ONE-STEP SMOOTHING SPINES INSTRUMENTAL VARIABLES REGRESSION
PAPIER DE RECHERCHE /STATISTIQUE

TUTEUR ELIA LAPENTA / CREST
OUMAR DIONE / ENSAE PARIS

[1]Universidade Tecnológica Federal do Paraná, CREST, Paris
[2]ENSAE Paris, Palaiseau

Inserir Logo: Evento

Inserir Logo: Org. Promotora

Inserir Logo: Inst. Parceira

ID: EVNT2024-0001

INTRODUCTION

Cette étude porte une nouvelle extension des Smoothing Splines de régression non paramétriques dans le contexte où les régresseurs sont endogénes et des variables instrumentales sont disponibles.

L'ÉQUATION DU MODÉLE

l'équation du modèle s'écrit comme suit .

$$Y = g_0(Z) + \epsilon, \quad \mathbb{E}[\epsilon \mid W] = 0$$

Dans cette équation, Y est la variable dépendante, Z est le vecteur des régresseurs endogènes, W représente les instruments, et ϵ, non corrélé à W, est le bruit ajouté au modèle.

CONDITION D'INDETIFICATION

Pour assurer l'unicité de la solution, sous réserve de son existence, la condition d'identification suivante est définie :

$$\mathbb{E}[Y - g(Z) \mid W] = 0 \quad \Longrightarrow \quad g = g_0 \quad \text{p.s.}$$

Cette condition est également connue sous le nom de "Completeness".
L'estimation de l'espérance conditionnelle étant plus complexes , les auteurs optent pour une approchent alternative en utilisant un résultat de Bierens(1982) selon lequel

$$\mathbb{E}[Y - g(Y) \mid W] = 0 \quad \Longleftrightarrow \quad \mathbb{E}[(Y - g(Y)) \exp(i(W)^T \mathbf{t})] = 0, \quad \forall\, \mathbf{t} \in \mathbb{R}^p$$

où i est le nombre complexe imaginaire

FERRAMENTAS PARA GERAR OU EDITAR ENTRADAS BIBTEX

- ZoteroBib.
- BibTEX Editor.

DÉFINITION DU PROBLÉME D'OPTIMISATION

Les auteurs considèrent alors g_0 défini par

$$g_0 = \arg\min_{g \in \mathcal{G}} M_n(g),$$

et montrent que la condition

$$M_n(g) = 0 \quad \Longleftrightarrow \quad g = g_0 \quad \text{p.s.}$$

Pour estimer g_0, ils adoptent un estimateur naturel de $M(g)$, noté $M_n(g)$, dans lequel l'espérance est remplacée par une moyenne empirique. Des calculs simples ont montré que

$$M_n(g) = (1/n) \sum_{i,j} (Y_i - g(Z_i)(Y_j - g(Z_j) w(W_i - X_j).$$

OPTIMISATION

Cependant, minimiser directement $M_n(g)$ peut conduire à un problème mal posé, notamment en raison de l'absence de solution ou d'une interpolation.
Pour contourner cette difficulté, les auteurs minimisent une version pénalisée de $M_n(g)$, définissant ainsi l'objectif comme étant $S_n(g)$. Cette fonction est donnée par

$$S_n(g) = M_n(g) + \lambda \int \|g''(x)\|^2 \, dx,$$

où λ est un paramètre de pénalisation, $g''(x)$ représente la dérivée seconde de g, et $\|\cdot\|$ est la norme au carré.

