C. R. Acad. Sci. Paris, Ser. I 334 (2002) 59–63

Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Estimation non-paramétrique de la régression dichotomique – application biomédicale

Gérard Derzko [a], Paul Deheuvels [b]

[a] Sanofi-Synthelabo Recherche, 371, rue du professeur Joseph Blayac, 34184 Montpellier cedex 04, France

[b] LSTA, Université Paris VI, 8A23, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 30 juin 2001 ; accepté après révision le 12 novembre 2001

Note présentée par Marc Yor.

Résumé Nous considérons un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de répliques indépendantes de (X, Y), où X est une v.a. possédant une densité conditionnellement à une v.a. discrète Y prenant les valeurs 0 ou 1. Nous estimons la régression dichotomique $R(x) = \mathbb{E}(Y \mid X = x)$ par un estimateur non-paramétrique $\widehat{R}_n(x)$ de type Nadaraya–Watson, dont nous décrivons le comportement limite. Ces résultats sont appliqués à l'exemple biomédical du pronostic de décès sur une période fixée, connaissant la variation de la capacité vitale chez des patients atteints de sclérose latérale amyotrophique. Pour citer cet article : G. Derzko, P. Deheuvels, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 59–63. © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Nonparametric estimation of dichotomic regression with biomedical applications

Abstract We consider a Nadaraya–Watson-type nonparametric estimator $\widehat{R}_n(x)$ of the dichotomic regression $R(x) = \mathbb{E}(Y \mid X = x)$, given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). We assume that X is a r.v. with continuous density, depending upon the values $Y = 0$ or 1 of an indicator r.v. We give a description of the large sample limiting behaviour of $\widehat{R}_n(x)$, and illustrate the method by a biomedical example: the prediction of death rate during a fixed period of time given the variation of vital capacity in patients suffering from amyothrophic lateral sclerosis. To cite this article: G. Derzko, P. Deheuvels, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 59–63. © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

1. Introduction

Soient X une variable aléatoire [v.a.] continue, et Y une v.a. binaire pouvant prendre deux modalités notées 0 et 1. L'estimation de la *régression dichotomique* $R(x) = \mathbb{E}(Y \mid X = x) = \mathbb{P}(Y = 1 \mid X = x)$, probabilité conditionnelle que $Y = 1$ sachant $X = x$, à partir d'un échantillon $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ de (X, Y), présente de l'intérêt pour de nombreuses applications biomédicales. Les valeurs de Y, codées ici

Adresses e-mail : gerard.derzko@sanofi-synthelabo.com (G. Derzko); pd@ccr.jussieu.fr (P. Deheuvels).

S1631-073X(02)02202-1/FLA

par 0 ou 1 par commodité, ne sont pas supposées ordonnées, et ce modèle se généralise à la *régression catégorielle* (ou *polytomique*), correspondant au cas où Y prend ses valeurs parmi un nombre fini de modalités. Nous considérerons dans cette Note la régression dichotomique pour simplifier, tout en notant que nos méthodes et résultats se généralisent sans difficulté au cas polytomique.

Un tel modèle est fréquent en recherche clinique et en épidémiologie, où l'on s'intéresse à la description d'événements cliniques ultimes (décès, crise, agravation,...) qui peuvent se produire ou non chez un malade au cours d'une période fixée d'observation, quelquefois avec une fréquence très faible. Le fait que ces événements aient lieu ou non est ici codé par la valeur 1 ou 0 de la v.a. Y. Ces événements étant difficiles à pronostiquer, on recherche des variables plus facilement observables, appelées *marqueurs*, ou *variables intermédiaires*. Une v.a. de ce type, soit X, n'a d'intérêt que si elle possède un caractère prédictif pour l'événement clinique ultime $Y = 1$. Cette liaison s'estime donc naturellement à l'aide de la régression dichotomique $R(x) = \mathbb{E}(Y \mid X = x)$.

