the Lorenz curve and the Concentration curve. Finally, we show that the testing procedure for the equality of these two curves can be turned into a test for auto-calibration. This provides an alternative to the approach developed by Delong and Wüthrich (2023) who derived confidence bands for lift, or reliability diagrams and critical values of miscalibration tests based on strictly consistent scoring functions (deviance from Exponential Dispersion family in their study). The present paper also reconciles the approaches basing model performances assessment on Lorenz curves and Gini coefficients to these recent methodological advances for auto-calibrated estimators.

The remainder of the text is structured as follows. Section 2 recalls the concept of auto-calibration and shows that Concentration and Lorenz curves coincide for auto-calibrated estimators. This offers a new characterization of this concept. In Section 3, we propose a testing procedure for the equality of Lorenz and Concentration curves. A simulation experiment is conducted there to assess its performances. Section 4 exploits the testing procedure obtained in Section 3 to design a testing procedure for auto-calibration. A case study with motor insurance claim data is worked out in Section 5 to illustrate the practical relevance of the results derived in this paper. The final Section 6 summarizes the contents of the paper and concludes. Proofs are gathered in appendix.

2. Lorenz and Concentration curves for auto-calibrated estimators

2.1. Regression for the mean

Consider a response Y of interest. It can be a discrete (e.g. count) or continuous random variable. To explain the mean response, the analyst has at his or her disposal a set of features $X_1, \dots, X_p$ gathered in the vector $\boldsymbol{X}$. The aim is to evaluate the expected response as accurately as possible, so that the target is the conditional expectation $\mu(\boldsymbol{X}) = \mathrm{E}[Y|\boldsymbol{X}]$ of the response Y given the available information $\boldsymbol{X}$.

The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is unknown to the analyst, and may exhibit a complex behavior in $\boldsymbol{x}$. This is why this function is approximated by a function $\boldsymbol{x} \mapsto m(\boldsymbol{x})$ with a simpler structure. For instance, $m(\boldsymbol{X})$ is a monotonic function of a linear combination of $X_1, \dots, X_p$ with Generalized Linear Models (GLMs). Once fitted on the training data set using an appropriate learning procedure, this produces estimates $\widehat{m}(\boldsymbol{x})$ for $\mu(\boldsymbol{x})$, or fitted values.

Assume that $\widehat{m}$ has been obtained by minimizing the empirical loss on some training set (all formulas in the text are meant given this training set). The respective performances of competing estimators can then be assessed on the basis of a validation set $\{(Y_i, \boldsymbol{X}_i), i = 1, 2, \dots, n\}$, that has not been used to obtain $\widehat{m}$. Provided the validation set comprises a large number of independent and identically distributed pairs $(Y_i, \boldsymbol{X}_i)$, this allows us to perform calculations with a generic random vector $(Y, \boldsymbol{X})$ distributed as the observations contained in the training set. This approach is meaningful in applications where the analyst is in a data-rich situation (like the insurance study conducted in Section 5).

The merits of $\widehat{m}$ can be assessed using the pair $\big(\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})\big)$ so that we are back to the bivariate case even if there were thousands of features comprised in $\boldsymbol{X}$. Notice that we allow here for general estimator $\widehat{m}(\boldsymbol{X})$ and conditional expectation $\mu(\boldsymbol{X})$ in the sense that they can be continuous random variables admitting probability density functions or discrete ones. The latter situation corresponds to classification and regression trees (CARTs) where $\widehat{m}$ is a piecewise constant function.

2.2. Auto-calibration and balance correction

We expect strong correlation between $\widehat{m}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ but as $\mu(\boldsymbol{X})$ is unobserved the analyst can only use its noisy version Y to reveal the agreement of $\mu(\boldsymbol{X})$ to $\widehat{m}(\boldsymbol{X})$. In many applications, correlation is important, but it is also essential that the sum of fitted values $\widehat{m}(\boldsymbol{X})$ matches the sum of actual observations as closely as possible. This naturally leads to the concept of auto-calibration proposed by Krüger and Ziegel (2021). Recall that an estimator $\widehat{m}$ is said to be auto-calibrated if $\widehat{m}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{m}(\boldsymbol{X})]$. Auto-calibration is a desirable property since it ensures that the information contained in $\widehat{m}$ is used without any bias. By definition, auto-calibration ensures that sums of estimates closely match sums of actual outcomes in any neighborhood defined by $\widehat{m}$. This is also true for the sum of unobserved conditional expectations. Auto-calibration thus implies $\mathrm{E}[\widehat{m}(\boldsymbol{X})] = \mathrm{E}[Y]$, meaning that $\widehat{m}$ is unbiased on the portfolio level, which should be a minimal requirement in insurance pricing. However, numerical evidence suggests that candidate premiums produced by machine learning tools, like boosted regression trees or neural networks, can depart a lot from observed claim totals. See e.g. Denuit et al. (2019) and the references therein. That is why, in the following, we do not assume that $\mathrm{E}[\widehat{m}(\boldsymbol{X})] = \mathrm{E}[Y]$, so that our approach can be applied to any predictor $\widehat{m}$ produced by machine learning tools.

Auto-calibration can be implemented by replacing the initial estimator with the conditional expectation of the response, given the estimator. Precisely, assume that $\widehat{m}$ is not auto-calibrated. Then, the estimator $\widehat{m}$ can be auto-calibrated by switching to $\widehat{m}_{\mathrm{auto}}$ defined as

$$\widehat{m}_{\mathrm{auto}}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{m}(\boldsymbol{X})].$$

Denuit et al. (2021a) proposed an effective method to implement auto-calibration in insurance studies, referred to as balance correction. In this approach, local polynomial regression allows the analyst to implement auto-calibration, that is, to switch from $\widehat{m}$ to $\widehat{m}_{\mathrm{auto}}$. See Ciatto et al. (2023) for an application to insurance data. Another approach, called isotonic recalibration, has been proposed by Wüthrich and Ziegel (2023). Their suggestion is to apply isotonic regression to form an adaptive partition of the feature space, determining the auto-calibrated estimator by averaging responses over the partition elements.

2.3. Lorenz and Concentration curves

The Concentration curve of $\mu(\boldsymbol{X})$ with respect to $\widehat{m}$ based on the information contained in the vector $\boldsymbol{X}$ is defined as the function

$$\alpha \mapsto \mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \frac{\mathrm{E}\left[\mu(\boldsymbol{X})\mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right]}{\mathrm{E}[\mu(\boldsymbol{X})]}, \ \alpha \in (0,1),$$

where $\mathrm{I}[\cdot]$ denotes the indicator function (equal to 1 if the event appearing within brackets is realized and to 0 otherwise) and $F_{\widehat{m}}^{-1}$ is the quantile function of $\widehat{m}$. We assume here that $F_{\widehat{m}}$ is either (i) known (which is generally not the case in practice) or (ii) estimated through the training set as $\hat{m}$ (see Section 2.1). For (ii), for instance if $\mathbf{X}_{n+1}, \dots, \mathbf{X}_{n+k}$ are the observed covariates in the training set, a possible expression for $F_{\widehat{m}}$ is

$$F_{\widehat{m}}(x) := \frac{1}{k} \sum_{i=n+1}^{n+k} \mathrm{I}[\widehat{m}(\mathbf{X}_i) \le x].$$

As for $\widehat{m}$ (again see Section 2.1), we will tacitly assume below that $F_{\widehat{m}}$ is known and fixed since it is completely independent from the observations in the validation set we will use in our analysis. In practice, we can replace the unknown $\mu(\boldsymbol{X})$ with the response Y in the Concentration curve. Precisely, the following identity holds

$$\mathrm{CC}\left[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha\right] = \mathrm{CC}[Y, \widehat{m}(\boldsymbol{X}); \alpha] = \frac{\mathrm{E}\left[Y\mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right]}{\mathrm{E}[Y]} \tag{2.1}$$

for every probability level α. Formula (2.1) shows that we can equivalently replace the unknown $\mu(\boldsymbol{X})$ with the observed Y, which opens the door to estimation.

The Lorenz curve associated with $\widehat{m}(\boldsymbol{X})$ is defined as the function

$$\begin{aligned}\alpha \mapsto \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha] &= \mathrm{CC}[\widehat{m}(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] \\ &= \frac{\mathrm{E}\left[\widehat{m}(\boldsymbol{X})\mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]}, \ \alpha \in (0,1).\end{aligned} \tag{2.2}$$

Notice that the Lorenz curve only depends on $\widehat{m}$ and not on the response Y (beyond its estimation on the training data set). For an exhaustive review of the properties of Concentration and Lorenz curves, we refer the interested reader to the book by Yitzhaki and Schechtman (2013).

