DOES AUTOCALIBRATION IMPROVE GOODNESS OF LIFT?

Nicolas Ciatto, Harrison Verelst, Julien Trufin, Michel Denuit

REPRINT | 2023 / 07

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

European Actuarial Journal (2023) 13:479–486
https://doi.org/10.1007/s13385-022-00330-4

LETTERS

Does autocalibration improve goodness of lift?

Nicolas Ciatto[1] · **Harrison Verelst**[1] · **Julien Trufin**[1] · **Michel Denuit**[1]

Received: 6 July 2022 / Revised: 14 September 2022 / Accepted: 14 September 2022 /
Published online: 21 September 2022

Abstract
Autocalibration is a desirable property since it ensures that the information contained in a candidate premium is used without any bias. It turns out to be intimately related to the method of marginal totals that predates modern risk classification methods. The present note aims to assess the impact of autocalibration on the goodness of lift. It is shown on a case study that autocalibration does not only restore global and local balances but also improve lift.

Keywords Risk classification · Ratemaking · Autocalibration · Lift curve

1 Introduction and motivation

The starting point of the present analysis is that balance is required in ratemaking. It is indeed 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 that sums of pure premiums must match corresponding claim totals as accurately as possible.

Flexible pricing tools like boosting and neural networks are able to produce premiums that correlate a lot with the response compared to models imposing a rigid form for the predictor (like Generalized Linear Models or GLMs, for instance). But more flexible models often do not impose any constraint on the replication of observed totals. With boosting techniques and neural networks, the sum of fitted values can depart from the observed totals to a large extent and this often confuses actuarial analysts (see e.g. [9, 10]). To solve this problem, [3] proposed a new strategy based on the concept of autocalibration (see, e.g., [4]) to restore global balance as well as local equilibrium in the spirit of the original method of marginal totals dating back to the 1960s. Specifically, after the analysis has been performed with a method that does not necessarily respect marginal totals, a local constant GLM fit is

✉ Julien Trufin
julien.trufin@ulb.ac.be

[1] Universite Libre de Bruxelles, Brussels, Belgium

achieved. The extra step should not be performed on the same data in order to avoid a potential source of overfitting (called nested-model bias). By using a local constant, or intercept-only GLM, uncorrected premiums define optimal neighborhoods to perform local averaging of observed losses.

The present note aims to explore the impact of autocalibration on lift charts that are often used in practice to assess the performances of a candidate premium. To draw such graphs, the data set is sorted based on the values of fitted premiums. The data are then bucketed into equally populated classes based on quantiles. The average predicted and average actual loss costs are then graphed for each bucket. This graph allows the analyst to check whether the actual response monotonically increases as we move to higher buckets. Property 3.5 in [2] identifies the condition required to observe an increasing trend in such a lift chart (that is, positive regression dependence). We know that autocalibration purposes to restore global and local balance but its impact on goodness of lift has not been assessed so far. Intuition suggests that autocalibration should also improve on lift and we confirm this guess on the basis on an analysis of a classical motor insurance data set. The case study performed in this note thus suggests that autocalibration is also beneficial in terms of lift.

The remainder of the text is structured as follows. Section 2 sets up the scene by recalling the ratemaking context. Section 3 presents the concept of autocalibration and the way it can be implemented. The case study in Sect. 4 assesses the effect of autocalibration on goodness of lift using a reference motor insurance database. The final Sect. 5 briefly concludes.

2 Ratemaking problem

Consider a claim-related response Y (frequency, severity or total losses) and a set of features $X_1, \dots, X_p$ gathered in the vector $\boldsymbol{X}$. In actuarial pricing, the aim is to extract the information contained in $\boldsymbol{X}$ about Y in order 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 given the available information $\boldsymbol{X}$. Henceforth, $\mu(\boldsymbol{X})$ is referred to as the true (pure) premium.

The function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x}) = \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ is approximated by a (working, or actual) premium $\boldsymbol{x} \mapsto \pi(\boldsymbol{x})$. Sometimes, the working premium has a relatively simple structure compared to the unknown regression function $\boldsymbol{x} \mapsto \mu(\boldsymbol{x})$. This is the case when actuaries resort to GLMs or GAMs (Generalized Additive Models) which are very transparent but highly constrained. Boosting for instance is more flexible and better captures the structure of the true pure premium, but at the cost of transparency. We refer the reader to [6] for more details about the very aim of ratemaking, which is not to predict the actual losses Y but to create accurate estimates $\widehat{\pi}(\boldsymbol{X})$ of unobserved $\mu(\boldsymbol{X})$.

