Tests for Auto-Calibration

Mario V. Wüthrich
Department of Mathematics
ETH zürich

Insurance Data Science Conference
Bayes Business School, London

June 20, 2025

AI Tools for Actuaries (work in progress)

Authors: Mario V. Wüthrich, Ronald Richman, Benjamin Avanzi, Mathias Lindholm, Marco Maggi, Michael Mayer, Jürg Schelldorfer, Salvatore Scognamiglio

About This Project

This project aims to empower the actuarial profession with modern machine learning and AI tools. We provide comprehensive teaching materials that consist of lecture notes (technical document) building the theoretical foundation of this initiative. Each chapter of these lecture notes is supported by notebooks and slides which give teaching material, practical guidance and applied examples. Moreover, hands-on exercises in both R and Python are provided in additional notebooks.

Lecture Notes (Technical Document)

Lecture Notes

Notebooks, Slides and Code

Chapter 1: Introduction and Preliminaries

Notebook PDF Slides

https://aitools4actuaries.com/

- **Section 1: The auto-calibration property**

Regression modeling

- **Actuarial pricing.** Find the (unknown) regression function $\boldsymbol{X} \mapsto \mu(\boldsymbol{X})$ that describes the conditionally expected claim

$$\mu(\boldsymbol{X}) = \mathbb{E}\left[Y \middle| \boldsymbol{X}\right],$$

  where $\boldsymbol{X}$ are the covariates (features) characterizing the claim (response) Y.

- **Solution.** For an i.i.d. learning sample $\mathcal{L} = (Y_i, \boldsymbol{X}_i)_{i=1}^n$, select the best candidate

$$\boldsymbol{X} \ \mapsto \ \widehat{\mu}_{\mathcal{L}}(\boldsymbol{X}) = \widehat{\mathbb{E}}\left[Y \middle| \boldsymbol{X}\right].$$

  from a pre-selected class $\mathcal{M} = \{\mu\}$ of candidate regression models.

- **Question.** Is the selected regression model $\widehat{\mu}_{\mathcal{L}}$ a suitable choice?
  This depends on:

(a) the selected model class $\mathcal{M} = \{\mu\}$,
(b) the observed data $\mathcal{L} = (Y_i, \boldsymbol{X}_i)_{i=1}^n$, and
(c) the model selection procedure.

Global unbiasedness

- **Global unbiasedness** of the estimated model (out-of-sample evaluation)

$$\mathbb{E}\left[\widehat{\mu}_{\mathcal{L}}(\boldsymbol{X})\right] = \mathbb{E}\left[Y\right].$$

  - ⋆ Charged insurance premium on average covers the expected claim.
  - ⋆ Difficult to verify because the true data generating model is unknown.

- A regression function $\boldsymbol{X} \mapsto \widehat{\mu}_{\mathcal{L}}(\boldsymbol{X})$ selection procedure satisfies **the balance property** if for a.e. realisation of $\mathcal{L} = (Y_i, \boldsymbol{X}_i)_{i=1}^n$

$$\sum_{i=1}^{n} \widehat{\mu}_{\mathcal{L}}(\boldsymbol{X}_i) = \sum_{i=1}^{n} Y_i.$$

  - ⋆ The balance property is an in-sample property that reflects a claims re-allocation.
  - ⋆ For the balance property, see Bühlmann–Gisler (2005) and Lindholm–W. (2024).

Local unbiasedness: auto-calibration

- A regression function $\boldsymbol{X} \mapsto \mu(\boldsymbol{X})$ is **auto-calibrated** for $(Y, \boldsymbol{X})$ if, a.s.,