La méthode d'élection utilisée par les expérimentateurs pour l'estimation de $R(x)$ est la *régression logistique*. Celle-ci, disponible dans les principaux logiciels statistiques, fait usage du modèle linéaire généralisé (*cf.* [8]), ou de modèles additifs (*cf.* [7]), et suppose le choix a priori d'une fonction de lien exprimant une relation fonctionnelle entre des paramètres liés à X et Y, sans oublier une modélisation des résidus de régression. L'usage de telles techniques en dehors de la connaissance précise des structures de dépendance entre X et Y peut mener à des évaluations erronées, et il convient d'utiliser d'autres méthodes ne dépendant pas de choix subjectifs de modèle.

Nous développons ici une famille de méthodes non-paramétriques (*cf.* [1,5,10–12]) dont le prototype est l'estimateur de la régression non-paramétrique de Nadaraya et Watson (*cf.* [9,13]). Nous présentons les résultats correspondants dont la généralisation dans le contexte des estimateurs à noyau discret (*cf.* [4]) est naturelle.

2. Estimation non-paramétrique de la régression dichotomique

Nous adopterons les notations et hypothèses suivantes. Soit (X, Y) un vecteur aléatoire. On note $p_0 = \mathbb{P}(Y = 0)$, $p_1 = \mathbb{P}(Y = 1)$ avec $p_0 + p_1 = 1$ et $0 < p_1 < 1$, $\mathbb{P}(\cdot)$ désignant la probabilité. On suppose que $X \in \mathbb{R}$ est une v.a. telle que, $\forall A$ borélien de $\mathbb{R}$,

$$\mathbb{P}(Y = 0, X \in A) = p_0 \int_A f_0(x)\,\mathrm{d}x, \qquad \mathbb{P}(Y = 1, X \in A) = p_1 \int_A f_1(x)\,\mathrm{d}x, \tag{2.1}$$

où f_0 (resp. f_1) est la densité de probabilité conditionnelle sur $\mathbb{R}$ de X sachant $Y = 0$ (resp. $Y = 1$). On note f la densité marginale de X sur $\mathbb{R}$ donnée, compte tenu de (2.1), par

$$f(x) = p_0 f_0(x) + p_1 f_1(x) \quad \text{pour } x \in \mathbb{R}. \tag{2.2}$$

La régression de Y sachant X est donnée ($\mathbb{E}(\cdot)$ désignant l'expérance mathématique) par

$$R(x) = \mathbb{E}(Y \mid X = x) = \frac{p_1 f_1(x)}{f(x)} = \frac{p_1 f_1(x)}{p_0 f_0(x) + p_1 f_1(x)}. \tag{2.3}$$

L'expression (2.3) n'a de sens que lorsque x appartient au support de f. Nous raisonnerons dans ce qui suit en supposant donnés des intervalles $I = [a, b] \subset I' = [a', b']$, tels que $-\infty < a' < a < b < b' < \infty$, et tels que les hypothèses suivantes soient satisfaites.

(F1) f_0 et f_1 sont continues sur I' ;

(F2) $\inf_{x \in I'} f_0(x) > 0$ et $\inf_{x \in I'} f_1(x) > 0$.

Nous sommes ici intéressés par l'étude de $R(\cdot)$ sur I. L'introduction du voisinage auxiliaire I' de I vise à permettre certains passages à la limite. En particulier, sous (F1)–(F2), la régression dichotomique $R(\cdot)$ est définie, positive et continue, sur I'. Nous estimons cette fonction à partir de l'échantillon

$(X_1, Y_1), \ldots, (X_n, Y_n)$ de répliques indépendantes de (X, Y), par l'estimateur de Nadaraya–Watson (*cf.* [9, 13])

$$\widehat{R}_n(x) = \frac{\sum_{i=1}^n Y_i K((x - X_i)/h_n)}{\sum_{i=1}^n K((x - X_i)/h_n)}. \tag{2.4}$$

Ici, $h_n > 0$ est une suite de constantes positives, et $K(\cdot)$ un *noyau*, fonction réelle, vérifiant les conditions ci-dessous, pour une constante convenable $M < \infty$.

(H1) $h_n \to 0$ lorsque $n \to \infty$;

(H2) $nh_n / \log n \to \infty$ lorsque $n \to \infty$.

(K1) $\int_{\mathbb{R}} K(t)\,\mathrm{d}t = 0$;

(K2) K est à variation bornée sur $\mathbb{R}$;

(K3) $K(t) = 0$ pour $|t| \geqslant M$.

