VI Congreso Galego de Estatística e Investigación de Operacións
Vigo 5–7 de Novembro de 2003

PRESMOOTHED CENSORED REGRESSION

de Uña-Álvarez, Jacobo[1], Rodríguez-Campos, M. Celia[1]

[1]Department of Statistics and OR
University of Vigo

ABSTRACT

It is known that Kaplan-Meier estimation may be improved *via* presmoothing methods. In this work we consider presmoothed least squares estimation for a regression parameter, when the response is subject to right-censoring. The approach is that in de Uña-Álvarez (2002). Properties of the proposed estimators are investigated *via* simulations.

Key words: Censoring, Kaplan-Meier, Least Squares, Presmoothing, Regression

1. INTRODUCTION

In Survivial Analysis, one is often interested in a time (hence nonnegative) response, say Y, which is subject to censoring from the right. If $X = (x^1, ..., x^p)^t$ is a vector of covariates, the basic regression model is represented as

$$\ln Y = f(X; \theta_0) + \varepsilon, \tag{1}$$

where $f(.; \theta_0)$ is some smooth function which depends on the "true" (unknown!) regression parameter $\theta_0 \in \Theta \subset \mathbb{R}^d$, and ε is a (usually zero-mean) error term. Due to the log-transformation for the Y in (1), the function $f(.; \theta_0)$ is not restricted to take positive values. The accelerated failure time model, very important in applications, is obtained as a special case of (1).

As a result of right-censoring effects, rather than a random sample of (X, Y), one is just able to observe $(X_1, Z_1, \delta_1), ..., (X_n, Z_n, \delta_n)$, iid data with the same distribution as (X, Z, δ). Here, $Z = \min(Y, C)$ is the recorded time, C is the censoring variable, and $\delta = 1_{\{Y \leq C\}}$ stands for a censoring indicator. In censored regression, the goal is the estimation of θ_0 from the (X_i, Z_i, δ_i); several approaches have been investigated to this aim.

A useful model, introduced by Stute (1993), is that based on the assumptions

H1. Y and C are independent;

H2. δ and X are independent conditionally on Y.

Assumption H1 is typical in censored scenarios, while H2 incorporates the covariate vector in a sensible way. See Stute (1993, 1996, 1999) for further discussion. Note that H1-H2 hold whenever C is independent of (X, Y). Under H1-H2, weighted LSE of θ_0 is defined through minimization of

$$\theta \mapsto \sum_{i=1}^{n} W_{i,n}^{KM} \left[\ln Z_{i:n} - f(X_{[i:n]}; \theta)\right]^2 \tag{2}$$

where $Z_{1:n} \leq ... \leq Z_{n:n}$ are the ordered Z_i,

$$W_{i,n}^{KM} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\delta_{[j:n]}}{n-j+1} \right] \tag{3}$$

is the jump size of the Kaplan-Meier estimator at $Z_{i:n}$, and $\left(X_{[i:n]}, \delta_{[i:n]}\right)$ stands for the (X, δ) vector attached to $Z_{i:n}$. Note that $W_{i,n}^{KM} = 1/n$ in the uncensored case. Consistency and distributional convergence for this weighted LS criterion was established in Stute (1993, 1996) (for a linear $f(.;\theta_0)$) and Stute (1999) (nonlinear case).

Set $\widetilde{m}(z) = P(\delta = 1 \mid Z = z)$, the conditional probability of uncensoring given $Z = z$, and put $\widetilde{m}_n(z)$ for an estimator of $\widetilde{m}(z)$ based on the (Z_i, δ_i)'s. Improved estimation of the marginal df of Y is obtained by substituting $\widetilde{m}_n(Z_{i:n})$ for $\delta_{[i:n]}$ in (3), provided that the smoother $\widetilde{m}_n$ is properly chosen. "Presmoothing the Kaplan-Meier" just means that such a substitution has taken place. Presmoothing ideas were introduced by Dikta (1998) under a parametric assumption on $\widetilde{m}$. Jácome and Cao (2002) provide important results when using a nonparametric fit to $\widetilde{m}$. However, the presence of covariates was not considered in these works.