2.4. A new characterization of auto-calibration

Remark 4.3 in Denuit et al. (2019) points out that Concentration and Lorenz curves coincide when $\mu(\boldsymbol{X})$ and $\widehat{m}(\boldsymbol{X})$ are perfectly positively dependent. Proposition 4.1 in Wüthrich (2023) states that if $\widehat{m}$ is auto-calibrated, then its Concentration and Lorenz curves coincide, so that the equality of both curves is a necessary condition for auto-calibration. The next proposition, which is the main result of this section, shows that the equality of both curves actually characterizes the concept of auto-calibration up to a multiplicative constant.

Proposition 2.1. *We have*

$\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha]$ *for every probability level* α

if, and only if $\widehat{m}_{unbiased}(\boldsymbol{X}) = \frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]}\widehat{m}(\boldsymbol{X})$ *is auto-calibrated.*

The proof of Proposition 2.1 can be found in Appendix A. Note that if $\widehat{m}$ is itself an unbiased estimator in the sense that

$$\mathrm{E}[\widehat{m}(\boldsymbol{X})] = \mathrm{E}[Y], \tag{2.3}$$

then $\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})$ in Proposition 2.1 does coincide with $\widehat{m}(\boldsymbol{X})$. The random function $\widehat{m}_{\mathrm{unbiased}}$ is therefore simply an unbiased version of $\widehat{m}$. As a result, the equality of the Lorenz and Concentration curves is equivalent to auto-calibration for unbiased estimators $\widehat{m}$. Bias may for instance come from the use of regularization techniques such as the classical Lasso, Ridge or Elastic-net techniques. From Proposition 2.1, we therefore directly get the following result.

Corollary 2.2. *The estimator* $\widehat{m}$ *is auto-calibrated if and only if (i) it is such that* (2.3) *holds and (ii)*

$\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha]$ *for every* $\alpha \in (0,1)$

or equivalently, if and only if

$$\mathrm{E}\left[(Y - \widehat{m}(\boldsymbol{X}))\mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right] = 0 \text{ for every } \alpha \in (0,1). \tag{2.4}$$

According to Definition 4.1 in Denuit et al. (2019), the best-performing estimator has smaller Lorenz curve and smaller Concentration curve. This ensures that the preferred estimator better correlates to the response and exhibits more variability (and thus contains more information). However, many empirical studies only use the Lorenz curve to evaluate the performances of $\widehat{m}$. This approach is generally flawed since it confuses the estimator $\widehat{m}$ with the conditional expectation (so that Lorenz curve and Concentration curve indeed coincide) whereas it is very likely that $\widehat{m}(\boldsymbol{X})$ at best only approximates $\mu(\boldsymbol{X})$. Therefore, analysts must resort to both Concentration and Lorenz curves to properly assess performances of a candidate unbiased estimator. The main message of Corollary 2.2 is that the use of Lorenz curves only, without reference to Concentration curves, is justified for auto-calibrated estimators.

2.5. ICC, ABC, and Gini metrics

In addition to the respective positions of the graphs of the Concentration and Lorenz curves, Denuit et al. (2019) suggest basing the comparison both on the integral of the Concentration curves (ICC) and on the area between the two curves (ABC). These metrics are defined as

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

and

$$\mathrm{ABC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})] = \int_0^1 \Big(\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] - \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha]\Big)\mathrm{d}\alpha,$$

respectively. A better model has smaller ICC and ABC. Similarly to the ICC for the Concentration curves, we can define the integral of the Lorenz curves (ILC) as

$$\mathrm{ILC}[\widehat{m}(\boldsymbol{X})] = \int_0^1 \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha]\mathrm{d}\alpha,$$

so that we get

$$\mathrm{ABC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})] = \mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})] - \mathrm{ILC}[\widehat{m}(\boldsymbol{X})].$$

The Gini index is defined as twice the area between the Lorenz curve and the diagonal line, so that it is closely related to ILC. Formally, Gini index for a non-negative random variable Z is defined as

$$\mathrm{Gini}[Z] = 2\int_0^1 \big(\alpha - \mathrm{LC}[Z; \alpha]\big)\mathrm{d}\alpha = 1 - 2\mathrm{ILC}[Z].$$

See e.g. Definition 2 in Haye and Zizler (2021). Whereas ICC and ABC have sound methodological foundation, dominant practice uses Gini coefficients to compare predictors, favoring mean estimators with larger Gini index.

The next property is a direct consequence of Corollary 2.2 and the very definition of Gini index. It derives expression of the ICC and ABC metrics for auto-calibrated estimators and relates ICC to Gini coefficient.

Property 2.3. *If* $\widehat{m}$ *is auto-calibrated, then we have*

$$\mathrm{ABC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})] = 0 \text{ and } \mathrm{ICC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X})] = \frac{1}{2}\big(1 - \mathrm{Gini}[\widehat{m}(\boldsymbol{X})]\big).$$

Notice that the relationship between ICC and Gini index stated under Property 2.3 is slightly different from the results derived in Denuit et al. (2019). This is because Gini mean difference is used there whereas here, we use Gini index defined in terms of the Lorenz curve. For auto-calibrated estimators $\widehat{m}_1$ and $\widehat{m}_2$, we get from Property 2.3 that $\widehat{m}_2$ outperforms $\widehat{m}_1$ when

$$\mathrm{ICC}[\mu(\boldsymbol{X}),\widehat{m}_2(\boldsymbol{X})]\leq\mathrm{ICC}[\mu(\boldsymbol{X}),\widehat{m}_1(\boldsymbol{X})]\Leftrightarrow\mathrm{Gini}\big[\widehat{m}_2(\boldsymbol{X})\big]\geq\mathrm{Gini}\big[\widehat{m}_1(\boldsymbol{X})\big].$$

We thus see that comparing model performances with the help of ICC is equivalent to comparing Gini coefficients when estimators are auto-calibrated. Again, this justifies the dominant practice provided auto-calibration has been implemented.

Also, Property 2.3 shows that a nonzero estimated ABC suggests that the estimator under consideration is not auto-calibrated. Initially proposed as an indicator of the performance of a given predictor in Denuit et al. (2019), it turns out that ABC is also useful for detecting the improper implementation of auto-calibration.

3. Testing for equality of concentration and Lorenz curves

3.1. Statistical test

The objective of the present part of the paper is to provide a test for the equality of the Lorenz and Concentration curves based on a validation set consisting of i.i.d. copies $(Y_1,\boldsymbol{X}_1),\dots,(Y_n,\boldsymbol{X}_n)$ of $(Y,\boldsymbol{X})$. More precisely, we want to test the null hypothesis

$$\mathcal{H}_0:\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]=\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha]\text{ for all }\alpha\in(0,1)$$

against the alternative

$$\mathcal{H}_1:\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]\neq\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha]\text{ for some }\alpha\in(0,1).$$

We know from Corollary 2.2 that $\mathcal{H}_0$ corresponds to auto-calibration for $\widehat{m}_{\text{unbiased}}(\boldsymbol{X})=\frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]}\widehat{m}(\boldsymbol{X})$ (or to autocalibration of $\widehat{m}(\boldsymbol{X})$ if it is unbiased). It is worth noticing that here, $\mathcal{H}_0$ is only a necessary condition for $\widehat{m}$ to be auto-calibrated. In Section 4, we provide a test for which $\mathcal{H}_0$ is a necessary and sufficient condition for $\widehat{m}$ to be auto-calibrated. Let $\bar{Y}=\frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{m}=\frac{1}{n}\sum_{i=1}^n\widehat{m}(\boldsymbol{X}_i)$ denote the averaged responses and averaged fitted values, respectively. The testing procedure is naturally based on the difference between sample versions of Concentration and Lorenz curves respectively defined as

$$\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]=\frac{1}{n\bar{Y}}\sum_{i=1}^n Y_i\,\mathrm{I}\left[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)\right],\ \alpha\in(0,1),$$

and

$$\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]=\frac{1}{n\bar{m}}\sum_{i=1}^n\widehat{m}(\boldsymbol{X}_i)\,\mathrm{I}\left[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)\right],\ \alpha\in(0,1).$$

These expressions can be seen as empirical analogues to (2.1)-(2.2). We refer to Gribkova et al. (2022) for related estimators of conditional expectations $\mathrm{E}[Y\,|\widehat{m}(\boldsymbol{X})>F_{\widehat{m}}^{-1}(\alpha)]$ known as tail conditional allocations in quantitative risk management.

