MODEL SELECTION WITH PEARSON'S CORRELATION, CONCENTRATION AND LORENZ CURVES UNDER AUTOCALIBRATION

Michel Denuit, Julien Trufin

LIDAM Discussion Paper ISBA
2022 / 33

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

Model selection with Pearson's correlation, concentration and Lorenz curves under autocalibration

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

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

November 4, 2022

Abstract

Wüthrich (2022) established that the Gini index is a consistent scoring rule in the class of autocalibrated predictors. This note further explores performances criteria in this class. Elementary Pearson's correlation turns out to be consistent when restricted to autocalibrated predictors. Also, any performance measure that is minimized for predictors that are comonotonic with the true regression model is consistent under autocalibration. This provides a new proof of the consistency for Gini index. In addition, it is established that the concentration curve of the true model is the lowest possible concentration curve under autocalibration and that the same property holds true for Lorenz curve.

Keywords: Gini index, consistent loss function, autocalibration, concentration curve, comonotonicity.

1 Introduction

There are many tools for model selection in machine learning. Let us mention deviance criteria, Gini index, concentration and Lorenz curves or correlation coefficients. While deviance is a consistent scoring rule for the mean, this is not the case in general for the other measures listed before; see Gneiting (2011). Working with these tools for model selection may thus lead to a wrong model choice.

Restricting the analysis to autocalibrated predictors makes several classical performances criteria consistent. It is shown that elementary Pearson's correlation coefficient between the response and any autocalibrated predictor is maximum if, and only if, the model is correct. Wüthrich (2022) shows that restricting the Gini index to the class of autocalibrated regression models makes it a consistent scoring rule. In this note, we generalize this result to any performance measure that is minimized for predictors that are comonotonic with the true regression model. Moreover, we show that the concentration curve of the true model is the lowest possible concentration curve for an autocalibrated predictor.

2 Regression modeling and autocalibrated predictors

In regression modeling, the aim is to extract the information contained in a set of features $X_1, \ldots, X_p$ gathered in the vector $\boldsymbol{X} \in \mathcal{X}$ (classically, $\mathcal{X} \subset \mathbb{R}^p$) about a response Y in order to evaluate as accurately as possible 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 and approximated by a function $\boldsymbol{x} \mapsto \pi(\boldsymbol{x})$ with a simpler structure. 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.

In some applications, it is essential that the sum of predictions $\widehat{\pi}(\boldsymbol{X})$ matches the sum of responses Y as closely as possible. This is the case in insurance ratemaking where $\widehat{\pi}(\boldsymbol{X})$ is supposed to be used as pure premium and Y is the actual loss. The concept of autocalibration is very useful in these situations. A predictor $\widehat{\pi}$ is said to be autocalibrated when $\widehat{\pi}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})]$. Autocalibration thus ensures that the average response in a neighborhood defined with the help of the autocalibrated predictor matches the value of the autocalibrated predictor. We refer the reader to Krüger and Ziegel (2021) for a general presentation of this concept. As mentioned in Wüthrich (2022), this is an important property in insurance pricing since every group of policyholders paying the same premium $\widehat{\pi}(\boldsymbol{X})$ is in average self-financing. The premium $\widehat{\pi}(\boldsymbol{X})$ exactly covers the expected loss Y of that group.

In the sequel, 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 assumption is not restrictive in insurance ratemaking provided $\boldsymbol{X}$ comprises at least one continuous feature and $\widehat{\pi}(\boldsymbol{X})$ is obtained with modern machine learning tools, like boosting, neural networks or random forests, for instance. We denote as $F_{\widehat{\pi}}$ the distribution function of $\widehat{\pi}(\boldsymbol{X})$ and as $F_{\widehat{\pi}}^{-1}$ the associated quantile function. Throughout the paper, we assume that all random variables admit finite first and second moments.

3 Pearson's linear correlation

Correlation coefficients are often used to assess the strength of dependence within the pair $(Y, \widehat{\pi}(\boldsymbol{X}))$, with the understanding that a better model produces stronger dependence. The most elementary correlation coefficient is Pearson's linear one, defined as

$$r(Y, \widehat{\pi}(\boldsymbol{X})) = \frac{\mathrm{Cov}[Y, \widehat{\pi}(\boldsymbol{X})]}{\sqrt{\mathrm{Var}[Y]\mathrm{Var}[\widehat{\pi}(\boldsymbol{X})]}}.$$

The next result shows that maximum correlation identifies the true model under autocalibration.