Notons que $K(t)$ peut être négatif pour certaines valeurs de t, ce qui ne garantit pas l'existence de $\widehat{R}_n(x)$ pour tout $x \in I'$ et pour des valeurs arbitraires de $n \geqslant 1$. Toutefois, il ressort du théorème 1.1 de [2] que, sous (H1)–(H2) et (K1)–(K3), l'estimateur de la densité $f(x)$ défini par

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \quad \text{pour } x \in \mathbb{R}, \tag{2.5}$$

est tel que (en notant $[\mathbb{P}]$ pour la convergence en probabilité)

$$\lim_{n\to\infty} \left\{ \sup_{x\in I} \left| \widehat{f}_n(x) - f(x) \right| \right\} = 0 \quad [\mathbb{P}]. \tag{2.6}$$

Sous (F1)–(F2), (2.6) implique que, $\forall \varepsilon > 0$, il existe un $n_\varepsilon < \infty$ tel que la probabilité que $R_n(x)$ soit défini, pour tout $n \geqslant n_\varepsilon$ et tout $x \in I$, soit supérieure ou égale à $1 - \varepsilon$. Dans la suite de l'exposé nous raisonnerons donc sur l'événement correspondant, en supposant implicitement que n soit assez grand pour que $R_n(x)$ soit défini $\forall x \in I$.

Posons, de plus,

$$R_n(x) = \frac{\int_{\mathbb{R}} p_1 f_1(x - h_n u) K(u)\,\mathrm{d}u}{\int_{\mathbb{R}} f(x - h_n u) K(u)\,\mathrm{d}u}. \tag{2.7}$$

Il est aisé de constater que sous les hypothèses (F1)–(F2), (H1) et (K1)–(K3), on a

$$\lim_{n\to\infty} \left\{ \sup_{x\in I} \left| R_n(x) - R(x) \right| \right\} = 0. \tag{2.8}$$

La vitesse de convergence dans (2.8) est un problème purement analytique dépendant d'hypothèses additionnelles sur la régularité de f_0 et f_1. Nous laissons de côté l'étude de celle-ci, pour concentrer notre intérêt sur le terme aléatoire $\widehat{R}_n(x) - R_n(x)$. Pour cela, nous introduisons une suite auxiliaire de fonctions $\Psi_n(x)$ (possiblement aléatoires) et une fonction $\Psi(x)$ continue et positive (non aléatoire), toutes deux définies sur $I' = [a', b']$, de sorte que

$$\lim_{n\to\infty} \left(\sup_{x\in I} \left| \frac{\Psi_n(x)}{\Psi(x)} - 1 \right| \right) = 0 \quad [\mathbb{P}]. \tag{Ψ}$$

THÉORÈME. – *Sous les conditions* (F1)–(F2), (H1)–(H2), (K1)–(K3) *et* (Ψ), *on a*

$$\lim_{n\to\infty} \left\{ \frac{nh_n}{2\log(1/h_n)} \right\}^{1/2} \sup_{x\in I} \left| \widehat{R}_n(x) - R_n(x) \right| \Psi_n(x)$$
$$= \left\{ \sup_{x\in I} \left(\frac{R(x)(1 - R(x))}{f(x)} \right) \Psi^2(x) \int_R K^2(t)\,\mathrm{d}t \right\}^{1/2} \quad [\mathbb{P}]. \tag{2.9}$$

Remarque. – Les choix suivants de Ψ_n et Ψ illustrent l'intérêt de la formulation (2.9). Ci-dessous, f_n est, soit l'estimateur $\widehat{f_n}$ de f donné en (2.5), soit un estimateur auxilaire de f choisi de sorte que (Ψ) soit vérifié.

$$\begin{aligned} &\Psi_n(x) = \sqrt{f_n(x)} \quad \text{et} \quad \Psi(x) = \sqrt{f(x)}\,; \quad \Psi_n(x) = 1 \quad \text{et} \quad \Psi(x) = 1\,; \\ &\Psi_n(x) = \left\{ \frac{f_n(x)}{\widehat{R}_n(x)(1-\widehat{R}_n(x))} \right\}^{1/2} \quad \text{et} \quad \Psi(x) = \left\{ \frac{f(x)}{R(x)(1-R(x))} \right\}^{1/2}. \end{aligned} \tag{2.10}$$