The test proposed in this section is naturally based on the difference between $\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]$ and $\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]$. More precisely, the null hypothesis is rejected for large values of the test statistic

$$\mathcal{T}=\sup_{\alpha\in(0,1)}|T_n(\alpha)|,$$

where

$$\begin{aligned}T_n(\alpha)&=\sqrt{n}\left(\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]\right)\\&=n^{-1/2}\sum_{i=1}^n\left(\frac{Y_i}{\bar{Y}}-\frac{\widehat{m}(\boldsymbol{X}_i)}{\bar{m}}\right)\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)],\quad\alpha\in(0,1).\end{aligned}\tag{3.1}$$

In the following result, we derive the asymptotic behavior of $T_n(\alpha)$. Define

$$Z_i(\alpha)=Y_i\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)]\text{ and }W_i(\alpha)=\widehat{m}(\boldsymbol{X}_i)\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)].$$

For $(\alpha_1,\alpha_2)\in(0,1)^2$, let $\Sigma(\alpha_1,\alpha_2)$ be the covariance matrix of the random vector

$$(\widehat{m}(\boldsymbol{X}_i),Y_i,Z_i(\alpha_1),W_i(\alpha_1),Z_i(\alpha_2),W_i(\alpha_2))',$$

assuming that the later exists and is finite. Define also (still assuming that the quantities involved exist)

$$\boldsymbol{v}(\alpha_1,\alpha_2)=\begin{pmatrix}-\frac{e_w(\alpha_1)}{e_m^2}&-\frac{e_z(\alpha_1)}{e_y^2}&e_y^{-1}&e_m^{-1}&0&0\\-\frac{e_w(\alpha_2)}{e_m^2}&-\frac{e_z(\alpha_2)}{e_y^2}&0&0&e_y^{-1}&e_m^{-1}\end{pmatrix}',$$

where

$$e_y=\mathrm{E}[Y_i],\ e_m=\mathrm{E}[\widehat{m}(\boldsymbol{X}_i)],\ e_z(\alpha)=\mathrm{E}[Z_i(\alpha)]\text{ and }e_w(\alpha)=\mathrm{E}[W_i(\alpha)].$$

We are now ready to state the main result of this section.

Proposition 3.1. *Assume that $(Y_i,\widehat{m}(\boldsymbol{X}_i))$, $i=1,2,\dots,n$, are such that $\mathrm{E}[Y_i]\neq0$, $\mathrm{E}[\widehat{m}(\boldsymbol{X}_i)]\neq0$, $\mathrm{E}[\widehat{m}^2(\boldsymbol{X}_i)]<\infty$ and $\mathrm{E}[Y_i^2]<\infty$. Then, under the null hypothesis, $T_n(\alpha)$ converges weakly to a Gaussian process with mean zero and covariance kernel*

$$C(\alpha_1,\alpha_2)=\boldsymbol{v}'(\alpha_1,\alpha_2)\boldsymbol{\Sigma}(\alpha_1,\alpha_2)\boldsymbol{v}(\alpha_1,\alpha_2).$$

The proof of Proposition 3.1 can be found in Appendix B.

As mentioned earlier, the proposed test rejects the null hypothesis when

$$\mathcal{T}=\sup_{\alpha\in(0,1)}|T_n(\alpha)|>c_\beta,$$

for some critical value c_β such that $\mathrm{P}[\sup_{\alpha\in(0,1)}|T_n(\alpha)|>c_\beta]=\beta$. However, Proposition 3.1 does not allow the analyst to compute the critical value c_β since the underlying distribution of $(Y_i,\widehat{m}(\boldsymbol{X}_i))$ remains unknown. Given the asymptotic Gaussian behavior of the process, the non-parametric Monte-Carlo methods of Zhu et al. (2016) can be used. For more details on the non-parametric Monte-Carlo methods used here, we refer the reader to Zhu (2005). The testing procedure $\phi_{\text{equal}}^{(n)}$ for the equality of the Lorenz and Concentration curves then proceeds as follows:

1. Generate B independent samples $\boldsymbol{U}_1,\dots,\boldsymbol{U}_B$, where $\boldsymbol{U}_b=(U_{b1},\dots,U_{bn})$ contains n independent standard Gaussian random variables.
2. Compute

$$\widehat{p}_{\text{equal}}=\frac{1}{B}\sum_{b=1}^B\mathrm{I}\left[\max_{0\leq\alpha\leq1}\left|\Delta\left(\alpha,T_n(\alpha),\boldsymbol{U}_b\right)\right|>\max_{0\leq\alpha\leq1}|T_n(\alpha)|\right],$$

where

$$\begin{aligned}&\Delta\left(\alpha,T_n(\alpha),\boldsymbol{U}_b\right)\\&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\left(\frac{Y_i}{\bar{Y}}-\frac{\widehat{m}(\boldsymbol{X}_i)}{\bar{m}}\right)\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\leq F_{\widehat{m}}^{-1}(\alpha)]-n^{-1/2}T_n(\alpha)\right)U_{bi}.\end{aligned}$$

3. Reject the null hypothesis at the level β when $\widehat{p}_{\text{equal}}<\beta$.

3.2. Simulation study

Let us conduct a simple simulation exercise to assess finite-sample performances of the proposed test. We generated 1500 independent samples of independent and identically distributed observations $(Y_1(\tau),\widehat{m}(\boldsymbol{X}_1)),\dots,(Y_n(\tau),\widehat{m}(\boldsymbol{X}_n))$ such that $\widehat{m}(\boldsymbol{X}_i)$ is uniformly distributed over the unit interval and

$$Y_i(\tau_\ell)=\widehat{m}(\boldsymbol{X}_i)+\tau_\ell(V_i+2)+V_i,\quad\tau_\ell=\ell/10,\ell=0,1,2,\tag{3.2}$$

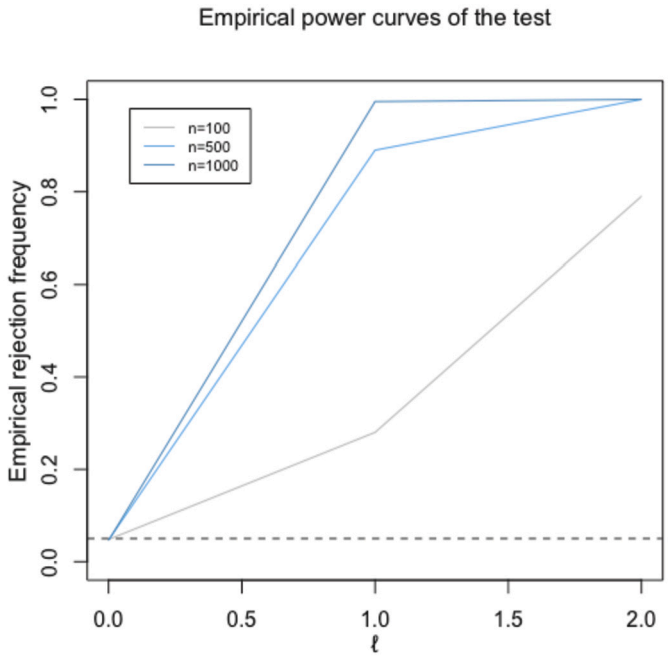

Fig. 3.1. Empirical power curves as functions of ℓ of the test for various sample sizes.

where $V_1, \dots, V_n$ are independent and uniformly distributed over $[-1, 1]$ (and independent of $\widehat{m}(\boldsymbol{X}_i)$). When $\ell = 0$, we have that $\mathrm{E}[Y_i(\tau_\ell)] = \mathrm{E}[\widehat{m}(\boldsymbol{X}_i)] = 1/2$ and

$$\mathrm{E}\big[Y_i\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big] = \mathrm{E}\big[\widehat{m}(\boldsymbol{X}_i)\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big],$$

while for $\ell = 1, 2$,

$$\begin{aligned}\mathrm{E}\big[Y_i\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big] &= \mathrm{E}\big[(\widehat{m}(\boldsymbol{X}_i) + \tau_\ell(V_i + 2) + V_i)\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big] \\ &= \mathrm{E}\big[\widehat{m}(\boldsymbol{X}_i)\mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big] + 2\tau_\ell \alpha \\ &> \mathrm{E}\big[\widehat{m}(\boldsymbol{X}_i)\mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le \alpha]\big],\end{aligned}$$

so that the larger τ_ℓ, the further the data generating process is from the null hypothesis. For each replication, we performed the test $\phi^{(n)}_{\text{equal}}$ at the nominal level $\alpha = .05$. Fig. 3.1 displays empirical rejection frequencies of the test for sample sizes $n = 100$, $n = 500$ and $n = 1000$. Inspection of Fig. 3.1 reveals that the test satisfies the level constraint; the rejection frequencies for $\ell = 0$ are close to the nominal level .05. We carried out another simulation study where we considered a discrete distribution in (3.2) for $\widehat{m}(\boldsymbol{X}_i)$. More precisely we generated observations $Y_i(\tau_\ell)$ following exactly the same scheme as in (3.2), the only difference being that we generated $\widehat{m}(\boldsymbol{X}_i)$'s with a Poisson distribution with parameter 2 (we took $F_{\widehat{m}}$ to be the distribution function of a Poisson with parameter 2). In Fig. 3.2 we plot the empirical rejection frequencies of the test for sample sizes $n = 100$ and $n = 200$. Clearly the test still satisfies the level constraint and behaves quite well. As expected, the proposed testing procedure exhibits higher power as n increases. Since the analysis proposed in this paper applies to data-rich situations where n ranges in the thousands (as in the numerical illustration of Section 5), we conclude that the proposed test performs efficiently in that setting.