Proposition 3.1. *Let $\widehat{\pi}(\boldsymbol{X})$ be an autocalibrated predictor. Then, $r(Y, \widehat{\pi}(\boldsymbol{X}))$ is maximum if, and only if, $\widehat{\pi}(\boldsymbol{X}) = \mu(\boldsymbol{X})$.*

Proof. Since

$$\mathrm{Cov}[Y, \widehat{\pi}(\boldsymbol{X})] = \mathrm{Cov}\big[\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})], \widehat{\pi}(\boldsymbol{X})\big] = \mathrm{Var}[\widehat{\pi}(\boldsymbol{X})]$$

when $\widehat{\pi}(\boldsymbol{X})$ is autocalibrated, we get

$$r(Y, \widehat{\pi}(\boldsymbol{X})) = \sqrt{\frac{\mathrm{Var}[\widehat{\pi}(\boldsymbol{X})]}{\mathrm{Var}[Y]}}.$$

Now, we know from Lemma 3.1 in Wüthrich (2022) that any autocalibrated predictor $\widehat{\pi}(\boldsymbol{X})$ satisfies $\widehat{\pi}(\boldsymbol{X}) \preceq_{\mathrm{CX}} \mu(\boldsymbol{X})$ where $\preceq_{\mathrm{CX}}$ is the convex order defined as $\mathrm{E}[\widehat{\pi}(\boldsymbol{X})] = \mathrm{E}[\mu(\boldsymbol{X})]$ and $\mathrm{E}[(\widehat{\pi}(\boldsymbol{X}) - t)_+] \leq \mathrm{E}[(\mu(\boldsymbol{X}) - t)_+]$ for all $t \geq 0$. The stochastic inequality $\widehat{\pi}(\boldsymbol{X}) \preceq_{\mathrm{CX}} \mu(\boldsymbol{X})$ implies $\mathrm{Var}[\widehat{\pi}(\boldsymbol{X})] \leq \mathrm{Var}[\mu(\boldsymbol{X})]$. Also,

$$\mathrm{Cov}[Y, \mu(\boldsymbol{X})] = \mathrm{Cov}\big[\mathrm{E}[Y|\boldsymbol{X}], \mu(\boldsymbol{X})\big] = \mathrm{Var}[\mu(\boldsymbol{X})] \text{ and } r(Y, \mu(\boldsymbol{X})) = \sqrt{\frac{\mathrm{Var}[\mu(\boldsymbol{X})]}{\mathrm{Var}[Y]}}.$$

Therefore,

$$r(Y, \widehat{\pi}(\boldsymbol{X})) \leq r(Y, \mu(\boldsymbol{X}))$$

for any autocalibrated predictor $\widehat{\pi}(\boldsymbol{X})$ and equality implies that $\widehat{\pi}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ are identically distributed (since two random variables ordered in the sense of $\preceq_{\mathrm{CX}}$ and having the same variances must have the same distribution). The announced statement then follows as in the last part of the proof of Lemma 3.1 in Wüthrich (2022), noticing that $\mathrm{E}[g(\widehat{\pi}(\boldsymbol{X}))] = \mathrm{E}[g(\mu(\boldsymbol{X}))]$ holds for every convex function g, provided the expectations exist (the reasoning there starts with the equality of these expectations). □

4 Bregman loss functions revisited

The objective function used for model training is often called loss function in machine learning. Conditional expectations 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 $L(\cdot,\cdot)$ 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

$$L(y,m) = \ell(y) - \ell(m) - \ell'(m)(y-m) \text{ for a convex function } \ell. \tag{4.1}$$

The loss function $L(\cdot,\cdot)$ in (4.1) is also referred to as q-loss function after Efron (1986). Savage (1971) showed, subject to weak regularity conditions, that the class of Bregman loss functions (4.1) is consistent for the (conditional) mean functional. This result is useful in insurance studies since with ℓ defined as $\ell(m) = 2\big(m\theta_m - a(\theta_m)\big)$ where θ_m is the solution of $a'(\theta_m) = m$, (4.1) corresponds to Tweedie deviance with cumulant function $a(\cdot)$.