3 Autocalibration

3.1 Concept

Assume that $\widehat{\pi}$ has been built from some training set (all formulas in the text are meant given this training set). The respective performances of competing models can then be assessed on the basis of a validation set $\{(Y_i, \boldsymbol{X}_i), i = 1, 2, \dots, n\}$, that has not been used to obtain $\widehat{\pi}$. Let $(Y, \boldsymbol{X})$ be a generic random vector distributed as the observations contained in the training set. 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 $\boldsymbol{X}$. To ease the exposition, we assume that the candidate premium $\widehat{\pi}(\boldsymbol{X})$ 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 $\boldsymbol{X}$.

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. This naturally leads to the concept of 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 [4] for a general presentation of this concept. Autocalibration ensures that the average response in a neighborhood defined with the help of the autocalibrated predictor under consideration matches premium income. This exactly corresponds to the method of marginal totals (except that balance is imposed in groups defined by proximity with respect to uncorrected premiums, not to features) and is thus desirable for any candidate premium.

3.2 Autocalibrating a given predictor

In [3] proposed an effective method to implement autocalibration in ratemaking studies. The authors assumed that $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ is continuously increasing, which is a reasonable requirement for any candidate premium and can be tested using the tools developed after [1]. Then, they showed that 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 their Property 5.1. Note that the increasingness of the regression function $s \mapsto \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X}) = s]$ is actually not needed in Property 5.1 in [3] since we directly have

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

A local polynomial regression approach allows the actuary to implement autocalibration, that is, to switch from $\widehat{\pi}$ to $\widehat{\pi}_{\mathrm{BC}}$. See e.g. [5] for a detailed account. To this

end, the only feature entering the analysis is the premium $\widehat{\pi}$ needing autocalibration. A local GLM is fitted to $(Y_i, e_i, \widehat{\pi}(\boldsymbol{x}_i))$ comprised in the validation data set, for some relevant exposure e_i. Let us consider a specific risk profile $\boldsymbol{x}$. Uniform weights are assigned to each $(Y_i, e_i, \widehat{\pi}(\boldsymbol{x}_i))$ in the neighborhood $\mathcal{V}(\boldsymbol{x})$ of $\boldsymbol{x}$. Thus, a rectangular (or uniform) weight function is specified. The neighborhood $\mathcal{V}(\boldsymbol{x})$ gathers the fraction α of the data with the closest premiums $\widehat{\pi}(\boldsymbol{x}_i)$ to $\widehat{\pi}(\boldsymbol{x})$. Here, α acts as a smoothing parameter and controls the size of sub-portfolios where local balance is imposed. The optimal value for the parameter α is selected by likelihood cross-validation.

The local GLM likelihood equation

$$\sum_{i \in \mathcal{V}(\boldsymbol{x})} y_i = \sum_{i \in \mathcal{V}(\boldsymbol{x})} e_i \widehat{\pi}_{\mathrm{BC}}(\boldsymbol{x})$$

transfers part of the experience at neighboring $\widehat{\pi}$ values to obtain $\widehat{\pi}_{\mathrm{BC}}$. A local constant, or intercept-only GLM thus implements local balance in sub-portfolios gathering policyholders with about the same predicted value.

4 Case study

4.1 Database

Let us consider the `freMTPL2freq` data set comprised in the `CASdatasets` package in `R`. It contains 678 013 observations of the number of claims (response Y) in a French motor third-party liability insurance portfolio, together with nine explanatory variables (features $\boldsymbol{X} = (X_1, \ldots, X_9)$). The latter correspond to several characteristics related to the policyholder (age, density of inhabitants in the home city, region, area, bonus-malus) and his or her vehicle (power, age, brand, fuel type). An exposure-to-risk is also available for each policy, here the duration of observation in the portfolio expressed in year. We refer to Noll et al. [7] for an accurate description of the data set.

4.2 Models under consideration

By setting expected claim severities all equal to 1 (thus, adopting the mean severity as monetary unit), we can interpret expected claim frequencies as pure premiums. We consider three candidate premiums corresponding to the following models in our case study: π_{glm} obtained from a Poisson GLM with a log-link function and all explanatory variables, π_{gam} obtained from a Poisson GAM with a log-link function and all explanatory variables, and π_{boost} obtained from a boosted Poisson model. The `h2o` package is used for boosting, precisely the `h2o.gbm` function. These models are fitted to the database under consideration as described on Arthur Charpentier's github page `github.com/freakonometrics/autocalibration`.