4. Testing for auto-calibration

The testing procedure described in Section 3 can be used to test the equality of the Concentration and Lorenz curves, which is equivalent to test for auto-calibration of unbiased estimators only. Nevertheless a test for autocalibration $\phi^{(n)}_{\text{auto}}$ of potentially biased estimators can also be obtained by simply adapting the procedures described in Section 3. Indeed using the empirical version of (2.4) defined as

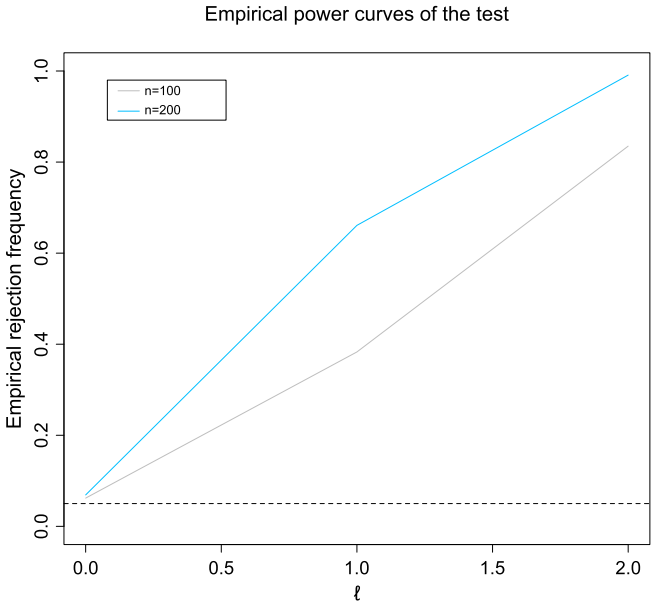

Fig. 3.2. Empirical power curves as functions of ℓ of the test for various sample sizes in the case where the $\widehat{m}(\boldsymbol{X}_i)$'s are Poisson.

$$A_n(\alpha) := n^{-1/2} \sum_{i=1}^{n} \big(Y_i - \widehat{m}(\boldsymbol{X}_i)\big)\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le F^{-1}_{\widehat{m}}(\alpha)],$$

our test will reject for large values of $\sup_{\alpha\in(0,1)}|A_n(\alpha)|$. Mimicking the practical test described at the end of the subsection 3.1, we can proceed as follows:

1. Generate B independent samples $\boldsymbol{U}_1, \dots, \boldsymbol{U}_B$, where $\boldsymbol{U}_b = (U_{b1}, \dots, U_{bn})$ contains n independent standard Gaussian random variables.
2. Compute

$$\widehat{p}_{\text{auto}} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{I}\left[\max_{0\le\alpha\le1} \big|\Delta\left(\alpha, A_n(\alpha), \boldsymbol{U}_b\right)\big| > \max_{0\le\alpha\le1} |A_n(\alpha)|\right],$$

where

$$\begin{aligned}&\Delta\left(\alpha, A_n(\alpha), \boldsymbol{U}_b\right) \\ &\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\big(Y_i - \widehat{m}(\boldsymbol{X}_i)\big)\, \mathrm{I}[\widehat{m}(\boldsymbol{X}_i) \le F^{-1}_{\widehat{m}}(\alpha)] - n^{-1/2} A_n(\alpha)\Big)\, U_{bi}.\end{aligned}$$

3. Reject the null hypothesis at the level β when $\widehat{p}_{\text{auto}} < \beta$.

5. Case study

5.1. Data set

Let us now apply the testing procedures proposed in Sections 3 and 4 to a Swiss motor insurance database used in Wüthrich and Buser (2016) and in Wüthrich (2020). We provide here only a brief description of the data set. The reader is referred to Appendix A of Wüthrich and Buser (2016) for an extensive presentation of the database. This reference also provides the link to the data.

This motor insurance data consists of 500 000 insurance policies. For each policy i ($i = 1, \dots, 500\,000$), the data set records the numbers of claims Y_i filed by policyholder i, the corresponding exposure-to-risk $e_i \le 1$ (i.e. the duration of observation expressed in years) and the following features $\mathbf{X}_i = (X_{i1}, \dots, X_{i8})$:

- X_{i1} = the age of the driver (`age`);
- X_{i2} = the age of the car (`ac`);
- X_{i3} = the power of the car (`power`);

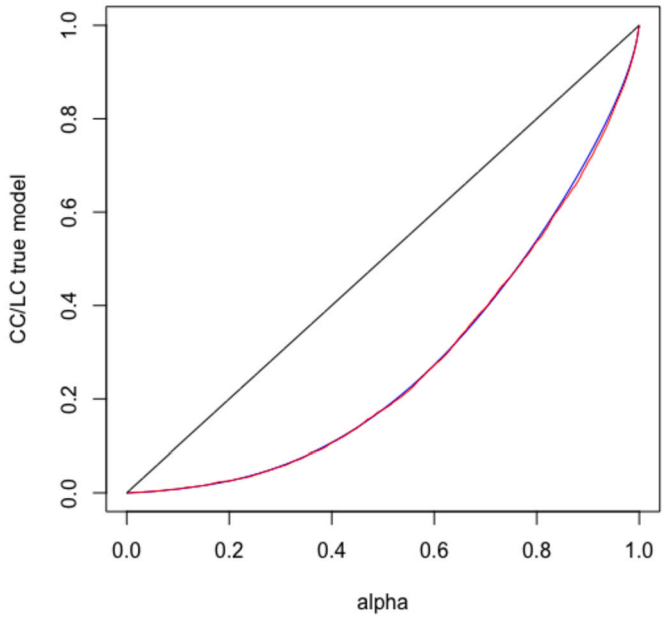

Fig. 5.1. Estimates of the Concentration curve (in red) and Lorenz curve (in blue) on $\overline{\mathcal{D}}$ for the true model. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

- X_{i4} = the fuel type of the car (`fuel`);
- X_{i5} = the vehicle brand (`vehicle brand`);
- X_{i6} = the area code of the living place of the driver (`area`);
- X_{i7} = the population density at the living place of the driver (`dens`);
- X_{i8} = the Swiss canton of the license plate of the car (`ct`).

This data set is special because the values $e_i\mu(\boldsymbol{X}_i)$ of the true model are also provided. Actually, the number of claims Y_i have been generated from expected true model frequencies $e_i\mu(\boldsymbol{X}_i)$.

We partition the data set into a training set $\mathcal{D}$ and a validation set $\overline{\mathcal{D}}$. The training set $\mathcal{D}$ is composed of 80% of the observations taken at random from the entire data set and the validation set $\overline{\mathcal{D}}$ is made of the 20% remaining observations.

Fig. 5.1 displays the estimates of the Concentration curve and Lorenz curve on the validation set $\overline{\mathcal{D}}$ for the true model. One sees that both curves nearly coincide, as expected. Applying our testing procedure to the true model on $\overline{\mathcal{D}}$ with $B = 500$, we get $\widehat{p}_{\text{equal}} = 0.29$ and $\widehat{p}_{\text{auto}} = 0.072$, meaning that the null hypothesis is not rejected at the level 5% for both tests. Our testing procedures thus confirm that we cannot reject that the true model is well auto-calibrated.

5.2. Models under consideration

The observations are assumed to be independent. Given $\boldsymbol{X} = \boldsymbol{x}$ and the exposure-to-risk e, the response Y is assumed to be Poisson distributed with mean $e\mu(\boldsymbol{x})$. Hence, $\mu(\boldsymbol{x}_i)$ represents the expected annual claim frequency for policyholder i. We aim to estimate the function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$ using the observations $(Y_i, \boldsymbol{X}_i)$ in the training set $\mathcal{D}$. To that end, the training set $\mathcal{D}$ is divided into two sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1$ includes 80% of the observations of $\mathcal{D}$ and $\mathcal{D}_2$ gathers the remaining 20%. The subset $\mathcal{D}_1$ is used to train the models while the subset $\mathcal{D}_2$ is used to implement the auto-calibration procedure.