Restricting the analysis to autocalibrated predictors $\widehat{\pi}(\boldsymbol{X})$, we see that

$$\mathrm{E}\big[\ell'\big(\widehat{\pi}(\boldsymbol{X})\big)\big(Y - \widehat{\pi}(\boldsymbol{X})\big)\big] = \mathrm{E}\big[\ell'\big(\widehat{\pi}(\boldsymbol{X})\big)\big(\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})] - \widehat{\pi}(\boldsymbol{X})\big)\big] = 0$$

so that

$$\mathrm{E}[L(Y,\widehat{\pi}(\boldsymbol{X}))] = \mathrm{E}[\ell(Y)] - \mathrm{E}[\ell(\widehat{\pi}(\boldsymbol{X}))].$$

For autocalibrated predictors, Bregman loss thus reduces to the expected difference of the response and predictors mapped to the scale induced by ℓ. Since $\widehat{\pi}(\boldsymbol{X}) \preceq_{\mathrm{CX}} \mu(\boldsymbol{X})$ is known to hold true, we get

$$\mathrm{E}[L(Y,\widehat{\pi}(\boldsymbol{X}))] \geq \mathrm{E}[L(Y,\mu(\boldsymbol{X}))]$$

with equality if, and only if, $\widehat{\pi}(\boldsymbol{X}) = \mu(\boldsymbol{X})$ provided ℓ is strictly convex. We thus recover Savage's (1971) general result by this very simple argument.

Notice that comparing two autocalibrated predictors $\widehat{\pi}_1(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ with Bregman loss functions does not involve the response as

$$\mathrm{E}[L(Y,\widehat{\pi}_2(\boldsymbol{X}))] - \mathrm{E}[L(Y,\widehat{\pi}_1(\boldsymbol{X}))] = \mathrm{E}[\ell(\widehat{\pi}_1(\boldsymbol{X}))] - \mathrm{E}[\ell(\widehat{\pi}_2(\boldsymbol{X}))].$$

The latter identity shows that $\widehat{\pi}_2(\boldsymbol{X})$ outperforms $\widehat{\pi}_1(\boldsymbol{X})$ for every Bregman loss function if, and only if, $\widehat{\pi}_1(\boldsymbol{X}) \preceq_{\mathrm{CX}} \widehat{\pi}_2(\boldsymbol{X})$. This result corresponds to Theorem 3.1 in Krüger and Ziegel (2021) and is directly derived here by a simple reasoning.

5 Consistent performance measures under autocalibration

5.1 Comonotonicity and autocalibration

Recall that two random variables Z_1 and Z_2 are said to be comonotonic if there exist non-decreasing functions g_1 and g_2 and a third random variable Z such that $Z_i = g_i(Z)$ holds for $i \in \{1,2\}$. Two comonotonic predictors $\widehat{\pi}_1(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ thus rank policyholders in the same way, from the cheapest to the most expensive pure premium. See e.g. Denuit et al. (2005) for a presentation of this concept with applications to actuarial science. The next result shows that if the autocalibrated predictor $\widehat{\pi}(\boldsymbol{X})$ and the conditional mean $\mu(\boldsymbol{X})$ are comonotonic then they must coincide.

Proposition 5.1. *Let $\widehat{\pi}(\boldsymbol{X})$ be an autocalibrated predictor. If $\widehat{\pi}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ are comonotonic, then we necessarily have $\widehat{\pi}(\boldsymbol{X}) = \mu(\boldsymbol{X})$.*

Proof. Since $\widehat{\pi}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ are comonotonic, we get $\mathrm{E}[Y|\widehat{\pi}(\boldsymbol{X})] = \mathrm{E}[Y|\mu(\boldsymbol{X})]$, so that

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

This ends the proof. □

Notice that comonotonicity implies that $\mu(\boldsymbol{X})$ is a non-decreasing function of $\widehat{\pi}(\boldsymbol{X})$ and the proof of Lemma 3.1 in Wüthrich (2022) can easily be adapted to provide an alternative proof of Proposition 5.1. Relating comonotonicity to ranking policyholders according to their pure premiums, Proposition 5.1 means that any autocalibrated predictor $\widehat{\pi}(\boldsymbol{X})$ ranking policyholders as the conditional mean $\mu(\boldsymbol{X})$ in fact coincides with the latter.

