Test non paramétrique d'adéquation pour un modèle de régression en présence de censure dépendant des variables explicatives

Olivier Lopez, Valentin Patilea

Test non paramétrique d'adéquation pour un modèle de régression en présence de censure dépendant des variables explicatives

Olivier Lopez & Valentin Patilea

Université Paris VI
Laboratoire de Statistique Théorique et Appliquée
175 rue du Chevaleret
75013 Paris
&
Insa-Irmar
Centre des Mathématiques
Institut National des Sciences Appliquées (INSA) de Rennes 20, Avenue des Buttes de Coësmes
CS 14315, 35043 Rennes Cedex

Résumé. Nous proposons une nouvelle méthode afin de tester contre une alternative non paramétrique l'hypothèse

$$H_0 \ : \ \exists \theta_0 \in \Theta, \text{ t.q. } E[m_{\theta_0}(Y,X)|X] = 0, \tag{1}$$

où m_θ est une famille de fonctions connue, $\theta_0 \in \Theta \subset \mathbf{R^k}$ un paramètre inconnu de dimension finie, $Y \in \mathbf{R}$ et $X \in \mathbf{R^d}$ des vecteurs aléatoires. La formulation générale du modèle (1) permet de recouvrir un nombre important de modèles classiquement utilisés en statistique. Cette formulation fournit notamment le cas particulier d'un modèle de régression paramétrique portant sur l'espérance conditionnelle (en considérant $m_{\theta_0}(Y,X) = Y - f_0(\theta_0, X)$ où f_0 est une fonction connue), ou encore un modèle paramétrique de régression quantile. Nous nous intéressons plus particulièrement à la détermination d'une procédure de test de (1) dans le cas où la variable Y est censurée. Nous considérerons essentiellement le cas d'une censure aléatoire à droite, et préciserons comment notre méthode peut être étendue à d'autres types de censure (par exemple censure bilatérale).

Dans un modèle de régression censurée à droite, les observations sont constituées de $(T_i, \delta_i, X_i^T)_{1\leq i\leq n}$ avec

$$\begin{aligned} T_i &= \inf(Y_i, C_i) \\ \delta_i &= \mathbf{1}_{Y_i \leq C_i}, \end{aligned}$$

où les variables Y_i, C_i, X_i sont i.i.d, les variables C_i étant dites variables de censure. Lopez et Patilea (2009) ont considéré ce problème dans le cas où $m_{\theta_0}(Y,X) = Y - f(\theta_0, X)$, et dans le cas où le modèle de censure satisfait une hypothèse d'identifiabilité forte, i.e. Y_i indépendant de C_i, et $\mathbf{P}(Y_i \leq C_i | X_i, Y_i) = \mathbf{P}(Y_i \leq C_i | Y_i)$, hypothèse notamment vérifiée

dans le cas où C_i est indépendante de X_i et Y_i. Cette hypothèse, même si elle reste adaptée à un bon nombre de situations (censure pour causes administratives notamment), pose d'importantes restrictions sur la loi de C_i sachant X_i.

Dans ce travail, nous nous intéressons au cas plus général où le modèle repose sur l'hypothèse d'identifiabilité suivante,

$$Y_i \perp C_i | X_i,$$

de sorte que le modèle impose peu de restrictions sur la loi conditionnelle de C_i sachant X_i. Nous considérons une statistique de test basée sur une adaptation de la démarche utilisée par Zheng (1996) et Horowitz et Spokoiny (2001) en l'absence de censure. Notre méthode repose sur un estimateur nonparamétrique de la fonction de répartition multivariée $F(x, y) = P(X \leq x, Y \leq y)$ proposé par Lopez (2007), et qui possède l'avantage de préserver les performances de notre méthode dans le cas où X est un vecteur de grande dimension.