Presmoothed Kaplan-Meier estimation with covariates was investigated in de Uña-Álvarez (2002). Introduce the presmoothed Kaplan-Meier weights as

$$W_{i,n}(m_n) = \frac{m_n(X_{[i:n]}, Z_{i:n})}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m_n(X_{[j:n]}, Z_{j:n})}{n-j+1} \right],$$

where $m_n(x,z)$ stands for an estimator of $m(x,z) = P(\delta = 1 \mid X = x, Z = z)$ based on the available (X_i, Z_i, δ_i). The associated weighted LSE of θ_0, say θ_n, is defined as the minimizer of

$$\theta \mapsto \sum_{i=1}^{n} W_{i,n}(m_n) \left[\ln Z_{i:n} - f(X_{[i:n]}; \theta) \right]^2. \tag{4}$$

The connection between the weights in (4) and those in (2) is that $W_{i,n}(m_n)$ collapses to $W_{i,n}^{KM}$ when severely undersmoothing the estimator m_n.

The consistency of θ_n follows from that of general integrals $\int \varphi d\widehat{F}_{X,Y}^{PKM}$, where

$$\widehat{F}_{X,Y}^{PKM}(x,y) = \sum_{i=1}^{n} W_{i,n}(m_n) 1_{\left\{ X_{[i:n]} \leq x, Z_{i:n} \leq y \right\}}$$

is an estimator of the multivariate df of (X, Y). Assumptions H1-H2 are crucial to this aim. Of course, some assumption on the smoother m_n is needed too. Roughly speaking, uniform strong convergence of m_n is enough for guaranteeing $\theta_n \to \theta_0$ almost surely. See de Uña-Álvarez (2002) for details.

In the present work, we investigate *via* simulations the performance of the presmoothed weighted LSE θ_n. Comparison with the minimizer of (2) is included. Parametric and nonparametric (kernel) estimates for m will be considered, leading to parametric and nonparametric presmoothed estimators for θ_0, respectively. As in previous works without covariates, it will be seen that presmoothing leads in general to better estimates. Moreover, presmoothing may result in fairly good estimators even when the parametric model (resp. the bandwidth) for m is not correctly chosen. We report some of these simulations in Section 2.

There exists some previous research on presmoothing methods with covariates under the alternative (conditional) model

H*. Y and C are independent conditionally on X.

Veraverbeke and Cadarso-Suárez (2000) considered the case in which $m(x,z)$ is free of z, providing a conditional version of the Koziol-Green censorship model. Nonparametric estimation of the conditional df of the Y was investigated in the referred paper. Also, Yuan (2002) introduced parametric presmoothing under H*, and proposed efficient estimation of the regression parameter in the scope of the Cox model. Model assumptions H1-H2 considered in this note lead to a quite different approach for regression.

2. SOME SIMULATIONS

We have considered a basic regression model as (1), where $f(x;\theta_0) = -x^t\theta_0$ is a linear function, and the (normalized) error term ε/σ follows an extreme value distribution, independent of the covariate vector. As a result, the conditional df of Y given $X = x$ is $Weibull(\alpha_1(x), \beta_1)$, with $\alpha_1(x) = \exp(x^t\theta_0)$ and $\beta_1 = 1/\sigma$. We have considered a single, real-valued covariate X, uniformly distributed on the unit interval. The resulting regression model is

$$\ln Y = -\theta_{10} - \theta_{20}X + \varepsilon.$$

Censoring was introduced following a $Weibull(\alpha_2, \beta_2)$ distribution, with $\alpha_2 = \exp(\theta_{10})$ and $\beta_2 = \beta_1$, independent of (X,Y). In this situation, the function m follows the logistic specification

$$m(x,z) = \frac{\exp(\beta_1\theta_{20}x)}{1+\exp(\beta_1\theta_{20}x)}. \tag{5}$$

