Comparison among Regression Methods for Censored Data

Liciana Vaz de A. S. Chalita
Instituto de Biociências
UNESP – Botucatu -São Paulo
CEP: 18618-000
Brazil

José Eduardo Corrente
E.S.A. "Luiz de Queiroz"
USP – Piracicaba – São Paulo
CEP: 13418-900
Brazil

ABSTRACT

In censored data analysis, it is common the use of nonparametrical techniques such as Kaplan-Meier as well as parametric models. When we have covariates, semi-parametrical models can be used like, for instance, Cox model. Another approach when we have covariates, is the use of linear regression models, suggested by Miller (1976).

Miller's method consists in fitting a linear function for the censored data set via least squares. If we use weighted least squares, where the weights are the jumps of the Kaplan-Meier function, this method is called Miller Modified method and it does not always give a good fit.

In this work, we introduce a further modification in the Modified Miller's method, changing the quadratic function by a ψ-Huber robust function in order to get a better fit.

Empirical comparison among Miller, Miller Modified and Cox's methods were made through simulated data from a Weibull distribution and an application to the data of the Stanford Heart Transplant was made. Also, a routine in S-Plus language was written for fitting such models. Discussions were made for the robust model fits and for the Miller Modified by analyzing the mean square error, and it was suggested as an alternative method for Cox model.

KEY WORDS AND PHRASES: survival analysis, Cox model, Miller model, ψ-Huber function.

1. Introduction

According to Miller (1981), survival analysis is a statistical term that encompasses a variety of statistical techniques for analyzing positive-valued random variables. Typically, the value of the random variable is the time to failure of a physical component or the time to death of a biological unit. The difference with another fields of statistics is censoring, i.e., the observation contains only partial information about the random variable of interest. So, let $T_1, \cdots, T_n$ be independent and identically

Comparison among Regression Methods for Censored Data

Liciana Vaz de A. S. Chalita
Instituto de Biociências
UNESP – Botucatu -São Paulo
CEP: 18618-000
Brazil

José Eduardo Corrente
E.S.A. "Luiz de Queiroz"
USP – Piracicaba – São Paulo
CEP: 13418-900
Brazil

ABSTRACT

In censored data analysis, it is common the use of nonparametrical techniques such as Kaplan-Meier as well as parametric models. When we have covariates, semi-parametrical models can be used like, for instance, Cox model. Another approach when we have covariates, is the use of linear regression models, suggested by Miller (1976).

Miller's method consists in fitting a linear function for the censored data set via least squares. If we use weighted least squares, where the weights are the jumps of the Kaplan-Meier function, this method is called Miller Modified method and it does not always give a good fit.

In this work, we introduce a further modification in the Modified Miller's method, changing the quadratic function by a ψ-Huber robust function in order to get a better fit.

Empirical comparison among Miller, Miller Modified and Cox's methods were made through simulated data from a Weibull distribution and an application to the data of the Stanford Heart Transplant was made. Also, a routine in S-Plus language was written for fitting such models. Discussions were made for the robust model fits and for the Miller Modified by analyzing the mean square error, and it was suggested as an alternative method for Cox model.

KEY WORDS AND PHRASES: survival analysis, Cox model, Miller model, ψ-Huber function.

1. Introduction

According to Miller (1981), survival analysis is a statistical term that encompasses a variety of statistical techniques for analyzing positive-valued random variables. Typically, the value of the random variable is the time to failure of a physical component or the time to death of a biological unit. The difference with another fields of statistics is censoring, i.e., the observation contains only partial information about the random variable of interest. So, let $T_1, \cdots, T_n$ be independent and identically

distributed (i.i.d.) random variables with distribution function F and let $C_1,\cdots,C_n$ be i.i.d. random variables with distribution function G. C_i is a censoring time associated with T_i, for $i=1,\cdots,n$. In this case, we only observe $(Y_1,\delta_1),\cdots,(Y_n,\delta_n)$, where

$$Y_i = min(T_i, C_i)$$

$$\delta_i = \begin{cases} 1 & if \quad T_i \le C_i \\ 0 & if \quad T_i > C_i \end{cases}$$

for $i=1,\cdots,n$. Notice that $Y_1,\cdots,Y_n$ are i.i.d. with some distribution function H. Also, $\delta_1,\cdots,\delta_n$ contain the censoring information, that is, if $\delta_i=1$, T_i is not censored. Otherwise, T_i is censored.

Nonparametric techniques, like the product-limit method, developed by Kaplan-Meier (1958), as well as parametric techniques, have been used in order to analyse these sort of data set. Also, in general, it is possible to consider covariates related to response variable. The semi-parametric model proposed by Cox (1972) can be used. The proportional hazard model assumes:

$$h(t,x) = h_0(t)\, e^{x'\beta}$$

where β is an unknown vector of p regression coefficients, x is a p x n matrix of covariates and $h_0(t)$ is an unknown positive function. It can be noted, that, when $x = 0$, $h(t,x) = h_0(t)$.

