Régression semi-paramétrique de variables explicatives de dénombrement

Belkacem Abdous, Célestin Kokonendji, Tristan Senga Kiessé

Regression semi-paramétrique de variables explicatives de dénombrement

Belkacem Abdous

Université Laval, Médecine Sociale et Préventive
Québec, Canada G1K 7P4
Email : belkacem.abdous@msp.ulaval.ca

Célestin C. Kokonendji

Université de Franche-Comté, UFR Sciences et Techniques
Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS
16 route de Gray – 25030 Besançon cedex, France
Email : celestin.kokonendji@univ-fcomte.fr

Tristan Senga Kiessé

Laboratoire Central des Ponts et Chaussées
Division ESAR – Section AGR
Route de Bouaye – BP 4129 – 44341 Bouguenais cedex, France
Email : tristan.senga-kiesse@lcpc.fr

RÉSUMÉ : Dans cette communication, nous proposons un estimateur semi-paramétrique d'une fonction discrète de régression. Nous combinons une approche paramétrique et une non-paramétrique pure qui met en oeuvre la méthode des noyaux associés discrets. Le nouvel estimateur semi-paramétrique discret de régression possède un biais qui peut être inférieur à celui de l' estimateur purement non-paramétrique discret de régression et une même variance que ce dernier.

ABSTRACT: We propose a semiparametric estimator for a count regression function. We combine a parametric approach and a pure nonparametric one which is realized by using discrete associated-kernel method. The new semiparametric count regression estimator can reduce the bias with respect to purely nonparametric count regression estimator, without affecting the variance.

Mots clés : Estimateur à noyau associé discret ; régression non-paramétrique

Dans ce travail, nous partons de la décomposition d'une fonction discrète de régression $m : \mathbb{N}^d \to \mathbb{R}$ comme le produit

$$\begin{aligned} m(x) &= r(x;\beta)\omega(x) \\ &= : m_\omega(x;\beta), \quad \forall x \in \mathbb{N}^d. \end{aligned} \tag{1}$$

La fonction $r(x;\beta)$ de (1) est la partie paramétrique de paramètre β et la fonction $\omega(x)$ de (1) est la partie non-paramétrique sur $\mathbb{N}^d$. Pour construire un estimateur semi-paramétrique $\widehat{m}$ de m en (1), l'idée est d'abord de déterminer une estimation paramétrique $\widehat{r}$ de la fonction r puis de corriger en multipliant par une estimation entièrement non-paramétrique de $\omega = m/\widehat{r}$; voir, Kokonendji *et al.* (2009a), pour l'estimation semi-paramétrique de distribution de données de dénombrement.

Tout d'abord, l'estimateur non-paramétrique d'une fonction de régression $m(x)$ sur $\mathbb{N}^d$ avec $d = 1$ a été défini par Kokonendji *et al.* (2009b) sur le modèle de celui de Nadaraya-Watson comme suit :

$$\widetilde{m}_n(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)}, \quad x \in \mathbb{N}, \tag{2}$$

avec Y_i la variable aléatoire (v.a.) réelle à expliquer, X_i la variable explicative de dénombrement, $h = h(n)$ une suite arbitraire de paramètres de lissage satisfaisant $\lim_{n\to\infty} h(n) = 0$ et $K_{x,h}(\cdot)$ un noyau associé discret (Kokonendji *et al.*, 2007, et Senga Kiessé, 2008) ; la variance $Var(Y|X = x)$ est finie pour tout $x \in \mathbb{N}$. Précisons qu'une fonction noyau associé discret $K_{x,h}(\cdot)$ est elle même une fonction de masse de probabilité (f.m.p.) sur le support discret S_x (ne dépendant pas de h). En plus, nous imposons les deux conditions suivantes :

$$\lim_{h\to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x \quad \text{et} \quad \lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) = 0, \tag{3}$$