We first fit generalized additive models (GAMs) on $\mathcal{D}_1$ with Poisson deviance loss and log-link function using the `R` package `mgcv`. More precisely, we fit two GAMs producing the following estimators:

- $\widehat{m}^{\text{GAM1}}(\boldsymbol{x})$, with only the feature X_1 (`age`);
- $\widehat{m}^{\text{GAM2}}(\boldsymbol{x})$, using all 8 available features.

Table 5.1
Values of $\widehat{p}_{\text{equal}}$ and $\widehat{p}_{\text{auto}}$ for $B = 500$.

$\widehat{m}$	$\widehat{p}_{\text{equal}}$	$\widehat{p}_{\text{auto}}$
$\widehat{m}^{\text{GAM1}}$	0.000	0.000
$\widehat{m}^{\text{GAM2}}$	0.744	0.516
$\widehat{m}^{\text{GBT1}}$	0.000	0.000
$\widehat{m}^{\text{GBT2}}$	0.000	0.000
$\widehat{m}^{\text{GBT3}}$	0.000	0.000
$\widehat{m}^{\text{GBT4}}$	0.000	0.000

The effects of the covariates `age`, `ac`, `power` and `dens` are captured by splines and we do not consider interaction terms between features.

Then, we train gradient boosting trees (GBTs) on $\mathcal{D}_1$ with Poisson deviance loss and log-link function with the help of the `R` package `gbm`. The bagging fraction γ (that is, the fraction of observations randomly selected from the training set to fit the next tree in the expansion) is set at 0.5. The shrinkage parameter κ is set at the low value 0.01. The size of the trees is controlled by the interaction depth `ID`. We fit four GBTs on 80% of the observations in $\mathcal{D}_1$ producing the following estimators:

- $\widehat{m}^{\text{GBT1}}(\boldsymbol{x})$, with only the feature X_1 (`age`);
- $\widehat{m}^{\text{GBT2}}(\boldsymbol{x})$, using all 8 available features and `ID` = 1;
- $\widehat{m}^{\text{GBT3}}(\boldsymbol{x})$, using all 8 available features and `ID` = 2;
- $\widehat{m}^{\text{GBT4}}(\boldsymbol{x})$, using all 8 available features and `ID` = 3.

The optimal values for the number of trees T are determined by minimizing the out-of-sample Poisson deviance loss computed on the remaining 20% of observations in $\mathcal{D}_1$. We get $T = 507$ for $\widehat{m}^{\text{GBT1}}$, $T = 2977$ for $\widehat{m}^{\text{GBT2}}$, $T = 2984$ for $\widehat{m}^{\text{GBT3}}$ and $T = 2943$ for $\widehat{m}^{\text{GBT4}}$.

5.3. Testing for equality of concentration and Lorenz curves

Fig. 5.2 displays the estimates of the Concentration and Lorenz curves on the validation set $\overline{\mathcal{D}}$ for the models under consideration. While it seems clear that the curves do not coincide for $\widehat{m}^{\text{GAM1}}$, $\widehat{m}^{\text{GBT1}}$, $\widehat{m}^{\text{GBT2}}$, $\widehat{m}^{\text{GBT3}}$ and $\widehat{m}^{\text{GBT4}}$, the conclusion is not so obvious for $\widehat{m}^{\text{GAM2}}$. Table 5.1 contains the values of $\widehat{p}_{\text{equal}}$ and $\widehat{p}_{\text{auto}}$ computed with $B = 500$. The null hypothesis $\mathcal{H}_0 : \text{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \text{LC}[\widehat{m}(\boldsymbol{X}); \alpha]$ for all $\alpha \in (0, 1)$ is always rejected for $\widehat{m}^{\text{GAM1}}$, $\widehat{m}^{\text{GBT1}}$, $\widehat{m}^{\text{GBT2}}$, $\widehat{m}^{\text{GBT3}}$, $\widehat{m}^{\text{GBT4}}$ whatever the level β. Turning to $\widehat{m}^{\text{GAM2}}$, we do not reject the null hypothesis for both tests at the level 5%, so that we do not reject that $\widehat{m}^{\text{GAM2}}$ is auto-calibrated.

5.4. Auto-calibrating models

A local polynomial regression approach allows the analyst to implement auto-calibration, that is, to switch from $\widehat{m}$ to $\widehat{m}_{\text{auto}}$. To this end, the only feature entering the analysis is the estimator $\widehat{m}$ (fitted on $\mathcal{D}_1$) needing auto-calibration. A local GLM is fitted to $(Y_i, e_i, \widehat{m}(\boldsymbol{x}_i))$ comprised in $\mathcal{D}_2$. We refer to Loader (1999) for an introduction to local polynomial models and to the `locfit` package in `R` implementing local polynomial regression.

Specifically, let us consider a specific risk profile $\boldsymbol{x}$. Uniform weights are assigned to each $(Y_i, e_i, \widehat{m}(\boldsymbol{x}_i))$ in the neighborhood $\mathcal{V}(\boldsymbol{x})$ of $\boldsymbol{x}$. Thus, a rectangular (or uniform) weight function is specified. The choice of this particular weight function is justified by the fact that we aim to implement local balance, that is, close agreement between sums of estimates and actual values at both global (unbiasedness) and local scales. This is particularly appealing in the present application to insurance since it ensures that every group of policyholders paying the same premium is on average self-financing. The neighborhood $\mathcal{V}(\boldsymbol{x})$ gathers the fraction δ of the data with the closest predictors $\widehat{m}(\boldsymbol{x}_i)$ to $\widehat{m}(\boldsymbol{x})$. Here, δ acts as a smoothing parameter and controls the size of sub-portfolios