$$\mu(\boldsymbol{X}) = \mathbb{E}\left[Y \middle| \mu(\boldsymbol{X})\right].$$

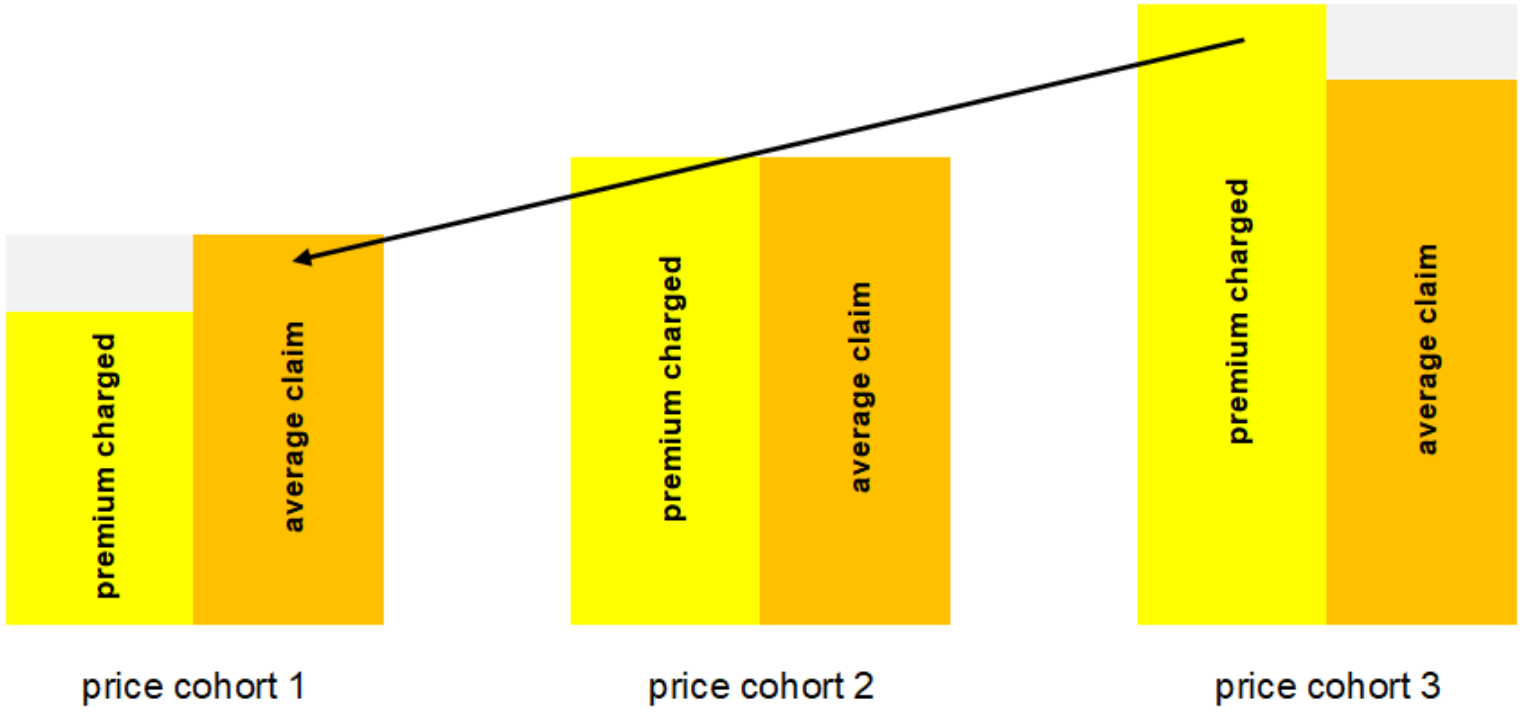
- Auto-calibration means that every price cohort $\mu(\boldsymbol{X})$ is on average self-financing for its claim Y, and there is no systematic cross-financing within the insurance portfolio.

- Auto-calibration was introduced by Schervish (1989) in the statistical literature, and it has been popularized by Gneiting–Resin (2023), Krüger–Ziegel (2021) and Denuit et al. (2021).

- Finding powerful tests for auto-calibration is an active field of research: Hosmer–Lemeshow (1980), Gneiting–Resin (2023), Dimitriadis et al. (2023), Lindholm et al. (2023), Denuit et al. (2024), Delong–W. (2024) and Delong et al. (2025).