5.2 Gini index and ICC

The concentration curve (CC) associated with the predictor $\widehat{\pi}(\boldsymbol{X})$ is defined as

$$\alpha \in (0,1) \mapsto \mathrm{CC}[Y, \widehat{\pi}(\boldsymbol{X}); \alpha] = \frac{\mathrm{E}\big[Y \mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)]\big]}{\mathrm{E}[Y]}. \tag{5.1}$$

The integral of the concentration curve (ICC) over the whole interval $[0,1]$ is then given by

$$\begin{aligned}\mathrm{ICC}[Y, \widehat{\pi}(\boldsymbol{X})] &= \int_0^1 \mathrm{CC}[Y, \widehat{\pi}(\boldsymbol{X}); \alpha] \mathrm{d}\alpha \\ &= \frac{\mathrm{E}\big[Y \int_0^1 \mathrm{I}[\widehat{\pi}(\boldsymbol{X}) \leq F_{\widehat{\pi}}^{-1}(\alpha)] \mathrm{d}\alpha\big]}{\mathrm{E}[Y]}.\end{aligned} \tag{5.2}$$

We refer the reader to Denuit et al. (2019) for more details about CC and ICC. As noticed in Wüthrich (2022), CC is also called cumulative accuracy profile (CAP) up to sign switches.

The Gini index can be defined as

$$\mathrm{Gini}[Y, \widehat{\pi}(\boldsymbol{X})] = \frac{\frac{1}{2} - \mathrm{ICC}[Y, \widehat{\pi}(\boldsymbol{X})]}{\frac{1}{2} - \int_0^1 \mathrm{CC}[Y, Y; \alpha] \mathrm{d}\alpha}. \tag{5.3}$$

As noticed by Wüthrich (2022), the denominator in (5.3) does not use $\widehat{\pi}(\boldsymbol{X})$ so that it has no impact on model selection. Maximizing the Gini index thus amounts to minimizing ICC. This means that model selection with Gini index is equivalent to model selection with ICC.

As pointed out by Wüthrich (2022), the Gini index is not in general a consistent scoring rule for the mean, so neither is the ICC. In his Theorem 3.5, Wüthrich (2022) proves that the Gini index gives a consistent scoring rule within the class of autocalibrated regression

functions. We can recover this result as a direct consequence of Proposition 5.1 which applies to ICC and Gini index. As shown in Denuit et al. (2019), ICC can be expressed as

$$\mathrm{ICC}[Y,\widehat{\pi}(\boldsymbol{X})]=\frac{1}{2}-\frac{\mathrm{Cov}[Y,\widehat{\Pi}]}{\mathrm{E}[Y]}, \tag{5.4}$$

where $\widehat{\Pi}=F_{\widehat{\pi}}\big(\widehat{\pi}(\boldsymbol{X})\big)$ is the rank induced by the predictor $\widehat{\pi}(\boldsymbol{X})$. The response Y in (5.4) can be replaced with the true model $\mu(\boldsymbol{X})$ as pointed out in Denuit et al. (2019) since

$$\mathrm{E}[Y\widehat{\Pi}]=\mathrm{E}\Big[\mathrm{E}\left[Y\widehat{\Pi}\big|\boldsymbol{X}\right]\Big]=\mathrm{E}\left[\mathrm{E}[Y|\boldsymbol{X}]\widehat{\Pi}\right]=\mathrm{E}\left[\mu(\boldsymbol{X})\widehat{\Pi}\right].$$

Minimizing the ICC over $\widehat{\pi}$ thus amounts to maximizing $\mathrm{Cov}[\mu(\boldsymbol{X}),\widehat{\Pi}]$ over $\widehat{\pi}$. It is well-known that $\mathrm{Cov}[\mu(\boldsymbol{X}),\widehat{\Pi}]$ is maximized when $\mu(\boldsymbol{X})$ and $\widehat{\Pi}$ are comonotonic. See e.g. Proposition 2.3 in Denuit and Dhaene (2003). Hence, Proposition 5.1 enables to conclude that ICC is a consistent scoring rule under autocalibration.