où $\mathcal{K}_{x,h}$ est la v.a. discrète de f.m.p. $K_{x,h}(\cdot)$; voir, Abdous et Kokonendji (2009) pour une revue sur les noyaux associés discrets. Les différences finies d'une f.m.p. g sont telles que, pour $k \in \mathbb{N} \setminus \{0\}$, on ait :

$$g^{(k)}(x) = \{g^{(k-1)}(x)\}^{(1)} \text{ et } g^{(1)}(x) = \begin{cases} \{g(x+1) - g(x-1)\}/2 & \text{si } x \in \mathbb{N} \setminus \{0\} \\ g(1) - g(0) & \text{si } x = 0. \end{cases} \tag{4}$$

De là, la différence finie d'ordre 2 est

$$g^{(2)}(x) = \begin{cases} \{g(x+2) - 2g(x) + g(x-2)\}/4 & \text{si } x \in \mathbb{N} \setminus \{0,1\} \\ \{g(3) - 3g(1) + 2g(0)\}/4 & \text{si } x = 1 \\ \{g(2) - 2g(1) + g(0)\}/2 & \text{si } x = 0. \end{cases} \tag{5}$$

Le résultat suivant présente le biais et la variance asymptotiques de l'estimateur non-paramétrique $\widetilde{m}_n$ de (2).

Théorème 1 *Soit X la v.a. discrète de f.m.p. telle que $f(x) = \Pr(X = x) > 0$, pour tout $x \in \mathbb{N}$ fixé. Soit $h = h(n)$ une suite de nombres positifs satisfaisant $\lim_{n\to\infty} h = 0$. Alors, le biais et la variance de $\widetilde{m}_n(x)$ admettent les expressions suivantes :*

$$biais\{\widetilde{m}_n(x)\} = \left\{ m^{(2)}(x) + 2m^{(1)}(x) \left(\frac{f^{(1)}}{f} \right)(x) \right\} \frac{Var(\mathcal{K}_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h), \tag{6}$$

$$Var\{\widetilde{m}_n(x)\} = \frac{Var(Y|X=x)}{nf(x)}\{\Pr(\mathcal{K}_{x,h}=x)\}^2 + o\left(\frac{1}{n}\right), \tag{7}$$

où $m^{(2)}$, $m^{(1)}$ and $f^{(1)}$ sont les différences finies définies comme dans (5) et (4).

Ensuite, l'approche non-paramétrique par noyau associé discret appliquée à la fonction discrète $\omega(x)$ permet de définir l'estimateur semi-paramétrique $\widehat{m}_n$ de m de la manière suivante :

$$\widehat{m}_n(x) = r(x;\widehat{\beta})\sum_{i=1}^{n}\frac{Y_i K_{x,h}(X_i)}{r(X_i;\widehat{\beta})\sum_{j=1}^{n}K_{x,h}(X_j)} \tag{8}$$

$$= \sum_{i=1}^{n}\frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n}K_{x,h}(X_j)} \times \frac{r(x;\widehat{\beta})}{r(X_i;\widehat{\beta})}, \quad x \in \mathbb{N}, \tag{9}$$

où $\hat{\beta}$ est l'estimateur de β, $h > 0$ est la fenêtre et $K_{x,h}(\cdot)$ est un noyau associé discret donné. Ainsi, dans le cas où la fonction $r_0(x) = r(x;\beta_0)$ est fixée, en notant $m = r_0\omega$ nous formulons le résultat suivant.

Théorème 2 *Pour tout $x \in \mathbb{N}$ fixé tel que $f(x) = \Pr(X = x) > 0$, soit $r_0(x) = r(x;\beta_0)$ un départ fixé dans (1). Pour $n \to \infty$ et $h = h(n) \to 0$, l'estimateur semi-paramétrique de régression $\widehat{m}_n(x)$ dans (8) possède le biais et la variance suivants :*

$$biais\{\widehat{m}_n(x)\} = \left\{r_0(x)\omega^{(2)}(x) + 2r_0(x)\omega^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}\frac{Var(\mathcal{K}_{x,h})}{2} + O\left(\frac{1}{n}\right) + o(h), \tag{10}$$

$$Var\{\widehat{m}_n(x)\} = \frac{Var(Y|X=x)}{nf(x)}\{\Pr(\mathcal{K}_{x,h}=x)\}^2 + o\left(\frac{1}{n}\right), \tag{11}$$

où $\omega^{(2)}$, $\omega^{(1)}$ et $f^{(1)}$ sont les différences finies définies comme dans (5) et (4).