Démonstration. – La démonstration utilise les méthodes de [6] et [2]. Nous notons tout d'abord qu'une modification des arguments de [6] permet d'obtenir la validité d'une version du corollaire 3, p. 7 de [6] sous les hypothèses (H1)–(H2). Des arguments semblables sont utilisés dans [2]. Par ailleurs, on se ramène comme en [2] au cas de $\Psi_n(x) = \Psi(x) = 1$ au prix d'un découpage convenable des intervalles I et I'. Bien que le théorème semble, à première vue, une conséquence directe de cette version modifiée du corollaire 3 de [6], tel n'est pas le cas, puisque ce dernier corollaire ne peut être appliqué que lorsque (X, Y) possède une densité continue sur un ouvert convenable de $\mathbb{R}^2$ (*voir* (G.i)–(G.ii), p. 4 de [2]). Pour pallier cette difficulté, il faut construire une variable accessoire Z, de telle sorte que (X, Z) ait une densité jointe sur $I \times \mathbb{R}$ vérifiant les hypothèses (G.i)–(G.ii), p. 4 de [6], et que Y s'exprime comme fonction monotone de Z. Or, ceci s'obtient grâce aux hypothèses admises (F1)–(F2).

Comme dans [4], des résultats semblables peuvent être obtenus pour des régressogrammes moyennés, ainsi que pour des estimateurs à *bande* $h_n = h_n(x)$ variable en fonction de x. Les théorèmes correspondants seront explicités dans des articles ultérieurs.

3. Application

Les données sont issues d'une étude clinique portant sur $n = 2046$ patients, atteints de *sclérose latérale amyotrophique* [SLA], et observés pendant 1 an et demi. La SLA est une pathologie très grave, qui endommage les neurones moteurs, et génère, entre autres, des troubles respiratoires, avant que le *décès* ne survienne (la médiane de la survie est comprise entre 3 et 5 ans). L'état respiratoire du patient peut être décrit par une mesure de pourcentage de sa *capacité vitale* [CV], effectuée à 1 % près. L'estimation du lien entre la *diminution* X de la capacité vitale d'un patient, et sa survie ($Y = 0$) ou son décès ($Y = 1$), pendant la période d'observation, permet d'établir le caractère pronostique de la première quantité sur la seconde. La figure 1 ci-dessous montre une estimation non-paramétrique $\widehat{R}_n(x)$ de la régression $R(x)$ telle qu'en (2.4), ainsi qu'une estimation $f_n(x)$ de la densité marginale $f(x)$ de la CV X par la méthode [4], et ce, pour une largeur de bande h_n de 5 %. Une bande de confiance déduite de (2.9) est également représentée. L'estimation $\widehat{R}_n(x)$ de $R(x)$ manque de précision dans les régions de faible densité (pour les valeurs de x pour lesquelles $f(x)$ est petit). Toutefois, les résultats fournis par cette analyse confirment des propriétés déjà observées dans des expériences antérieures. En particulier, l'existence, a priori inattendue, de taux *élevés* de décès pour des valeurs *faibles* de la variation X de CV (qui correspondent à une *amélioration* de la CV) sont connus des experts, sans pour autant qu'une explication logique ait pu en être donnée. De l'autre côté de l'échelle, le plafonnement du taux de mortalité pour des valeurs élevées de la variation X de CV s'interprète par l'effet positif de traitements tels que la mise sous assistance respiratoire des patients, ou la réalisation de trachéotomies. En l'absence de tels actes médicaux, les patients dont la fonction respiratoire est la plus détériorée ne pourraient survivre, et leur réalisation devient systématique. À l'inverse de ces caractéristiques *non linéaires* de $R(\cdot)$ aux extrémités du domaine de variation de X, la partie centrale révèle une variation quasi-linéaire de $R(\cdot)$, s'interprétant comme une augmentation moyenne de la mortalité de 1 % pour chaque pourcent de diminution de capacité vitale, lorsque celle-ci se situe entre 0 % et 50 % (ceci concerne 1490 patients sur 2046). On peut s'assurer par ailleurs que cette relation linéaire demeure très stable pour les différents sous-groupes de patients. La comparaison, sur la figure 1, de ces résultats avec les estimations fournies par la *régression logistique* sur l'ensemble des données montre que cette dernière estime mal la

