A nonparametric goodness-of-fit test for two-component mixture cure models in survival analysis

Valentin Patilea

Ensai, Campus de Ker-Lann, Bruz. E-mail: valentin.patilea@ensai.fr

Résumé. Les modèles de régression prenant en compte une proportion des individus avec un temps de vie infini sont souvent utilisés en analyse de survie, fiabilité ou économétrie. Ces modèles, qu'on appellera modèles de cure (*cure models* en anglais), sont utiles pour les situations où pour certaines entités d'observations la durée est très longue ou l'évènement étudié n'est jamais observé. Dans certaines études médicales où les patients sont gardés sous surveillance pour remarquer une éventuelle rechute suite à une maladie, la rechute n'a pas lieu et ces patients sont considérés comme guéris. Dans l'économie du travail, on étudie souvent le retour sur le marché de travail des certaines catégories de travailleurs, par exemple le retour des femmes après un arrêt pour cause de maternité. Il est bien connu qu'une proportion de travailleurs ne retourne plus sur le marché de travail. Dans de telles situations, une question importante est l'estimation de la probabilité conditionnelle d'un temps de vie infini, sachant les variables explicatives. En général, la tâche du statisticien est rendue plus difficile par la présence d'une censure à droite. Par exemple, la surveillance est arrête avant la rechute ou le travailleur sort du marché de travail pour raison d'émigration. Plusieurs types de modèles de cure ont été proposés dans la littérature, le plus souvent la probabilité conditionnelle d'un temps de vie infini est modélisée par une régression logistique. Dans ce travail on pose la question de test de l'adéquation d'une modélisation choisie, logistique ou autre. La difficulté réside dans le fait que pour une observation censurée on ne saura pas préciser s'il elle correspond à un temps de vie fini ou infini. A notre connaissance, aucune solution n'a été proposée pour l'instant. Nous introduisons un test d'adéquation non paramétrique basé sur un lissage à noyau. Ce test est proposé sous des hypothèses minimales d'identification et de conditions techniques habituelles en analyse de survie. Les valeurs critiques asymptotiques du test sont données par la loi normale standard.

Mots-clés. analyse de survie, test d'adéquation en régression, lissage par noyau, $U-$statistiques

Abstract. Cure regression models are a special topic in lifetime analysis. Such models are designed to take into account situations where a proportion of subjects will never experience the event under study. In such a case the lifetime is considered infinite. For instance, medical studies could reveal a proportion of patients for whom the disease under surveillance will never recur, and these patients could be considered as cured. A

well studied topic in Labor Economics is the time to get a new job after a permanent layoff. It is commonly accepted that a proportion of the labor force will withdraw and never get a new job. The crucial issue in cure models is to estimate the conditional probability of an infinite lifetime. In most of the applications the analysis is made more difficult by the presence of a finite random right censorship. Several cure regression models have been considered in the literature and most of them consider a logistic regression for the conditional probability of an infinite lifetime. To our best knowledge, no goodness-of-fit procedure has been proposed yet. The difficulty comes from the fact that it is impossible to know whether a censored observation has a finite or infinite lifetime. In this contribution we propose a kernel smoothing based model check procedure that is able to detect general (nonparametric) alternatives. The assumptions on the lifetime of interest and the censorship are very general and the critical values are given by a standard normal distribution.

Keywords. cure regression models, goodness-of-fit, kernel-smoothing, $U-$statistics

1 Introduction

Cure models are a special topic in survival analysis model designed to model situations where there exists a proportion of subjects who will never experience the event. In this case the time to event is considered infinite. In many applications, it is reasonable to admit that a certain event under study will never occur. For instance, medical studies could reveal a proportion of patients for whom the disease under surveillance will never recur, and these patients could be considered as cured. There is large biostatistical literature that considered this type of model; see for instance Tsodikov, Ibrahim & Yakovlev (2003), Zheng, Yin & Ibrahim (2006) and the references therein. The same models appears in economics and econometrics under the name of split-population models; see Schmidt and Witte (1989).