The proportion of uncensored responses results in

$$E(\delta) = \frac{1}{\beta_1\theta_{20}} \ln\left[\frac{1+\exp(\beta_1\theta_{20})}{2}\right].$$

We have considered the cases $\beta_1 = 1$ and $\beta_1 = 2$. Given β_1, the value of the slope θ_{20} was fixed in order to obtain five censoring percentages, about 10, 28, 45, 50 and 62% of censoring. The intercept θ_{10} was chosen to be zero. Sample sizes 25, 50, 100 and 200 were considered. For each case, 1,000 trials were performed.

Mean squared errors (MSE) of estimators for $\theta_0 = (\theta_{10}, \theta_{20})^t$ along the 1,000 trials were computed. We considered three estimation methods. First, the minimizer of the weighted LS criterion (2) was computed. This involves the ordinary Kaplan-Meier weights. Then, parametric presmoothing was considered, under the logistic assumption

$$m(x,z;\beta_0) = \frac{\exp(\beta_{10}+\beta_{20}x+\beta_{30}z)}{1+\exp(\beta_{10}+\beta_{20}x+\beta_{30}z)}. \tag{6}$$

Note that the true m given in (5) belongs to this parametric family. In this case, the smoother m_n was chosen as the maximizer of the conditional likelihood

$$\mathcal{L}_n(\beta) = \prod_{i=1}^{n} m(X_i,Z_i;\beta)^{\delta_i}\left[1-m(X_i,Z_i;\beta)\right]^{1-\delta_i}. \tag{7}$$

Finally, nonparametric presmoothing was introduced. Nadaraya-Watson type regression with (bivariate) product kernel function was considered to this aim. The Epanechnikov kernel with smoothing parameter $h = 0.3$ was used in all the cases. Both parametric and nonparametric smoothers for m were used in order to get presmoothed LSE for θ_0 *via* (4). The results for the estimator of the slope θ_{20} are displayed in Table 1. Empty cells correspond to situations in which the maximization of (7) gave no solution. The intercept estimator behaved quite similarly, and the corresponding results will be reported elsewhere.

		$\beta_1 = 1$			$\beta_1 = 2$		
n	CP	KM	PP	NP	KM	PP	NP
25	10%	-	-	-	-	-	-
	28%	-	-	-	-	-	-
	45%	1.8159	1.0189	1.6014	0.5386	0.2642	0.4306
	50%	2.1077	1.0075	1.7839	0.5274	0.2930	0.4447
	62%	3.9718	1.4843	3.0948	0.9365	0.3643	0.7255
50	10%	-	-	-	-	-	-
	28%	0.6703	0.5258	0.6173	0.1730	0.1279	0.1513
	45%	0.9585	0.4888	0.8318	0.2522	0.1301	0.2117
	50%	1.1448	0.5137	1.0160	0.2858	0.1373	0.2361
	62%	1.9766	0.8892	1.7608	0.4681	0.2069	0.3896
100	10%	0.2499	0.2507	0.2731	0.0641	0.0611	0.0685
	28%	0.3670	0.2678	0.3249	0.0854	0.0622	0.0751
	45%	0.5425	0.2821	0.5029	0.1323	0.0756	0.1124
	50%	0.6264	0.2812	0.5890	0.1592	0.0772	0.1311
	62%	1.2059	0.5206	1.1134	0.2773	0.1310	0.2349
200	10%	0.1118	0.1085	0.1323	0.0306	0.0284	0.0358
	28%	0.1706	0.1370	0.1596	0.0397	0.0298	0.0362
	45%	0.2943	0.1569	0.2710	0.0758	0.0409	0.0619
	50%	0.3198	0.1692	0.2898	0.0878	0.0401	0.0735
	62%	0.7152	0.3512	0.6966	0.1870	0.0885	0.1593

Table 1. Mean squared error of LSE for the slope along 1,000 trials, with ordinary Kaplan-Meier (KM), parametric presmoothed (PP), and nonparametric presmoothed (NP) weights.