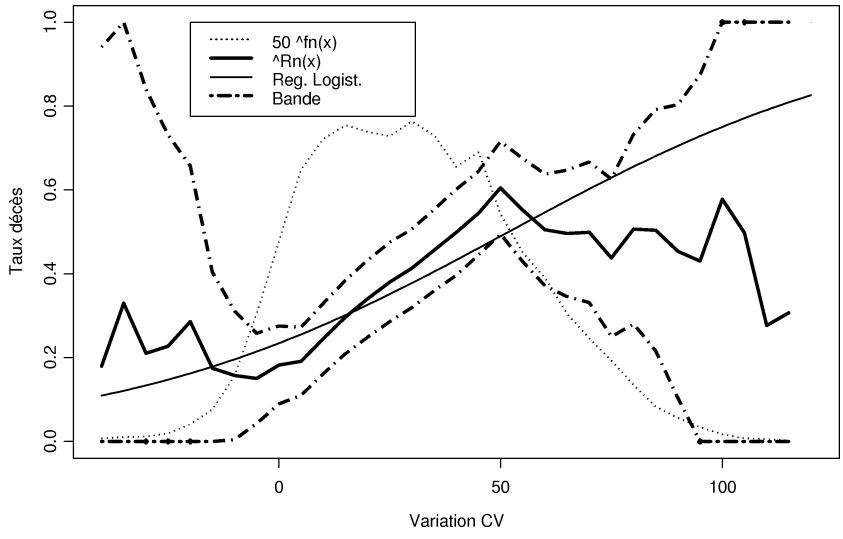

Figure 1.

relation entre la variation de CV et la mortalité dans la partie centrale, tout en ne reflétant pas correctement les particularités du phénomène aux extrémités de l'échelle.

Un tel exemple justifie pleinement l'intérêt d'une utilisation systématique de ces méthodes non-paramétriques en complément des techniques habituelles. Leur capacité à mettre en évidence des particularités cliniquement importantes des relations liant X à Y apparaît comme indéniable. Une quantification des effets observés dans le cadre non-paramétrique peut naturellement être obtenue par un usage ultérieur de modèles paramétriques appliqués dans les zones de variation de X où le nombre d'observations permet une estimation précise.

Références bibliographiques

[1] Bosq D., Lecoutre J.P., Théorie de l'estimation fonctionnelle, Economica, Paris, 1987.
[2] Deheuvels P., Einmahl J.H.J., Functional limit laws for the increments of Kaplan–Meier product-limit processes and applications, Ann. Probab. 28 (2000) 1301–1335.
[3] Deheuvels P., Mason D.M., Functional laws of the iterated logarithm for the increments of empirical and quantile processes, Ann. Probab. 22 (1992) 1619–1661.
[4] Derzko G., Une approche intrinsèque de l'estimation non paramétrique de la densité, C. R. Acad. Sci. Paris, Série I 327 (1998) 985–988.
[5] Devroye L., Györfi L., Nonparametric Density Estimation: The L_1 View, Wiley, New York, 1985.
[6] Einmahl U., Mason D.M., An empirical approach to the uniform consistency of kernel-type function estimators, J. Theoret. Probab. 13 (2000) 1–37.
[7] Hastie T., Tibshirani R., Generalized Additive Models, Chapman and Hall, London, 1990.
[8] McCullagh P., Nelder J.A., Generalized Linear Models, Chapman and Hall, London, 1989.
[9] Nadaraya E.A., On estimating regression, Theoret. Probab. Appl. 9 (1964) 141–142.
[10] Rosenblatt M., Curve estimates, Ann. Math. Statist. 42 (1971) 1815–1841.
[11] Scott D.W., Multivariate Density Estimation, Theory, Practice, and Visualization, Wiley, New York, 1992.
[12] Tapia R.A., Thompson J.R., Nonparametric Probability Density Estimation, John Hopkins University Press, Baltimore, 1978.
[13] Watson G.S., Smooth regression analysis, Sankhyā 26 (1964) 359–372.