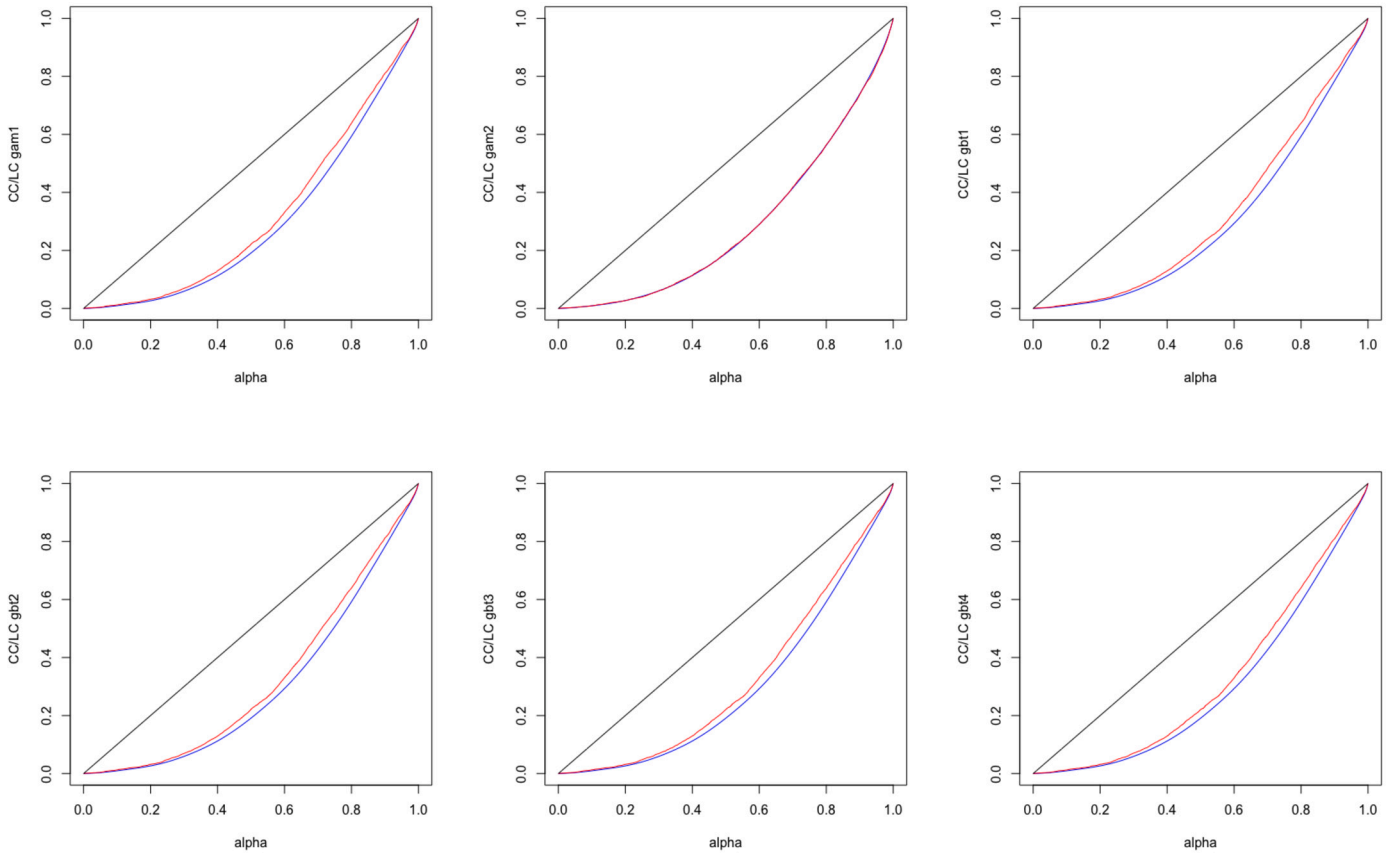

Fig. 5.2. Estimates of the Concentration curves (in red) and Lorenz curves (in blue) on $\overline{D}$ for $\widehat{m}^{\text{GAM1}}$ (top-left), $\widehat{m}^{\text{GAM2}}$ (top-middle), $\widehat{m}^{\text{GBT1}}$ (top-right), $\widehat{m}^{\text{GBT2}}$ (bottom-left), $\widehat{m}^{\text{GBT3}}$ (bottom-middle), $\widehat{m}^{\text{GBT2}}$ (bottom-right).

where local balance is imposed. The optimal value for the parameter δ is selected by likelihood cross-validation (LCV).

The local GLM likelihood equation

$$\sum_{i\in\mathcal{V}(\boldsymbol{x})} y_i = \sum_{i\in\mathcal{V}(\boldsymbol{x})} e_i \widehat{m}_{\text{auto}}(\boldsymbol{x})$$

transfers part of the experience at neighboring $\widehat{m}$ values to obtain $\widehat{m}_{\text{auto}}$. The latter equality implements the self-financing constraint, as explained before. A local constant, or intercept-only GLM thus implements local balance in sub-portfolios gathering policyholders with about the same estimated value.

The selection of the optimal value for δ is based on a LCV plot produced with the help of the `lcvplot` function in `locfit`, which calls the `lcv` function for a grid of smoothing parameters δ. The LCV plot graphs the likelihood cross-validation statistic against the equivalent degree of freedom of the fit, for each smoothing parameter δ. We only implement the auto-calibration procedure for estimators $\widehat{m}^{\text{GAM1}}$, $\widehat{m}^{\text{GBT1}}$, $\widehat{m}^{\text{GBT2}}$, $\widehat{m}^{\text{GBT3}}$ and $\widehat{m}^{\text{GBT4}}$ since those are the only estimators for which the null hypothesis is rejected. Optimal values are respectively equal to $\delta = 3.70\%$ for $\widehat{m}^{\text{GAM1}}$, $\delta = 2.90\%$ for $\widehat{m}^{\text{GBT1}}$, $\delta = 3.20\%$ for $\widehat{m}^{\text{GBT2}}$, $\delta = 3.35\%$ for $\widehat{m}^{\text{GBT3}}$ and $\delta = 3.15\%$ for $\widehat{m}^{\text{GBT4}}$.

Fig. 5.3 displays the estimates of the Concentration curves and Lorenz curves on the validation set $\overline{D}$ for $\widehat{m}^{\text{GAM1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT2}}_{\text{auto}}$, $\widehat{m}^{\text{GBT3}}_{\text{auto}}$ and $\widehat{m}^{\text{GBT4}}_{\text{auto}}$. Compared to Fig. 5.2 for $\widehat{m}^{\text{GAM1}}$, $\widehat{m}^{\text{GBT1}}$, $\widehat{m}^{\text{GBT2}}$, $\widehat{m}^{\text{GBT3}}$ and $\widehat{m}^{\text{GBT4}}$, we can see that the curves have moved closer to each other in all cases. Table 5.2 contains the values of $\widehat{p}_{\text{equal}}$ and $\widehat{p}_{\text{auto}}$ computed with $B = 500$ for $\widehat{m}^{\text{GAM1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT2}}_{\text{auto}}$, $\widehat{m}^{\text{GBT3}}_{\text{auto}}$ and $\widehat{m}^{\text{GBT4}}_{\text{auto}}$. As expected, the testing procedures do not reject anymore at the level 5% that estimators $\widehat{m}^{\text{GAM1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT1}}_{\text{auto}}$, $\widehat{m}^{\text{GBT2}}_{\text{auto}}$, $\widehat{m}^{\text{GBT3}}_{\text{auto}}$ and $\widehat{m}^{\text{GBT4}}_{\text{auto}}$ are auto-calibrated.

Table 5.2
Values of $\widehat{p}_{\text{equal}}$ and $\widehat{p}_{\text{auto}}$ for $B = 500$.

$\widehat{m}$	$\widehat{p}_{\text{equal}}$	$\widehat{p}_{\text{auto}}$
$\widehat{m}^{\text{GAM1}}_{\text{auto}}$	0.072	0.322
$\widehat{m}^{\text{GBT1}}_{\text{auto}}$	0.074	0.328
$\widehat{m}^{\text{GBT2}}_{\text{auto}}$	0.052	0.208
$\widehat{m}^{\text{GBT3}}_{\text{auto}}$	0.092	0.262
$\widehat{m}^{\text{GBT4}}_{\text{auto}}$	0.238	0.252

6. Discussion

The developments in this paper apply to any regression for the mean setting where balance is desirable at both global and local scales, making auto-calibration relevant. This is for instance the case with the insurance application proposed in this paper, where it is important that sums of estimates (considered as pure premiums) closely match sums of corresponding observations (actual insurance losses) to ensure equilibrium of insurance operations.

The new testing procedure proposed in this paper can be applied to check whether a candidate estimator is auto-calibrated. If not, balance correction can be performed with the help of a locally constant fit to make it auto-calibrated. This technique is shown to be effective, since the proposed test does not reject the null hypothesis supporting auto-calibration for the resulting estimators.

Once auto-calibrated, performances of estimators can be properly assessed with the help of Lorenz curves and Gini indices. Several competing estimators can be compared in this way, whereas Concentration curves and ICC must be used in general. This finding is relevant for practice since it questions the wide use of Lorenz curves even for estimators which are not auto-calibrated.