Nous obtenons une représentation asymptotique de notre statistique de test se rapprochant de celle obtenue sous des hypothses plus restrictives par Lopez et Patilea (2009) sous l'hypothèse plus restrictive, avec des termes additionnels provenant de la loi conditionnelle de C_i sachant X_i. Cette représentation asymptotique nous permet de proposer des méthodes de simulations permettant d'améliorer notre procédure à distance finie, ainsi que de montrer la consistance de notre test envers des alternatives locales se rapprochant de l'hypothèse nulle.

Mots clés : Tests non paramétriques, censure aléatoire, estimateur de Kaplan-Meier conditionnel, modèles de régression

Abstract. We introduce a new method for nonparametric testing of the following null hypothesis

$$H_0 \ : \ \exists \theta_0 \in \Theta, \text{ t.q. } E[m_{\theta_0}(Y, X)|X] = 0, \tag{2}$$

where m_θ is a family of known functions, $\theta_0 \in \Theta \subset \mathbf{R^k}$ an unknown finite dimensional parameter, $Y \in \mathbf{R}$ and $X \in \mathbf{R^d}$ are random vectors. The general expression of the model (2) allows to cover a large number of classical situations considered in statistics. This especially covers the case of parametric mean-regression (considering $m_{\theta_0}(Y, X) = Y - f_0(\theta_0, X)$ where f_0 is a known function), and the case of a parametric quantile regression model. We are interested in providing a test procedure for (2) in the case where the random variable Y is censored. We essentially consider the case of randomly right-censored data, and give some indications on how to extend this procedure to more general censoring models (such as bilateral censoring for example).

In a regression model with right-censored responses, the observations are made of $(T_i, \delta_i, X_i^T)_{1 \leq i \leq n}$ where

$$\begin{aligned} T_i &= \inf(Y_i, C_i) \\ \delta_i &= \mathbf{1}_{Y_i \leq C_i}, \end{aligned}$$

where Y_i, C_i, X_i are i.i.d, and C_i are the censoring variables. Lopez and Patilea (2009) considered this problem in the case where $m_{\theta_0}(Y, X) = Y - f_{\theta_0}(Y, X)$, and with a strong identifiability assumption, i.e. Y_i independent of C_i, and $\mathbf{P}(Y_i \leq C_i | X_i, Y_i) = \mathbf{P}(Y_i \leq C_i | Y_i)$, assumption which holds in the particular case where C_i is independent from X_i and Y_i. This assumption, although it covers a large number of situations (administrative censoring for instance), relies on strong restrictions on the conditional law of C_i given X_i.

In this work, we consider the more general case of a model relying on the following identifiability assumption,

$$Y_i \perp C_i | X_i,$$

so that this assumption does not require lots of restrictions on the conditional law of C_i given X_i. We consider a test statistic based on an adaptation of the procedure considered by Zheng (1996) and Horowitz and Spokoiny (2001) in absence of censoring. Our method relies on a nonparametric estimator of the multivariate distribution function $F(x, y) = P(X \leq x, Y \leq y)$ introduced by Lopez (2007), which preserves the performances of our method in the case where X is an high-dimensional vector.

We obtain an asymptotic i.i.d. representation of our test statistic which is close to the one obtained under the more restrictive identifiability assumption by Lopez and Patilea (2009), but with additional terms which depend on the conditional law of C_i given X_i. This asymptotic representation allows us to propose simulations methods which improve the behavior of our procedure for a finite sample-size, and allows us to derive results on consistency of our test against local alternatives.

Keywords : Nonparametric testing, random censoring, conditional Kaplan-Meier, regression models

Bibliographie

[1] J.L. Horowitz, V.G. Spokoiny, (2001) *An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative*, Econometrica 69, pages 599-631.

[2] O. Lopez (2007) *On the estimation of the joint distribution in a censored regression model*, Document Crest 2007–11.

[3] O. Lopez, V. Patilea (2009) *Nonparametric lack-of-fit tests for parametric mean-regression models with censored data*, Journal of Multivariate Analysis, Volume 100, Issue 1, January 2009, pages 210–230.

[4] J.X. Zheng (1996), *A consistent test of functional form via nonparametric estimation techniques*, J. Econometrics 75 pages 263-289.