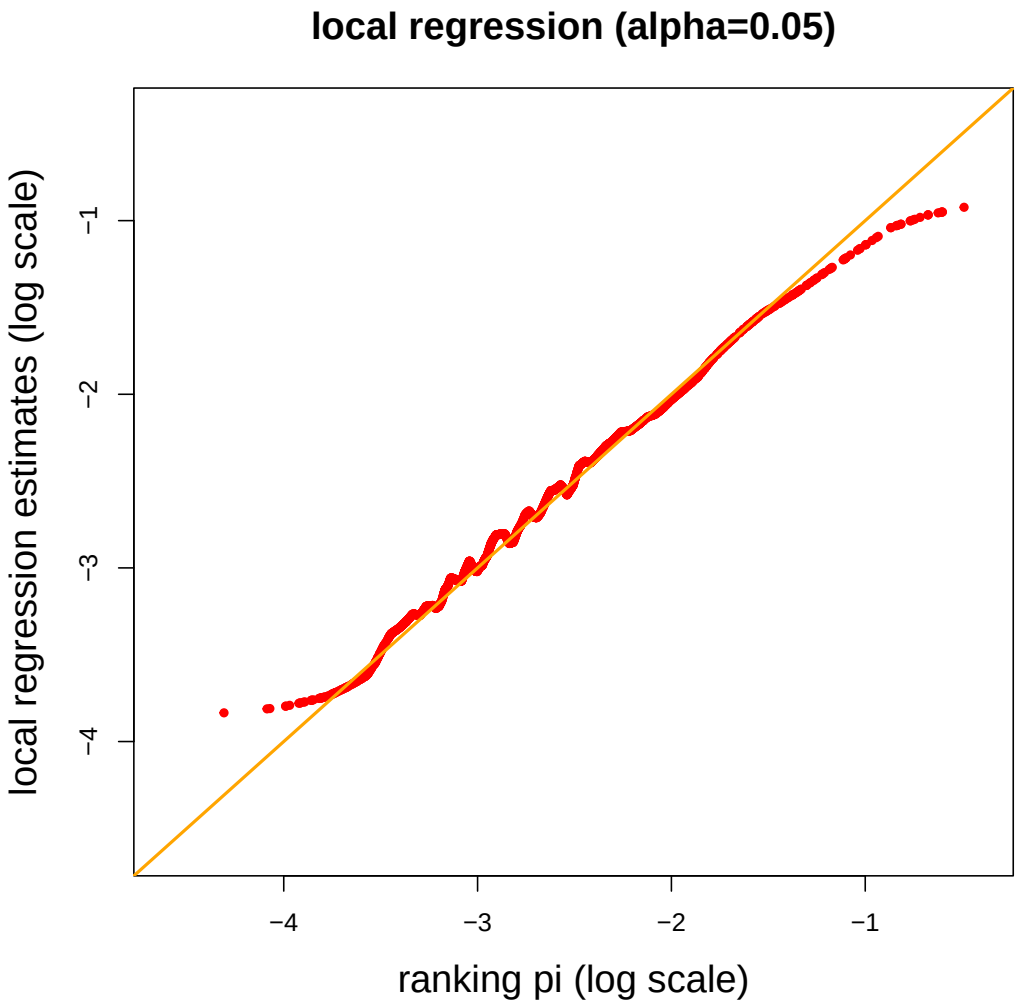
- We present additional insight on the results of Denuit et al. (2024); see W. (2025).

Violation of auto-calibration



Price cohort 1 is subsidized by price cohort 3.

French MTPL data: network regression model



Auto-calibration violation at the boundaries. What about the general fluctuations?

- **Section 2: Tests for auto-calibration**

Finite discrete regression functions

- Assume the selected regression function $\mu : \mathcal{X} \to \mathbb{R}$ only takes finitely many (ordered) values $-\infty < \mu_1 < \cdots < \mu_K < \infty$.

- This partitions the covariate space $\mathcal{X}$ into K different sets with

$$\mathbb{P}\left[\mu(\boldsymbol{X}) = \mu_k\right] = p_k > 0 \qquad \text{for all } 1 \leq k \leq K.$$

- In this finite partition case, auto-calibration of $\mu(\cdot)$ for $(Y, \boldsymbol{X})$ reads as

$$\mu_k = \mathbb{E}\left[Y \middle| \, \mu(\boldsymbol{X}) = \mu_k\right] \qquad \text{for all } 1 \leq k \leq K.$$

Test statistics

- For a given i.i.d. test sample $\mathcal{T} = (Y_i, \boldsymbol{X}_i)_{i=1}^n$, consider the test statistics

$$S_n^{(k)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(\boldsymbol{X}_i)) \, \mathbb{1}_{\{\mu(\boldsymbol{X}_i) = \mu_k\}} \qquad \text{for } 1 \leq k \leq K.$$

- **Proposition.** Under auto-calibration of $\mu(\cdot)$ for $(Y, \boldsymbol{X})$

$$\sqrt{n} \left(S_n^{(1)}, \ldots, S_n^{(K)} \right)^\top \Longrightarrow \mathcal{N}\left(0, \operatorname{diag}\left(p_k \tau_k^2\right)_{k=1}^K\right) \qquad \text{as } n \to \infty,$$

with conditional variances $\tau_k^2 = \operatorname{Var}\left(Y \middle| \, \mu(\boldsymbol{X}) = \mu_k\right)$ for $1 \leq k \leq K$.