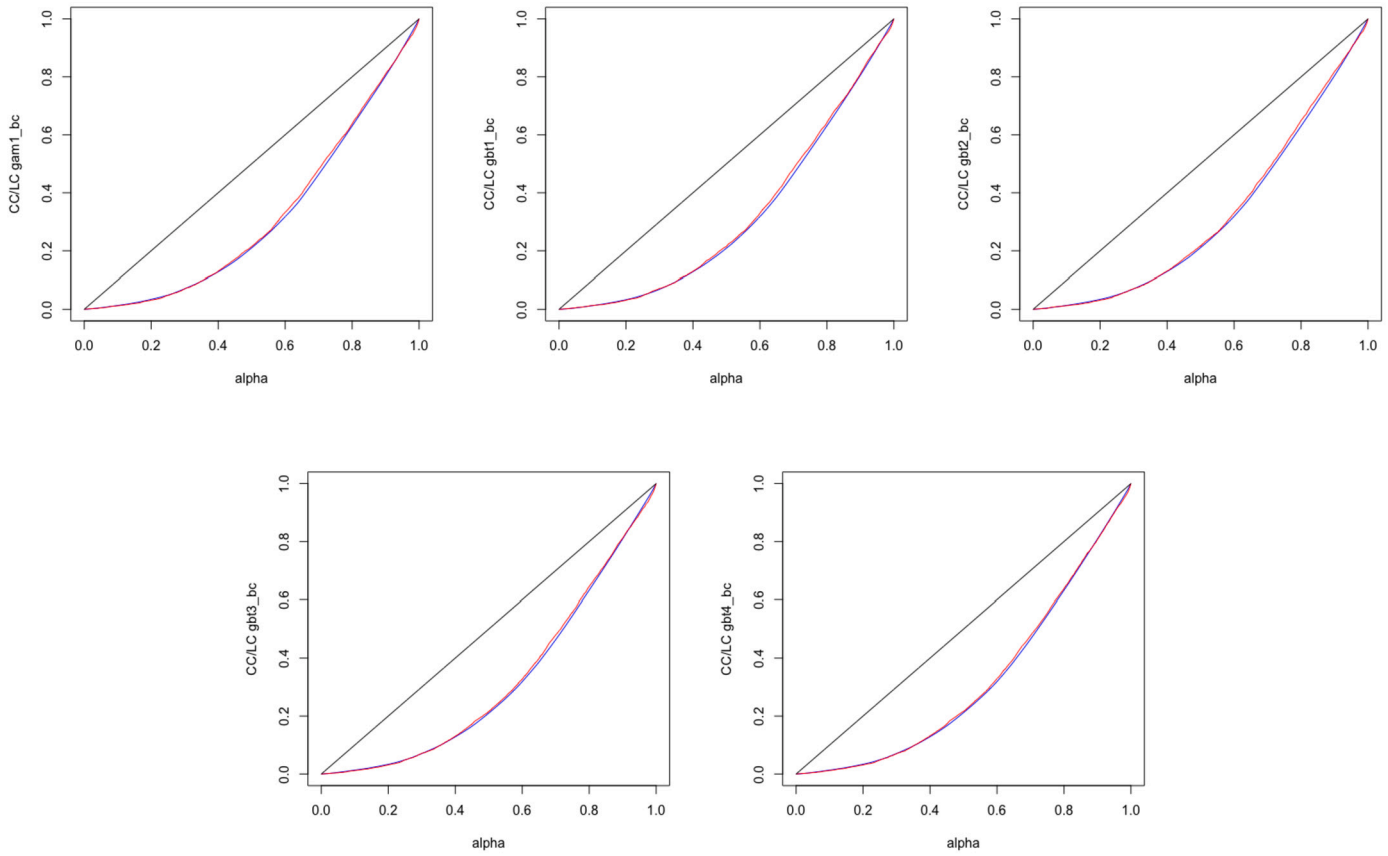

Fig. 5.3. Estimates of the Concentration curves (in red) and Lorenz curves (in blue) on $\overline{\mathcal{D}}$ for $\widehat{m}_{\text{auto}}^{\text{GAM1}}$ (top-left), $\widehat{m}_{\text{auto}}^{\text{GBT1}}$ (top-middle), $\widehat{m}_{\text{auto}}^{\text{GBT2}}$ (top-right), $\widehat{m}_{\text{auto}}^{\text{GBT3}}$ (bottom-left) and $\widehat{m}_{\text{auto}}^{\text{GBT4}}$ (bottom-right).

CRediT authorship contribution statement

Michel Denuit: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Julie Huyghe:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Julien Trufin:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Thomas Verdebout:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Data availability

Data will be made available on request.

Appendix A. Proof of Proposition 2.1

For any $\alpha \in (0,1)$, we have

$$
\begin{aligned}
&\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha] \\
&\Leftrightarrow \frac{\mathrm{E}\left[Y \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right]}{\mathrm{E}[Y]} = \frac{\mathrm{E}\left[\widehat{m}(\boldsymbol{X}) \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]} \\
&\Leftrightarrow \mathrm{E}\left[Y \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right] = \frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]} \mathrm{E}\left[\widehat{m}(\boldsymbol{X}) \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right] \\
&\Leftrightarrow \mathrm{E}\left[Y \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right] = \mathrm{E}\left[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \mathrm{I}[\widehat{m}(\boldsymbol{X}) \le F_{\widehat{m}}^{-1}(\alpha)]\right] \\
&\Leftrightarrow \mathrm{E}\left[Y \mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \le F_{\widehat{m}_{\text{unbiased}}}^{-1}(\alpha)]\right] \\
&\quad = \mathrm{E}\left[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \le F_{\widehat{m}_{\text{unbiased}}}^{-1}(\alpha)]\right]
\end{aligned}
$$

since $F_{a\widehat{m}}^{-1}(\alpha) = a F_{\widehat{m}}^{-1}(\alpha)$ for any constant $a > 0$. Therefore, we get

$$
\begin{aligned}
&\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha] \text{ for all } \alpha \in (0,1) \\
&\Leftrightarrow \mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) \mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \le F_{\widehat{m}_{\text{unbiased}}}^{-1}(\alpha)]\right] \\
&\quad = 0 \text{ for all } \alpha \in (0,1) \\
&\Leftrightarrow \mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) \mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \le t]\right] = 0 \text{ for all real } t. \qquad \text{(A.1)}
\end{aligned}
$$

Let us now show that (A.1) is equivalent to

$$
\mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) g\left(\widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right)\right] = 0 \qquad \text{(A.2)}
$$

for all measurable functions $g(\cdot)$. Assume that (A.1) holds and that g is non negative (it suffices to use the classical decomposition of g into its positive and negative parts to extend the result). Then the monotone approximation Theorem entails that there exists a sequence of simple functions g_k such that $g_k \le g$ and $\lim_{k\to\infty} g_k = g$. It is well known that any simple function g_k can be decomposed as

$$
g_k(\widehat{m}_{\text{unbiased}}(\boldsymbol{X})) = \sum_{i=1}^{k} a_i (\mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) \le b_i] - \mathrm{I}[\widehat{m}_{\text{unbiased}}(\boldsymbol{X}) < b_{i-1}])
$$

for some real numbers $a_1, \ldots, a_k$ and $b_0, \ldots, b_k$. Then the monotone convergence Theorem entails that

$$
\begin{aligned}
&\mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) g\left(\widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right)\right] \\
&\quad = \mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) \lim_{k\to\infty} g_k\left(\widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right)\right] \\
&\quad = \lim_{k\to\infty} \mathrm{E}\left[\left(Y - \widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right) g_k\left(\widehat{m}_{\text{unbiased}}(\boldsymbol{X})\right)\right] \\
&\quad = 0,
\end{aligned}
$$

where the last line directly follows from (A.1). Therefore (A.2) holds. Obviously, (A.2) implies (A.1), so that we have shown that

$$\begin{aligned}
&\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]=\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha] \text{ for all } \alpha\in(0,1)\\
\Leftrightarrow\ &\mathrm{E}\left[\left(Y-\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\right)\mathrm{I}[\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\le t]\right]=0 \text{ for all real } t\\
\Leftrightarrow\ &\mathrm{E}\left[\left(Y-\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\right)g\left(\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\right)\right]=0
\end{aligned}$$

for all measurable functions $g(\cdot)$. The result then follows from the classical orthogonal projection property of conditional expectations which tells us that $\widehat{m}_{\mathrm{unbiased}}$ is auto-calibrated if, and only if,

$$\mathrm{E}\left[\left(Y-\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\right)g\left(\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X})\right)\right]=0$$

for all measurable functions $g(\cdot)$.

Appendix B. Proof of Proposition 3.1

The process

$$\alpha\to T_n(\alpha)=n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i}{\bar{Y}}-\frac{\widehat{m}(\boldsymbol{X}_i)}{\bar{m}}\right)\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(\alpha)]$$

is a weighted empirical process. Under the null hypothesis, we have

$$\begin{aligned}
T_n(\alpha)&=\sqrt{n}\Big(\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]\\
&\qquad-\left(\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha]\right)\Big)\\
&=\sqrt{n}\left(\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]\right)\\
&\qquad-\sqrt{n}\left(\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha]\right)\\
&=T_{1n}(\alpha)+T_{2n}(\alpha).
\end{aligned}$$

Let us first consider

$$T_{1n}(\alpha)=\sqrt{n}\left(\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]\right).$$

Denoting as $\hat{e}_{y,n}$ and $\hat{e}_{m,n}$ the empirical estimators of e_y and e_m, respectively, we have that

$$\begin{aligned}
T_{1n}(\alpha)&=\sqrt{n}\left(\widehat{\mathrm{CC}}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{CC}[\mu(\boldsymbol{X}),\widehat{m}(\boldsymbol{X});\alpha]\right)\\
&=n^{-1/2}\sum_{i=1}^{n}\Big(\hat{e}_{y,n}^{-1}Y_i\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(\alpha)]\\
&\qquad-e_y^{-1}\mathrm{E}\left[Y_i\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(\alpha)]\right]\Big).
\end{aligned}$$