A crucial aspect in cure models is to estimate the conditional probability of an infinite lifetime. In general the task is made more difficult by the presence of a finite random right censorship. Several cure regression models have been considered in the literature and most of them consider a logistic regression for the conditional probability of an infinite lifetime. It seems that no goodness-of-fit procedure has been proposed yet. The difficulty comes from the fact that it is impossible to know whether a censored observation has a finite or infinite lifetime. In this contribution we propose a kernel smoothing based model check procedure that is able to detect general (nonparametric) alternatives. The procedure is based on the explicit representations of the conditional probability of being cured as functions of the law of the observations; see Patilea & Van Keilegom (2014). These representations are obtained under usual conditional independence assumptions that are required for the identification of the law of the lifetime of interest.

2 A general framework for cure models

Let T denote the lifetime (or time to event) of interest that takes values in $[0, \infty]$. A cured observation corresponds to the event $\{T = \infty\}$, such that in the following this event is allowed to have a positive probability.

Consider the situation where one observes independent copies of Y, δ and X where Y is a nonnegative real-valued random variable, δ is an indicator variable and X is a $d-$dimension covariate vector with support $\mathcal{X}$. The indicator variable reveals whether Y is precisely the lifetime of interest, or Y is only a random quantity smaller than T. In other words $\delta = 1$ if $Y = T$ and $\delta = 0$ if $Y < T$.

The observations are characterized by the conditional sub-probabilities

$$\begin{aligned} H_1([0,t] \mid x) &= \mathbb{P}(Y \leq t, \delta = 1 \mid X = x) \\ H_0([0,t] \mid x) &= \mathbb{P}(Y \leq t, \delta = 0 \mid X = x), \qquad 0 \leq t < \infty,\ x \in \mathcal{X}. \end{aligned}$$

Then $H([0,t] \mid x) = \mathbb{P}(Y \leq t \mid X = x) = H_0([0,t] \mid x) + H_1([0,t] \mid x)$. Since we assume that Y is finite, we have $H([0,\infty) \mid x) = 1,\ \forall x \in \mathcal{X}$. For $j \in \{0,1\}$ and $x \in \mathcal{X}$, let $\tau_{H_j}(x) = \sup\{t : H_j([t,\infty) \mid x) > 0\}$ denote right extreme of the support of the conditional sub-probability H_j. Let us define $\tau_H(x)$ is a similar way and note that $\tau_H(x) = \max\{\tau_{H_0}(x), \tau_{H_1}(x)\}$.

The usual way to model this situation in order to identify and estimate the conditional law T is to consider that there exists a nonnegative random variable C, the right-censoring time, and

$$Y = T \wedge C, \qquad \delta = \mathbf{1}\{T \leq C\}.$$

Here and in the following $\mathbf{1}\{\}$ denotes an indicator function. For $0 \leq t \leq \infty$, let us define the conditional probabilities

$$F_C([0,t] \mid X) = \mathbb{P}(C \leq t \mid X) \quad \text{and} \quad F_T([0,t] \mid x) = \mathbb{P}(T \leq t \mid X),$$

and let $\Lambda_C(\cdot \mid X)$ and $\Lambda_T(\cdot \mid X)$ be the corresponding conditional cumulative hazard measures.