Auto-calibration tests

- **Test 1.** Under the null hypothesis of $\mu(\cdot)$ being auto-calibrated for $(Y, \boldsymbol{X})$, we have for $s > 0$ and n large

$$\mathbb{P}\left[\max_{1\leq k\leq K} \sqrt{n}|S_n^{(k)}| \leq s\right] \approx \prod_{k=1}^{K}\left(2\Phi\left(\frac{s}{\sqrt{p_k}\tau_k}\right) - 1\right).$$

- Often, it is beneficial to test for the maximum of the normalized quantities

$$\mathbb{P}\left[\max_{1\leq k\leq K} \sqrt{n}\,\frac{|S_n^{(k)}|}{\sqrt{p_k}\tau_k} \leq s\right] \approx \left(2\Phi\left(s\right) - 1\right)^K.$$

Test statistics

- Consider the aggregate (random walk) version, for $1 \leq k \leq K$,

$$T_n^{(k)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(\boldsymbol{X}_i)) \, \mathbb{1}_{\{\mu(\boldsymbol{X}_i) \leq \mu_k\}} = \sum_{j=1}^{k} S_n^{(j)}.$$

- **Corollary.** Under auto-calibration of $\mu(\cdot)$ for $(Y, \boldsymbol{X})$

$$\sqrt{n} \left(T_n^{(1)}, \ldots, T_n^{(K)} \right)^\top \Longrightarrow \mathcal{N} \left(0, \left(\sum\nolimits_{j=1}^{k \wedge m} p_j \tau_j^2 \right)_{1 \leq k,m \leq K} \right) \quad \text{as } n \to \infty.$$

Auto-calibration tests

- **Test 2.** Under the null hypothesis of $\mu(\cdot)$ being auto-calibrated for $(Y, \boldsymbol{X})$, we have for $s > 0$ and n large

$$\mathbb{P}\left[\max_{1\leq k\leq K} \sqrt{n}|T_n^{(k)}| \leq s\right] \approx \mathbb{P}\left[\max_{1\leq k\leq K} |Z_k| \leq s\right],$$

  with random walk

$$Z_k = \sum_{j=1}^{k} \sqrt{p_j}\, \tau_j\, \varepsilon_j,$$

  for i.i.d. standard Gaussian innovations $\varepsilon_j \sim \mathcal{N}(0,1)$ for $1 \leq j \leq K$.

- Essentially, this reflects the finite regression version of Proposition 3.1 of Denuit et al. (2024). In that reference, the authors have not been able to fully identify the limiting distribution of the test statistics and a non-parametric Monte Carlo simulation was proposed. From our results, it becomes clear that this test statistics studies the maximum absolute value of a Brownian motion.

References

[1] Bühlmann, H., Gisler, A. (2005). *A Course in Credibility Theory and its Applications*. Springer.

[2] Delong, Ł, Gatti, S., Wüthrich, M.V. (2025). Calibration bands for mean estimates within the exponential dispersion family. *arXiv*:2503.18896, 2025.

[3] Delong, Ł, Wüthrich, M.V. (2024). Isotonic regression for variance estimation and its role in mean estimation and model validation. *North American Actuarial Journal*, online.

[4] Denuit, M., Charpentier, A., Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing in machine learning. *Insurance: Mathematics and Economics* **101/B**, 485-497.

[5] Denuit, M., Huyghe, J., Trufin, J., Verdebout, T. (2024). Testing for auto-calibration with Lorenz and concentration curves. *Insurance: Mathematics and Economics* **117**, 130-139.

[6] Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., Ziegel, J. (2023). Honest calibration assessment for binary outcome predictions. *Biometrika* **110/3**, 663-680.

[7] Gneiting, T., Resin, J. (2023). Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics* **17**, 3226-3286.

[8] Hosmer, D.W., Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9**, 1043-1069.

[9] Krüger, F., Ziegel, J.F. (2021). Generic conditions for forecast dominance. *Journal of Business and Economics Statistics* **39/4**, 972-983.

[10] 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.

[11] Lindholm, M., Wüthrich, M.V. (2024). The balance property in insurance pricing. *SSRN Manuscript* ID 4925165.

[12] Schervish, M.J. (1989). A general method of comparing probability assessors. *The Annals of Statistics* **17/4**, 1856-1879.

[13] Wüthrich, M.V. (2025). Auto-calibration tests for discrete finite regression functions. *European Actuarial Journal* **15/1**, 335-341 .