CLOSED FORM SOLUTION

Grâce au théorème de Green et Silverman, le choix de la pénalisation – basé sur le carré de la norme de la dérivée seconde de g – implique qu'une solution doit être cherchée parmi les splines cubiques naturelles. Cela signifie qu'une optimisation sur l'ensemble $\mathcal{G}$ est équivalente à une optimisation sur un ensemble de fonctions paramétrées par $n + 2$ paramètres.
Par la suite, une version matricielle de S_n est obtenue, et elle s'écrit :

$$S_n(g) = \left(Y - g(Y) - \hat{E}\right)^\top \Omega \left(Y - g(Y) - \hat{E}\right),$$

où $\hat{E}$ est un terme d'ajustement et Ω est une matrice de pondération.
L'optimisation consiste alors à minimiser une fonction quadratique. La solution est ainsi définie, sous réserve d'existence, par les équations normales, c'est-à-dire que les coefficients optimaux s'obtiennent en résolvant le système linéaire

$$\begin{pmatrix} \tilde{E} & Z \\ Z^\top & 0 \end{pmatrix} \begin{pmatrix} \hat{\delta} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ 0 \end{pmatrix}$$

CONCLUSION

En résolvant ce système d'équations linéaires, nous obtenons les coefficients $\hat{a}$ et $\hat{\delta}$ qui caractérisent notre estimateur – un spline cubique naturelle – défini par

$$\hat{g}(Z) = Z\hat{a} + \hat{\delta}E,$$

où $\hat{g}(Z)$ est la fonction spline cubique naturelle, Z est le vecteur des régresseurs, $\hat{a}$ est le vecteur des coefficients estimés, et $\hat{\delta}$ est le vecteur d'ajustement.
La relation précédente fournit également un calcul direct de la dérivée première de $\hat{g}$. On montre que

$$\hat{g}'(Z) = O\hat{a} + D\hat{\delta},$$

où O et D sont des matrices qui dépendent des dérivées premières des splines cubiques naturelles.

- Primeiro item de conclusão.
- Segundo item de conclusão.
- Terceiro item de conclusão.

CROSS-VALIDATION

Le but de la validation croisée est de trouver la meilleure valeur de λ qui permet un équilibre entre un modèle qui sur-apprend (overfitting) et un modèle trop simpliste (underfitting). Ensuite, cette valeur sera utilisée pour définir notre estimateur final.
Pour cela, on parle d'une solution "closed-form"qui définit l'estimateur pour chaque valeur de λ. Premièrement, considérons $B_1 = \{(Y_i, Z_i, W_i) \mid 1 \leq i \leq n/2\}$. Dans le cas trivial, $\lambda = 0$, on obtient un modèle trivial (overfitting), et pour des valeurs de λ faibles, on obtient un modèle très simple (underfitting).
Pour mettre en œuvre la validation croisée, on commence, après avoir défini une grille de valeurs de λ, par diviser aléatoirement le jeu de données en deux "folds". Soient sans perte de généralité $B_1 = \{(Y_i, Z_i, W_i) \mid 1 \leq i \leq n/2\}$ et $B_2 = \{(Y_i, Z_i, W_i) \mid n/2 + 1 \leq i \leq n\}$.
Deuxièmement, pour chaque valeur de λ_j, avec $1 \leq j \leq n_\lambda$, nous calculons tour à tour notre estimateur sous forme close sur chaque fold et faisons des prédictions sur l'autre fold. Ensuite, nous regroupons ces valeurs estimées dans un vecteur $\tilde{g}_{\lambda_j}$, qui sera utilisé pour attribuer un score à ce λ_j à travers la fonction M_n.
Le score $s(\lambda_j)$ est défini par :

$$s(\lambda_j) = M_n(\lambda_j)$$

Le λ optimal est celui qui minimise ce score, soit :

$$\hat{\lambda} = \arg\min_{\lambda_j} s(\lambda_j)$$

L'objectif de la validation croisée est de déterminer la meilleure valeur de λ, permettant de trouver un équilibre entre un modèle qui sur-apprend (overfitting) et un modèle trop simple (underfitting). Une fois cette valeur optimale de λ identifiée, elle sera utilisée pour définir l'estimateur final.
Pour cela, on utilise une solution "closed-form"qui définit l'estimateur pour chaque valeur de λ. Tout d'abord, considérons $B_1 = \{(Y_i, Z_i, W_i) \mid 1 \leq i \leq n/2\}$. Lorsque $\lambda = 0$, le modèle est trivial (sur-apprentissage), et lorsque λ est très faible, le modèle devient trop simple