Some identification assumptions are required to be able to identify the conditional law of T from the observations. Let us assume that

$$C \perp T \mid X \qquad \text{and} \qquad \mathbb{P}(C < \infty) = 1. \tag{1}$$

For the sake of simplicity, let us consider also the condition

$$\mathbb{P}(T = C) = 0 \tag{2}$$

Let us point out that for any x the support of $\Lambda_T(dt \mid x)$ and $F_T(\cdot \mid x)$ (resp. $\Lambda_C(dt \mid x)$ and $F_C(\cdot \mid x)$) coincides with the support of $H_1(dt \mid x)$ (resp. $H_0(dt \mid x)$). Moreover, if $\tau_{H_1}(x) < \infty$,

$$\mathbb{P}(T > \tau_{H_1}(x) \mid x) = \prod_{t \in (0, \tau_{H_1}(x)]} \{1 - \Lambda_T(dt \mid x)\},$$

but there is no way to identify the conditional distribution of T beyond $\tau_{H_1}(x)$. Therefore, we will impose $\mathbb{P}(T > \tau_{H_1}(X) \mid X) = \mathbb{P}(T = \infty \mid X)$. Here and in the following, $\prod_t$ denotes the product integral. Finally, we will also assume that $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$ are such that

$$\mathbb{P}(C = \infty \mid X) = \prod_{t \in (0,\infty)} \{1 - \Lambda_C(dt \mid X)\} = 0, \qquad a.s. \tag{3}$$

Let us point out that this condition is satisfied only if $\tau_{H_1}(X) \leq \tau_{H_0}(X) \leq \infty$, a.s. Finally, note that in the case without covariates, $H_j([0,t] \mid X)$ should replaced by $H_j([0,t]) = \mathbb{P}(Y \leq t, \delta = j)$, $j = 0, 1$, in all formulae above.

3 The two-component mixture cure model

Let B be an indicator function for the event T *is not cured*, that is

$$B = \mathbf{1}\{T \text{ is not cured}\} = \mathbf{1}\{T < \infty\}.$$

Define the conditional cumulative hazard measure for the finite values of the lifetime of interest

$$\Lambda_T(dt \mid X, B = 1) = \frac{F_T(dt \mid X, B = 1)}{F_T([t, \infty) \mid X, B = 1)}$$

Patilea & Van Keilegom (2014) showed that

$$\Lambda_T(dt \mid X, B = 1) \;=\; \frac{H_1(dt \mid X)}{H([t, \infty) \mid X) - \mathbb{P}(B = 0 \mid X) F_C([t, \infty) \mid X)}. \tag{1}$$

Let us note that $F_C(\cdot \mid x)$ can be written as a transformation of $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$. Such a representation of $F_C(\cdot \mid x)$ plugged into equation (1), allows to express $\Lambda_T(\cdot \mid x, B = 1)$, and thus $F_T^{(1)}(\cdot \mid x)$ the associated conditional distribution function, as maps of $\mathbb{P}(B = 0 \mid x)$ and the measures $H_0(\cdot \mid x)$ and $H_1(\cdot \mid x)$.

Let

$$\phi(x; \beta) = \mathbb{P}(B = 1 \mid X = x),$$

where $\phi(\cdot; \cdot)$ is a given function. A typical choice that could be found in the literature is

$$\phi(x; \beta) = \frac{\exp(a + x^\top b)}{1 + \exp(a + x^\top b)}$$

where $\beta = (a, b^\top)^\top$ with $a \in \mathbb{R}$ and $b \in \mathbb{R}^d$. For identification purposes we impose the following mild condition: there exists $\mathcal{B} \subset \mathcal{X}$ such that

$$\phi(X; \beta)\mathbf{1}\{X \in \mathcal{B}\} = \phi(X; \widetilde{\beta})\mathbf{1}\{X \in \mathcal{B}\}, \text{ almost surely} \quad \Rightarrow \beta = \widetilde{\beta}.$$

A set $\mathcal{B} \subset \mathcal{X}$ will be helpful to trim the observations in regions of low density of X. Next, let β_0 such that

$$\mathbb{P}(B = 1 \mid x) = \phi(x; \beta_0).$$

Finally, for a fixed value of the parameter β, and $0 \leq t < \infty$, $x \in \mathcal{X}$, let

$$F^{(1)}_{T,\beta}((t, \infty) \mid x) = \prod_{0<s\leq t} \left\{ 1 - \frac{H_1(ds \mid x)}{H([s, \infty) \mid x) - [1 - \phi(x, \beta)] F_C([s, \infty) \mid x)} \right\}.$$