With $Z_i(\alpha)=Y_i\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(\alpha)]$, the delta method together with the law of large numbers allows us to write

$$\begin{aligned}
T_{1n}(\alpha)&=n^{-1/2}\sum_{i=1}^{n}\left(\hat{e}_{y,n}^{-1}Z_i(\alpha)-e_y^{-1}\mathrm{E}[Z_i(\alpha)]\right)\\
&=n^{-1/2}\sum_{i=1}^{n}\left(\hat{e}_{y,n}^{-1}Z_i(\alpha)-e_y^{-1}Z_i(\alpha)+e_y^{-1}Z_i(\alpha)-e_y^{-1}\mathrm{E}[Z_i(\alpha)]\right)\\
&=\sqrt{n}(\hat{e}_{y,n}^{-1}-e_y^{-1})\left(n^{-1}\sum_{i=1}^{n}Z_i(\alpha)\right)\\
&\qquad+e_y^{-1}\left(n^{-1/2}\sum_{i=1}^{n}Z_i(\alpha)-\mathrm{E}[Z_i(\alpha)]\right)\\
&=-e_y^{-2}\sqrt{n}(\hat{e}_{y,n}-e_y)\left(n^{-1}\sum_{i=1}^{n}Z_i(\alpha)\right)\\
&\qquad+e_y^{-1}\left(n^{-1/2}\sum_{i=1}^{n}Z_i(\alpha)-\mathrm{E}[Z_i(\alpha)]\right)+o_{\mathrm{P}}(1)\\
&=-\frac{e_z(\alpha)}{e_y^2}\sqrt{n}(\hat{e}_{y,n}-e_y)\\
&\qquad+e_y^{-1}\left(n^{-1/2}\sum_{i=1}^{n}Z_i(\alpha)-\mathrm{E}[Z_i(\alpha)]\right)+o_{\mathrm{P}}(1)
\end{aligned}\tag{B.1}$$

as $n\to\infty$.

Working along the same lines with

$$W_i(\alpha)=\widehat{m}(\boldsymbol{X}_i)\,\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(\alpha)],$$

we have that

$$\begin{aligned}
T_{2n}(\alpha)&=\sqrt{n}\left(\widehat{\mathrm{LC}}[\widehat{m}(\boldsymbol{X});\alpha]-\mathrm{LC}[\widehat{m}(\boldsymbol{X});\alpha]\right)\\
&=n^{-1/2}\sum_{i=1}^{n}\left(\hat{e}_{m,n}^{-1}W_i(\alpha)-e_m^{-1}\mathrm{E}[W_i(\alpha)]\right)\\
&=-\frac{e_w(\alpha)}{e_m^2}\sqrt{n}(\hat{e}_{m,n}-e_m)\\
&\qquad+e_m^{-1}\left(n^{-1/2}\sum_{i=1}^{n}W_i(\alpha)-\mathrm{E}[W_i(\alpha)]\right)+o_{\mathrm{P}}(1),
\end{aligned}\tag{B.2}$$

as $n\to\infty$.

Putting together (B.1) and (B.2), we therefore obtain that

$$\begin{aligned}
T_n(\alpha)&=T_{1n}(\alpha)+T_{2n}(\alpha)\\
&=n^{-1/2}\sum_{i=1}^{n}D_i(\alpha)-\mathrm{E}[D_i(\alpha)]+o_{\mathrm{P}}(1),
\end{aligned}$$

where

$$D_i(\alpha)=-\frac{e_z(\alpha)}{e_y^2}Y_i-\frac{e_w(\alpha)}{e_m^2}\widehat{m}(\boldsymbol{X}_i)+e_y^{-1}Z_i(\alpha)+e_m^{-1}W_i(\alpha).$$

The central-limit theorem then ensures that

$$\boldsymbol{T}_n(\alpha_1,\alpha_2)=n^{-1/2}\begin{pmatrix}\sum_{i=1}^{n}\widehat{m}(\boldsymbol{X}_i)-\mathrm{E}[\widehat{m}(\boldsymbol{X}_i)]\\ \sum_{i=1}^{n}Y_i-\mathrm{E}[Y_i]\\ \sum_{i=1}^{n}Z_i(\alpha_1)-\mathrm{E}[Z_i(\alpha_1)]\\ \sum_{i=1}^{n}W_i(\alpha_1)-\mathrm{E}[W_i(\alpha_1)]\\ \sum_{i=1}^{n}Z_i(\alpha_2)-\mathrm{E}[Z_i(\alpha_2)]\\ \sum_{i=1}^{n}W_i(\alpha_2)-\mathrm{E}[W_i(\alpha_2)]\end{pmatrix}$$

converges weakly to a Gaussian random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}(\alpha_1,\alpha_2)$. Therefore since

$$\begin{pmatrix}T_n(\alpha_1)\\ T_n(\alpha_2)\end{pmatrix}=\mathbf{v}'(\alpha_1,\alpha_2)\mathbf{T}_n(\alpha_1,\alpha_2)+o_{\mathrm{P}}(1)$$

as $n\to\infty$, the convergence in finite distribution follows.

The announced result then follows if we are able to show that the class of functions

$$(Y,m,t)\to-\frac{e_z(t)}{e_y^2}Y-\frac{e_w(t)}{e_m^2}m+(e_y^{-1}Y+e_m^{-1}m)\mathrm{I}[m\le t]$$

is a Donsker class. First note that taking $(t_1,t_2)\in(0,1)^2$ (without loss of generality, we can consider $t_1>t_2$), the Cauchy-Schwartz inequality yields

$$\begin{aligned}
|e_z(t_1)-e_z(t_2)|&=\left|\mathrm{E}\left[Y_i\left(\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(t_1)]-\mathrm{I}[\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(t_2)]\right)\right]\right|\\
&\le\left(\mathrm{E}[Y_i^2]\right)^{1/2}\left(\mathrm{P}[F_{\widehat{m}}^{-1}(t_2)\le\widehat{m}(\boldsymbol{X}_i)\le F_{\widehat{m}}^{-1}(t_1)]\right)^{1/2}\\
&=\left(\mathrm{E}[Y_i^2]\right)^{1/2}(t_1-t_2)^{1/2},
\end{aligned}$$

so that since Y_i has finite second order moments, the class $(Y,m,t)\to-\frac{e_z(t)}{e_y^2}Y$ is Donsker. Working along the same lines $(Y,m,t)\to-\frac{e_w(t)}{e_m^2}m$ is also Donsker. Finally, it is easy to check that $(Y,m,t)\to(e_y^{-1}Y+e_m^{-1}m)\mathrm{I}[m\le t]$ is a Vapnik–Chervonenkis, or VC class and is therefore Donsker, too. Since a sum of Donsker classes is Donsker, the announced result follows.

References

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

Delong, L. Wüthrich, M.V., 2023. The role of the variance function in mean estimation and validation. Available at SSRN 4477677.

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

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

Denuit, M., Trufin, J., 2022. Autocalibration by balance correction in nonlife insurance pricing. LIDAM Discussion Paper 2022/41, available from dial.uclouvain.be.

Denuit, M., Trufin, J., Verdebout, Th., 2021b. Testing for more positive expectation dependence with application to model comparison. Insurance: Mathematics and Economics 101, 163–172.

Ehm, W., Gneiting, T., Jordan, A., Krüger, 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.

Frees, E.W., Meyers, G., Cummings, A.D., 2011. Summarizing insurance scores using a Gini index. Journal of the American Statistical Association 106, 1085–1098.

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

Gribkova, N.V., Su, J., Zitikis, R., 2022. Inference for the tail conditional allocation: large sample properties, insurance risk assessment, and compound sums of concomitants. Insurance: Mathematics and Economics 107, 199–222.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning. Springer, New York.

Haye, R.L., Zizler, P., 2021. Applications of the Gini index beyond economics and statistics. In: Handbook of the Mathematics of the Arts and Sciences. Springer, pp. 1661–1689.

Henriksen, S., 2023. Modelling the insurance pure premium using autocalibrated machine learning models. Available from https://medium.com/eika-tech/modelling-the-insurance-pure-premium-using-autocalibrated-machine-learning-models-0bd398b8a70f.

Kowalczyk, T., Pleszczynska, E., 1977. Monotonic dependence functions of bivariate distributions. The Annals of Statistics 5, 1221–1227.

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

Lindholm, M., Lindskog, F., Palmquist, J., 2023. Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. Scandinavian Actuarial Journal 2023 (10), 946–973.

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

Savage, L.J., 1971. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66, 783–801.

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

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

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

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

Wüthrich, M.V. Buser, C., 2016. Data analytics for non-life insurance pricing. SSRN Manuscript ID 2870308, version of June 4, 2019.

Wüthrich, M.V., Ziegel, J., 2023. Isotonic recalibration under a low signal-to-noise ratio. arXiv preprint arXiv:2301.02692.

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

Zhu, L., 2005. Nonparametric Monte Carlo Tests and Their Applications, vol. 182. Springer Science and Business Media.

Zhu, X., Guo, X., Lin, L., Zhu, L., 2016. Testing for positive expectation dependence. Annals of the Institute of Statistical Mathematics 68, 135–153.