La comparaison du résultat précédent avec celui correspondant à l'estimateur non-paramétrique dans le Théorème 1 montre que les variances en (11) et en (7) des deux estimateurs sont égales. Ainsi, le nouvel estimateur $\widehat{m}_n$ en (8) pourrait être plus performant que l'estimateur $\widetilde{m}_n$ en (2) selon la valeur de leur biais respectif en (10) et en (6). Pour un noyau associé discret choisi, la comparaison provient des termes des différences finies d'ordre k présents dans les expressions (10) et (6) ; en effet, selon le départ r_0 fixé , on a :

$$m^{(1)} = (r_0\omega)^{(1)} = r_0\omega^{(1)} + r_0^{(1)}\omega \lesseqgtr r_0\omega^{(1)} \text{ et}$$

$$m^{(2)} = (r_0\omega)^{(2)} = r_0\omega^{(2)} + 2r_0^{(1)}\omega^{(1)} + r_0^{(2)}\omega \lesseqgtr r_0\omega^{(2)},$$

où le symbole $\lesseqgtr$ signifie $\leq$ ou $\geq$.

Un résultat similaire au Théorème 2 se formule dans le cas général où le départ n'est pas fixé. Pour cela, la fonction $r(x;\beta_0)$ est déterminée comme étant la fonction approchant le mieux $m(x)$ au sens de la distance de Kullback-Leibler.

Enfin, les estimateurs en (8) et en (2) peuvent s'étendre au cas multidimensionnel. Il est alors nécessaire d'utiliser un noyau associé multiplicatif

$$K_{x,h}(X_i) = \prod_{k=1}^{d} K^{[k]}_{x_k,h_k}(X_{ik}),$$

avec $x = (x_1, \ldots, x_d)^T$, $h = (h_1, \ldots, h_d)^T$, $X_i = (X_{i1}, \ldots, X_{id})^T$ et $K^{[k]}_{x_k,h_k}(\cdot)$ un noyau associé univarié satisfaisant (3) pour tout $k = 1, \ldots, d$. Il faut aussi utiliser les versions multidimensionnelles des fonctions $m(\cdot)$, $r(\cdot;\beta)$ et $\omega(\cdot)$ avec les dérivées partielles correspondantes. Il s'en suit la formulation des différents résultats présentés dans le cas multivarié.

Bibliographie

[1] Abdous, B., Kokonendji, C.C. (2009). Consistency and asymptotic normality for discrete associated-kernel estimator. *African Diaspora Journal of Mathematics* 8, 63–70.

[2] Kokonendji, C.C., Senga Kiessé, T., Balakrishnan, N. (2009a). Semiparametric estimation for count data through weighted distributions. *Journal of Statistical Planning and Inference* 139, 3625–3638.

[3] Kokonendji, C.C., Senga Kiessé, T., Demétrio, C.G.B. (2009b). Appropriate kernel regression on a count explanatory variable and applications. *Advances and Applications in Statistics* 12, 99–126.

[4] Kokonendji, C.C., Senga Kiessé, T. et Zocchi, S.S. (2007). Discrete triangular distributions and non-parametric estimation for probability mass function. *Journal of Nonparametric Statistics* 19, 241–254.

[5] Senga Kiessé, T. (2008). Approche non-paramétrique par noyaux associés discrets des données de dénombrement. Thèse de Doctorat de Statistique de l'Université de Pau. URL http://tel.archives-ouvertes.fr/tel-00372180/fr/