6 Concentration and Lorenz curves

The concentration curve CC associated with a predictor $\widehat{\pi}(\boldsymbol{X})$ cannot be written as the expectation of a loss function. In that sense, it cannot be considered as a scoring rule (which needs an associated loss function). Two predictors might well be incomparable because their respective CC intersect. The next proposition shows that the lowest possible CC for an autocalibrated model is the one associated with the true model.

Proposition 6.1. *For any autocalibrated predictor $\widehat{\pi}(\boldsymbol{X})$, we have*

$$\mathrm{CC}[Y,\widehat{\pi}(\boldsymbol{X});\alpha]\geq\mathrm{CC}[Y,\mu(\boldsymbol{X});\alpha] \tag{6.1}$$

for any $\alpha\in(0,1)$, with an identity if, and only if, $\widehat{\pi}(\boldsymbol{X})=\mu(\boldsymbol{X})$.

Proof. Given two predictors $\widehat{\pi}_1(\boldsymbol{X})$ and $\widehat{\pi}_2(\boldsymbol{X})$ with the same means, we know from Property 3.4.41 of Denuit et al. (2005) that

$$\widehat{\pi}_1(\boldsymbol{X})\preceq_{\mathrm{CX}}\widehat{\pi}_2(\boldsymbol{X})\Leftrightarrow\mathrm{LC}[\widehat{\pi}_1(\boldsymbol{X});\alpha]\geq\mathrm{LC}[\widehat{\pi}_2(\boldsymbol{X});\alpha],$$

where $\mathrm{LC}[\widehat{\pi}(\boldsymbol{X});\alpha]$ is the Lorenz curve associated with the predictor $\widehat{\pi}(\boldsymbol{X})$ defined as

$$\alpha\in(0,1)\mapsto\mathrm{LC}[\widehat{\pi}(\boldsymbol{X});\alpha]=\frac{\mathrm{E}\big[\widehat{\pi}(\boldsymbol{X})\mathrm{I}[\widehat{\pi}(\boldsymbol{X})\leq F^{-1}_{\widehat{\pi}}(\alpha)]\big]}{\mathrm{E}[\widehat{\pi}(\boldsymbol{X})]}.$$

Moreover, we known from Proposition 3.1 in Denuit and Trufin (2021) that

$$\widehat{\pi}(\boldsymbol{X})\text{ autocalibrated}\Rightarrow\mathrm{LC}[\widehat{\pi}(\boldsymbol{X});\alpha]=\mathrm{CC}[Y,\widehat{\pi}(\boldsymbol{X});\alpha]\text{ for all }\alpha\in(0,1).$$

Therefore, we directly get the announced result by Lemma 3.1 of Wüthrich (2022) which states that $\widehat{\pi}(\boldsymbol{X})\preceq_{\mathrm{CX}}\mu(\boldsymbol{X})$ holds true for any autocalibrated $\widehat{\pi}(\boldsymbol{X})$. If (6.1) is an identity then $\widehat{\pi}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ are identically distributed and the announced result follows as in the proof of Proposition 3.1. □

Notice that the result stated in Proposition 6.1 is stronger than the consistency of the ICC in the class of autocalibrated models. Also, since Lorenz curve coincides with concentration curve for autocalibrated predictors, Proposition 6.1 shows that the true model also has the smallest Lorenz curve. This legitimates model assessment based on Lorenz curve under autocalibration, as it is often performed in practice.

References

- Denuit, M., Dhaene, J. (2003). Simple characterizations of comonotonicity and countermonotonicity by extremal correlations. Belgian Actuarial Bulletin 3, 22-27.

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

- Denuit, M., Trufin, J. (2021). Lorenz curve, Gini coefficient, and Tweedie dominance for autocalibrated predictors. LIDAM Discussion Paper ISBA 2021/36.

- 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.

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

- Krüger, F., Ziegel, J.F. (2021). Generic conditions for forecast dominance. Journal of Business & Economic Statistics 39/4, 972-983.

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

- Wüthrich, M.V. (2022). Model selection with Gini indices under auto-calibration. European Actuarial Journal, in press.