To perform our analysis, the database has been divided into three sets. The training set comprising 60% of the data is used to train the "base" model (Poisson GLM, Poisson GAM or boosted Poisson model). The validation (or correcting) set with

20% of the data is used to calibrate the local GLM applied in the auto-calibration procedure. The test (or final) set with the remaining 20% of the data allows us to compare the different models on data that have not been used to train or calibrate the models. All the graphs discussed hereafter are based on the data contained in the test set.

4.3 Autocalibration

The optimal value for α is selected with the help of the likelihood cross-validation, or LCV plot. This is performed with the help of the `lcvplot` function, which calls the `lcv` function for a grid of smoothing parameters α. Those functions are built in the `locfit` package implementing local polynomial regression. We obtain the likelihood cross-validation statistic against the degrees of freedom (DF) of the fit, for each smoothing parameter α. Results are displayed in Fig. 1. The optimum LCV chosen correspond respectively to $\alpha = 4\%$ for the Poisson GLM, $\alpha = 2.5\%$ for the Poisson GAM, and $\alpha = 2.5\%$ for the boosted Poisson model.

4.4 Lift charts

These lift charts are described in [8]. To draw such graphs, the data set is sorted based on the values of $\widehat{\pi}(\boldsymbol{X})$. The data are then bucketed into equally populated classes based on quantiles. Within each bucket, the average expected loss is calculated with the help of $\widehat{\pi}$ as well as the actual loss Y. The average predictions and average costs are then graphed for each class. To make it more readable and interpretable, we normalized the values by dividing them by the average predicted value of the whole test set.

The results for Poisson GLM are displayed in Fig. 2. It is important to realize that bands on each graph depend on the premiums produced by each model and so do not contain the same data. We see that average premiums and average observed values follow the same pattern and nearly coincide. This suggests that autocalibration has little impact on $\widehat{\pi}_{\text{glm}}$. This was expected since Poisson GLMs with log link impose local and global balance constraints. The lifts (value for last band – value for first band) for the models before and after auto-calibration are respectively equal to 1.53 and 1.40.

The results for the Poisson GAM are displayed in Fig. 3. The results are very similar to those obtained with the Poisson GLM and the same comments apply. Again,

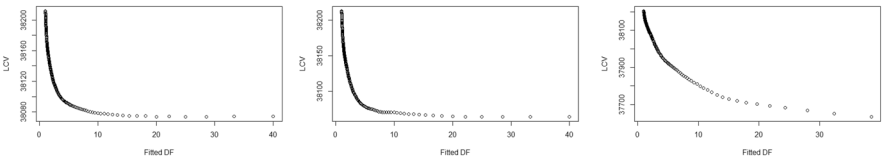

Fig. 1 LCV plots for the Poisson GLM (left panel), for the Poisson GAM (middle panel), and for the boosted Poisson model (right panel)

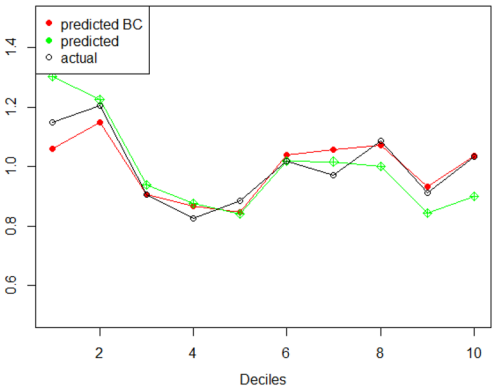

Fig. 2 Lift chart for the Poisson GLM, before (predicted, in green) and after autocalibration (predicted BC, in red)

autocalibration has a very limited impact on $\widehat{\pi}_{\text{gam}}$. The lifts for the models before and after auto-calibration are respectively equal to 1.55 and 1.51.

The results for boosted Poisson regression are displayed in Fig. 4. Marked differences before and after autocalibration are clearly visible there so that autocalibration greatly impacts on $\widehat{\pi}_{\text{boost}}$. Before autocalibration, the boosting model under-estimates on average the observed values. This can be seen as the red curve is below the black curve for all bands. This difference increases when the premiums become higher. Autocalibration corrects for this bias since both the model curve and the observed values curve coincide once it has been implemented. Only for the last bucket we do see a small difference. The lifts for the model before and after autocalibration are equal to 1.31 and 2.20, respectively. Once autocalibration has been implemented, $\widehat{\pi}_{\text{boost}}$ clearly outperforms $\widehat{\pi}_{\text{glm}}$ and $\widehat{\pi}_{\text{gam}}$ in terms of goodness of lift.