4 Cure regression models check

Let (Y_i, δ_i, X_i), $1 \leq i \leq n$, an independent sample from (Y, δ, X). Consider $\tau > 0$ such that $\tau < \tau_H(x)$ for all $x \in \mathcal{B}$. Let

$$U_i(y, X_i; \beta) = \frac{\delta_i}{F_C([y, \infty) \mid X_i)} \left\{ \mathbf{1}\{T_i < y\} - \phi(X_i; \beta) F^{(1)}_{T,\beta}([0, y] \mid X_i) \right\}.$$

Under our identification conditions, and if the cure regression model is correct,

$$\mathbb{E}[U_i(y, X_i; \beta_0) \mid X_i] = 0 \ \ a.s., \ \ \forall y \in [0, \tau].$$

Let $\widehat{U}_i(y, X_i; \beta)$ be a nonparametric estimation of $U_i(y, X_i; \beta)$ obtained from a nonparametric estimation of H_0 and H_1. Let $\widehat{\beta}$ an estimator of β_0; see Patilea & Van Keilegom (2014). Consider the statistic

$$T_n = \frac{1}{n(n-1)h^d} \sum_{1\leq i\neq j\leq n} \left\langle \widehat{U}_i(\cdot, X_i; \widehat{\beta}), \ \ \widehat{U}_j(\cdot, X_j; \widehat{\beta}) \right\rangle K_{ij}$$

where

$$\left\langle \widehat{U}_i(\cdot, X_i; \widehat{\beta}), \ \ \widehat{U}_j(\cdot, X_j; \widehat{\beta}) \right\rangle = \int_0^1 \widehat{U}_i(s\tau, X_i; \widehat{\beta}) \widehat{U}_j(s\tau, X_j; \widehat{\beta}) ds$$

and

$$K_{ij} = K\left(\frac{X_i - X_j}{h} \right) \mathbf{1}\{X_i, X_j \in \mathcal{B}\},$$

with h a bandwidth.

We show that under suitable technical conditions, for some random sequence μ_n and some positive value σ^2, the quantity $nh^{d/2}(T_n - \mu_n)$ converges in law to a $N(0, \sigma^2)$ variable provided the cure regression $\phi(\cdot; \beta)$ is correctly specified. The bias correction μ_n and the variance σ^2 have to be estimated nonparametrically, following the ideas of Lopez & Patilea (2013). Eventually, a test statistic with standard normal asymptotic critical values is derived.

Bibliographie

[1] Fang, H.B., Li, G., & Sun, J. (2005). Maximum likelihood estimation in a semiparametric Logistic/proportional-hazards mixture model. *Scand. J. Statist.* **32**, 59–75.
[2] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046.
[3] Kuk, A.Y.C. & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531–541.
[4] Lopez, O. & Patilea, V. (2013). Testing conditional moment restrictions in the presence of right-censoring depending on the covariates. Working Paper.
[5] Maller, R. A. & Zhou, S. (1996). *Survival analysis with long term survivors.* New York: Wiley.
[6] Patilea, V. & Van Keilegom, I. (2014). A general approach for cure models in survival analysis. Working Paper.
[7] Schmidt, P. & Witte, A.D. (1989). Predicting Criminal Recidivism Using Split Population Survival Time Models. *J. Econometrics* **40**, 141–159.
[8] Tsodikov, A.D., J. G Ibrahim, J.G. & Yakovlev, A.Y. (2003). Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models. *J. Amer. Statist. Assoc.* **98(464)**, 1063–1078.
[9] Zheng, D., Yin, G. & Ibrahim, J.G. (2006). Semiparametric Transformation Models for Survival Data With a Cure Fraction. *J. Amer. Statist. Assoc.* **101(474)**, 670–684.