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*

Résumé. Dans le prolongement des travaux de Frees et al. (2011), Denuit et al. (2019, 2021a, 2021b) ont introduit des relations de dominance et des outils de diagnostic basés sur les courbes de Lorenz et de Concentration afin de comparer différents estimateurs pour la fonction de régression. Nous montrons que cette approche est équivalente à la notion de dominance telle que définie par Ehm et al. (2016) lorsque les estimateurs considérés sont auto-calibrés (Kruger et Ziegel, 2021). Nous établissons également une nouvelle caractérisation de l'auto-calibration en utilisant les graphiques des courbes de Lorenz et de Concentration. Enfin, nous exploitons ce résultat pour proposer une procédure de test efficace pour l'auto-calibration.

Mots-clés. Courbe de Concentration, Courbe de Lorenz, Courbe de Concentration Intégrée, Aire Entre les Courbes, Estimateurs auto-calibrés.

Abstract. Following Frees et al. (2011), Denuit et al. (2019, 2021a, 2021b) proposed dominance relations and diagnostic tools based on Lorenz and Concentration curves in order to compare competing estimators of the regression function. This approach turns out to be equivalent to forecast dominance as defined by Ehm et al. (2016) when the estimators under consideration are auto-calibrated (Kruger and Ziegel, 2021). 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.

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

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

Cette présentation est divisée en deux parties. La première partie porte sur les courbes de concentration et de Lorenz d'un estimateur, ainsi que sur la notion d'auto-calibration et son utilité dans l'évaluation des performances des estimateurs de la moyenne conditionnelle. Dans la deuxième partie, nous nous intéressons spécifiquement à la classe des estimateurs auto-calibrés et aux fonctions de perte strictement consistantes pour cette classe.

L'estimation de la fonction moyenne est une tâche fondamentale en statistiques et en apprentissage automatique, où nous cherchons à estimer l'espérance conditionnelle d'une variable dépendante en fonction de variables explicatives. Traditionnellement, nous utilisons des fonctions de perte pour minimiser les erreurs de prédiction sur un ensemble de données d'entraînement.

En nous appuyant sur les travaux de Frees et al. (2011), nous introduisons une approche novatrice basée sur les courbes de Lorenz et de concentration pour comparer les mérites relatifs des estimateurs. Ces courbes offrent une perspective différente de la comparaison des estimateurs par rapport aux critères classiques d'ajustement.

Soit $F_{\widehat{m}}^{-1}$ la fonction quantile de $\widehat{m}$. La courbe de concentration de $\mu(\boldsymbol{X})$ par rapport à $\widehat{m}$, basée sur l'information contenue dans le vecteur $\boldsymbol{X}$, est définie par la fonction :

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

La courbe de Lorenz associée à $\widehat{m}(\boldsymbol{X})$ est définie comme la fonction :

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

Pour une revue des propriétés des courbes de concentration et de Lorenz, nous renvoyons le lecteur au livre de Yitzhaki and Schechtman (2013).

Un estimateur $\widehat{m}$ est dit auto-calibré si $\widehat{m}(\boldsymbol{X}) = \mathrm{E}[Y|\widehat{m}(\boldsymbol{X})]$. Nous montrons que :

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

si et seulement si $\widehat{m}_{\mathrm{unbiased}}(\boldsymbol{X}) = \frac{\mathrm{E}[Y]}{\mathrm{E}[\widehat{m}(\boldsymbol{X})]}\widehat{m}(\boldsymbol{X})$ est auto-calibré.

Cette caractérisation de l'auto-calibration nous permet de proposer une procédure de test efficace 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 :

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

et

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

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

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

La classe des estimateurs auto-calibrés est intéressante à étudier car certaines mesures de performances classiques en statistiques deviennent des mesures strictement consistantes pour la moyenne. Si l'on choisit une fonction de score $S : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ nous fournissant le score $\mathrm{E}[S(Y, \hat{m}(X))]$ pour toute fonction de régression $X \to \hat{m}(X)$, alors une fonction de score est dite strictement consistante pour la fonction de moyenne conditionnelle T si, pour toute fonction de régression $X \to \hat{m}(X)$, la condition suivante est satisfaite :

$$\mathrm{E}[S(Y, T(F))] \geq \mathrm{E}[S(Y, \hat{m}(X))]$$

et l'égalité tient si et seulement si $\hat{m}(X) = \mu^\star(X)$ presque sûrement. La stricte consistance implique donc que la maximisation du score nous conduit à la véritable fonction de régression $\mu^\star(X)$. En particulier, nous présenterons les résultats suivants :

- L'indice de Gini : $Gini[Y, \hat{\pi}(X)] = \dfrac{\frac{1}{2} - ICC[Y, \hat{m}(X)]}{\frac{1}{2} - \int_0^1 CC[Y, Y; \alpha] d\alpha}$ est une fonction de score strictement consistante pour la moyenne conditionnelle.
- L'ICC (Courbe de concentration intégrée) est une fonction de score strictement consistante pour la moyenne conditionnelle.
- Le coefficient de corrélation linéaire de Pearson est une fonction de score strictement consistante pour la moyenne conditionnelle.

- Si $\hat{m}(X)$ est un prédicteur auto-calibré et que $\hat{m}(X)$ et $\mu(X)$ sont comonotones, alors nécessairement $\hat{m}(X) = \mu(X)$.
- Pour tout prédicteur auto-calibré $\hat{m}(X)$, $CC[Y, \hat{m}(X); \alpha] \geq CC[Y, \mu(X), \alpha]$ pour tout $\alpha \in (0, 1)$, avec égalité si, et seulement si, $\hat{m}(X) = \mu(X)$.

En conclusion, notre étude des estimateurs auto-calibrés a révélé des propriétés intéressantes en matière de mesures de performances et de stricte consistance pour la moyenne conditionnelle.

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.