4.5 Out-of-sample Poisson deviance

In order to compare the different models, one can also use the out-of-sample deviance. The smaller this deviance, the "better" the model. Figure 5 displays out-of-sample deviance for the Poisson GLM, the Poisson GAM, and the boosted

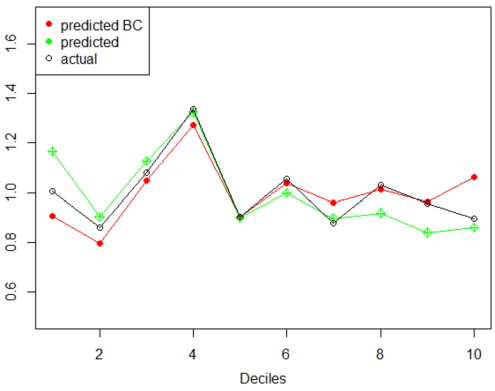

Fig. 3 Lift chart for the Poisson GAM, before (predicted, in green) and after autocalibration (predicted BC, in red)

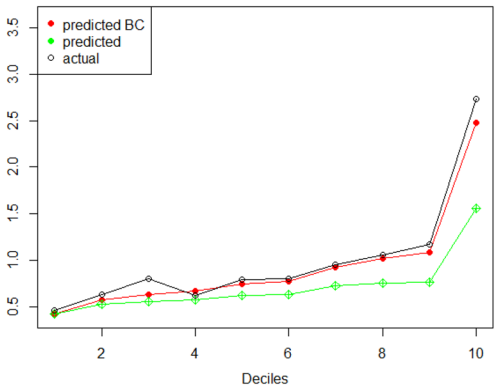

Fig. 4 Lift chart for the boosted Poisson regression model, before (predicted, in green) and after autocalibration (predicted BC, in red)

Poisson regression model, before and after autocalibration. We can see there that GAM does not outperform GLM on the database under consideration, whereas boosting achieves better performances. For the GLM and GAM models, there is no difference in deviance before or after the autocalibration procedure. For the boosted Poisson regression model, out-of-sample deviance is greatly reduced by applying the autocalibration method.

5 Discussion

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 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 often substantially increase overall bias.

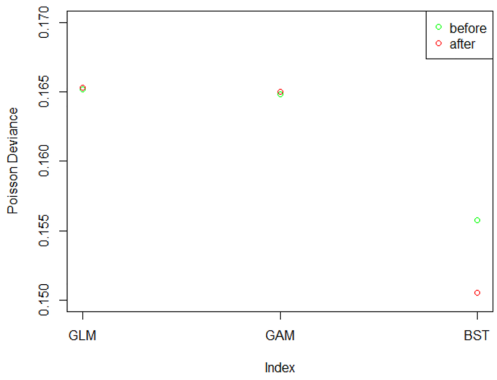

Fig. 5 Predictive Poisson deviance before (in green) and after autocalibration (in red) for the GLM, GAM and boosted regression model

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 $\widehat{\pi}$, reducing the high-dimensional feature space to the real line to compute $\widehat{\pi}_{\mathrm{BC}}$. The learning model is thus 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.

The case study performed in this note shows that autocalibration not only restores global and local balance, but also improves performances measured in terms of lift induced by the candidate premium. This suggests that autocalibration is a key ingredient of the ratemaking process with advanced machine learning tools, reconciling modern tools with old actuarial recipes.

Acknowledgements The authors are grateful to Professor Mario Wüthrich for having pointed out that increasingness of the regression function was not needed in Property 5.1 in [3].

References

1. Bowman AW, Jones MC, Gijbels I (1998) Testing monotonicity of regression. J Comput Graph Stat 7:489–500
2. Denuit M, Sznajder D, Trufin J (2019) Model selection based on Lorenz and concentration curves, Gini indices and convex order. Insur Math Econ 89:128–139
3. Denuit M, Charpentier A, Trufin J (2021) Autocalibration and Tweedie-dominance for insurance pricing with machine learning. Insur Math Econ 101:485–497
4. Kruger F, Ziegel JF (2021) Generic conditions for forecast dominance. J Bus Econ Stat 39:972–983
5. Loader C (1999) Local regression and likelihood. Springer, New York
6. Meyers G, Cummings AD (2009) Goodness of fit vs. goodness of lift. Actuarial Rev 36:16–17
7. Noll A, Salzmann R, Wüthrich M (2018). Case study: French motor third-party liability claims. Available at SSRN: https://ssrn.com/abstract=3164764
8. Tevet D (2013) Exploring model lift: is your model worth implementing. Actuarial Rev 40:10–13
9. Wüthrich MV (2019) From generalized linear models to Neural Networks, and back. SSRN Manuscript ID 3491790
10. Wüthrich MV (2020) Bias regularization in neural network models for general insurance pricing. Eur Actuar J 10:179–202