Procédure de test pour l'auto-calibration basée sur les courbes de Lorenz et de concentration

Huyghe Julie[1] & Julien Trufin [2] & Thomas Verdebout [3]

[1] *Department of Mathematics Université Libre de Bruxelles (ULB) Brussels, Belgium - julie.huyghe@ulb.be*

[2] *Department of Mathematics Université Libre de Bruxelles (ULB) Brussels, Belgium - julien.trufin@ulb.be*

[2] *Department of Mathematics Université Libre de Bruxelles (ULB) Brussels, Belgium - thomas.verdebout@ulb.be*

1 Courbes de Lorentz, de concentration et auto-calibration.

Nous considérons dans ce travail un contexte de régression classique dans lequel on a (i) une variable aléatoire **réponse** Y à valeurs réelles et (ii) un ensemble de **caractéristiques** $X_1, \ldots, X_p$ regroupées dans un vecteur aléatoire $\mathbf{X}$ à valeurs dans $\mathbb{R}^p$. La structure de dépendance à l'intérieur du vecteur aléatoire $(Y, X_1, \ldots, X_p)$ est ici exploitée pour extraire l'information contenue dans $\mathbf{X}$ à propos de Y. Soit

$$\mu(\mathbf{X}) = \mathrm{E}[Y|\mathbf{X}]$$

l'espérance conditionnelle de Y sachant $\mathbf{X}$ et soit

$$\begin{aligned} \hat{m} : \mathbb{R}^p &\to \mathbb{R} \\ \mathbf{x} &\to \hat{m}(\mathbf{x}) \end{aligned}$$

un estimateur de $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$. On peut associer à cet estimateur $\hat{m}$, deux courbes qui seront importantes ici:

- la **courbe de concentration** de $\mu(\boldsymbol{X})$ par rapport à $\widehat{m}(\mathbf{X})$ définie par

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

  où $F^{-1}_{\widehat{m}}$ est la fonction quantile associée à $\widehat{m}(\mathbf{X})$ et

- la **courbe de Lorenz** associée à $\widehat{m}(\boldsymbol{X})$ définie par

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

En assurance, L'estimateur $\hat{m}$ de la fonction de régression a pour objectif d'être utilisée comme un prime et il est donc naturel que la somme des primes correspondent le plus possible a la somme des pertes réelles. Ceci conduit naturellement au concept d'autocalibration. L'estimateur $\hat{m}$ est dit auto-calibré si et seulement si

$$\mathrm{E}[Y|\hat{m}(\mathbf{X}) = m] = m$$

pour tout $m \in \mathbb{R}$. Comme nous allons le décrire dans la prochaine Section, les courbes de concentration et de Lorenz vont être utilisée pour tester l'auto-calibration. Plus précisément, nous fournissons ci-dessous un test fondé sur la comparaison de fonctions aléatoires.

1.1 Procédure de test pour l'auto-calibration

Nous avons que

$$\mathrm{CC}[\mu(\boldsymbol{X}), \widehat{m}(\boldsymbol{X}); \alpha] = \mathrm{LC}[\widehat{m}(\boldsymbol{X}); \alpha] \text{pour tout } \alpha$$

si et seulement si la version non-biaisée de $\widehat{m}$ donnée par $\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X}) := \frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]}\widehat{m}(\boldsymbol{X})$ est auto-calibrée. Nous utilisons ci-dessous cette caractérisation en comparant des version empirique des courbes de concentration et de Lorenz. Cette caractérisation de l'auto-calibration via les courbes de performance nous permet de proposer une procédure de test pour l'auto-calibration à partir de n copies i.i.d. $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ de $(Y, \boldsymbol{X})$. Plus précisément, nous souhaitons tester l'hypothèse nulle :

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

contre l'hypothèse alternative :

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

La procédure de test est donc naturellement basée sur la différence entre les versions empiriques des courbes de Lorenz et de concentration, définies comme suit :

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

et

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

Plus précisément, l'hypothèse nulle est rejetée pour de grandes valeurs de la statistique de test suivante :

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

où

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

On peut ensuite dériver le comportement assymptotique du processus $T_n(\alpha)$. En supposant que $(Y_i, \widehat{m}(\boldsymbol{X}_i))$, $i = 1, 2, \ldots, n$, sont tels que : $\mathrm{E}[Y_i] \neq 0$, $\mathrm{E}[\widehat{m}(\boldsymbol{X}_i)] \neq 0$, $\mathrm{E}[\widehat{m}^2(\boldsymbol{X}_i)] < \infty$ et $\mathrm{E}[Y_i^2] < \infty$. Alors sous l'hypothèse nulle, $T_n(\alpha)$ converge vers un processus gaussien de moyenne zero et de covariance

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

avec (en supposant que toutes ces quantités existent et sont finies) :

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

$\Sigma(\alpha_1, \alpha_2)$ la matrice de covariance du vecteur aléatoire

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

pour $(\alpha_1, \alpha_2) \in (0, 1)^2$, et enfin

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

où

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

La procédure de test rejette donc l'hypothèse nulle si $\mathcal{T} > c_\beta$ où β est un niveau de confiance fixé tel que $\mathrm{P}[\sup_{\alpha \in (0,1)} |T_n(\alpha)| > c_\beta] = \beta$ sous l'hypothèse nulle. On ne peut pas calculer c_β car la distribution de $(Y_i, \widehat{m}(\boldsymbol{X}_i))$ reste inconnue. Nous utilisons néanmoins ce résultat pour fournir une procédure bootstrap.

Bibliographie

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

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

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.

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.