From Table 1 it can be seen that the MSE decreases as the sample size increases. Also, the error is increasing as a function of the censoring percentage (CP). When comparing the three estimation techniques, it should be noted that parametric presmoothing results in a lower MSE (this is particularly true under heavy censoring). Importantly, nonparametric presmoothing outperforms in general ordinary Kaplan-Meier LS, *even* when the bandwidth for m_n is not correctly chosen (it should be a decreasing function of n). Hence, wrong presmoothing still may lead to sensible estimators.

This robustness property of nonparametric presmoothing is shared by its parametric analogue (as previously noted by other authors). In order to illustrate this, Table 2 reports (for sample size $n = 100$) the MSE of the slope estimators considered above, when the censoring variable is generated according to a uniform distribution on $(0, 3)$. In this case, the function m is no longer (5), rather being equal to

$$m(x, z) = \frac{\exp(\beta_1 \theta_{20} x)}{(\beta_1 z^{\beta_1 - 1}(3 - z))^{-1} + \exp(\beta_1 \theta_{20} x)}.$$

Hence, parametric presmoothing in initially misleaded by the (untrue) logistic specification (6). Given values for β_1 and θ_{20}, the censoring proportion is obtained as

$$\frac{\Gamma(1/\beta_1)}{3\beta_1^2\theta_{20}} \int_1^{\exp(\beta_1\theta_{20})} x^{-1-1/\beta_1}(1-e^{-3x^{1/\beta_1}})dx$$

if $\theta_{20} \neq 0$, and by

$$\frac{1}{3}\int_0^3 e^{-x^{\beta_1}}\,dx$$

if $\theta_{20} = 0$. As a result, for the special values of these parameters in Table 2, the corresponding censoring percentages are about 5, 14, 27, 32 and 46% for $\beta_1 = 1$, and about 8, 18, 26, 30, and 34% for $\beta_1 = 2$. As in Table 1, we see that the parametric presmoothed LSE is the best estimator (in the sense of the MSE). However, we observed that ordinary Kaplan-Meier weights do it slightly better than presmoothed ones when estimating the intercept. Further discussion and more simulations will be reported elsewhere.

	$\beta_1 = 1$			$\beta_1 = 2$		
$\beta_1\theta_{20}$	KM	PP	NP	KM	PP	NP
7	-	-	-	-	-	-
2	0.2340	0.2030	0.2191	0.0616	0.0572	0.0576
.4	0.3192	0.2209	0.2930	0.0704	0.0644	0.0692
0	0.3595	0.2170	0.3247	0.0823	0.0695	0.0820
-1	0.7349	0.4229	0.6692	0.1032	0.0770	0.1001

Table 2. Mean squared error of LSE for the slope along 1,000 trials of size $n = 100$, with uniform censoring distribution.

3. ACKNOWLEDGEMENTS

First author was supported by the Grants PGIDIT02PXIA30003PR and BFM2002-03213.

4. REFERENCES

Dikta, G. (1998). "On semiparametric random censorship models". *Journal of Statistical Planning and Inference* 66, 253-279.

Jácome, M. A. and Cao, R. (2002). "Presmoothed kernel density estimator for censored data". Unpublished manuscript.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis* 45, 89-103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* 23, 461-471.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica* 9, 1089-1102.

de Uña-Álvarez, J. (2002). Strong consistency of presmoothed Kaplan-Meier integrals when covariables are present. Unpublished manuscript.

Veraverbeke, N. and Cadarso-Suárez, C. (2000). Estimation of the conditional distribution in a conditional Koziol-Green model. *Test* 9, 97-122.

Yuan, M. (2002). Semiparametric censorship model with covariates. Unpublished manuscript.