In the same sense, Miller (1976) proposed a standard linear mode for the survival time, where its expected value could be for a linear regression function of the covariates, i.e.

$$E(T_i) = \alpha + \beta' x_i$$

for $i=1,\cdots,n$.

For simplicity, we are going to assume $p = 1$. When censoring is present, Miller (1976) proposed to minimize

$$n\int_{-\infty}^{+\infty} z^2\, d\hat{F}(z) = \sum_{i=1}^{n} \hat{w}_i(\beta)(y_i - \alpha - \beta x_i)^2$$

where $\hat{F}$ is the Kaplan-Meier product-limit based on $(z_1,\delta_1),\cdots,(z_n,\delta_n)$, $z_i = y_i - \alpha - \beta x_i$, and $\hat{w}_1(\beta),\cdots,\hat{w}_n(\beta)$ are the jumps of the product-limit estimator. Notice that if $\delta_i = 0$, corresponding to a censored observation, then $\hat{w}_i(\beta) = 0$. So, at first glance, the weighted sum of squares does not

depend on the censored observation. If $\delta_{(n)}=0$, that is, the last ordered observation is censored, change it to be uncensored. Then $\sum_{i=1}^{n}\hat{w}_i(\beta)=1$.

In order to calculate $\hat{\alpha}$ and $\hat{\beta}$, we differentiate the above function with respect to α and β. To estimate β, we must use a search procedure. Miller (1981) suggests the following modified approach. Define the initial estimate

$$\hat{\beta}^0=\frac{\sum_u y_i(x_i-\bar{x}_u)}{\sum_u (x_i-\bar{x}_u)^2}$$

which is the slope of the least-squares line through the uncensored observations. With this guess $\hat{\beta}^0$, form

$$\hat{z}_i^0=y_i-\hat{\beta}^0 x_i \text{ for } i=1,\cdots,n.$$

Let $\hat{F}^0$ be the product-limit estimator based on $(\hat{z}_1^0,\delta_1),\cdots,(\hat{z}_n^0,\delta_n)$, and let $\hat{w}_1(\hat{\beta}^0),\cdots,\hat{w}_n(\hat{\beta}^0)$ be the jumps of $\hat{F}^0$. Now, define the new estimate

$$\hat{\beta}^1=\frac{\sum_u \hat{w}_i^*(\hat{\beta}^0)y_i(x_i-\bar{x}_u^*)}{\sum_u \hat{w}_i^*(\hat{\beta}^0)(x_i-\bar{x}_u^*)^2}$$

where

$$\hat{w}_i^*(\hat{\beta}^0)=\frac{\hat{w}_i(\hat{\beta}^0)}{\sum_u \hat{w}_i(\hat{\beta}^0)}$$

and

$$\bar{x}_u^*=\sum_u \hat{w}_i^*(\hat{\beta}^0)x_i$$

for $i=1,\cdots,n$. Using the renormalized weights $\hat{w}_i(\hat{\beta}^0)$, allows us to ignore whether the last ordered $\hat{z}_i^0$ is censored or not. The procedure continues until convergence, which does it is not take place necessarily. It can be expected that the estimate of β may become trapped in a loop oscillating between two values. In such case, we take the average of the two vales. This procedure does not always provide good results for the linear model. Thus, we propose a modification of this method, using a ψ-Huber robust function, given by Bustos (1981). In Section 2, we present the theoretical development, in order to find estimators for α and β for the straight line. In Section 3, we present some results applying the

suggested methodology for a simulated data form Weibull distribution and an application using the Stanford Heart Transplant Data, from 1967 to 1976. Discussions were made in Section 4.

2. Methodology

As is was seen in the previous section, Miller (19810 provides an algorithm to fit a linear model for survival time when censoring is present, by minimizing a weighted sum of squares. The basic idea here is to replace the quadratic function in the sum by a robust ψ-Huber function, in order to get a better fit for the linear model, as we can see in Corrente (1984). Then we proposed to minimize

$$S(\alpha,\beta)=\sum_{i=1}^{n}\hat{w}_i(\hat{\beta})\psi\left(\frac{y_i-\alpha-\beta_1 x_{i1}-\cdots-\beta_p x_{ip}}{\sigma(\alpha,\beta)}\right)$$

where $\sigma(\alpha,\beta)=median\{|y_i-\alpha-\beta_1 x_{i1}-\cdots-\beta_p x_{ip}|\}$ / var(y) , and ψ is a Huber function given by

$$\psi(t)=\begin{cases}\dfrac{t^2}{2} & if \quad |t|\le k\\ k|t|-\dfrac{k^2}{2} & if \quad |t|>k\end{cases}$$

Considering $p = 1$ for simplicity, we have

$$S(\alpha,\beta)=\sum_{i=1}^{n}\hat{w}_i(\hat{\beta})\psi\left(\frac{y_i-\alpha-\beta x_i}{\sigma(\alpha,\beta)}\right)$$

To find $\hat{\alpha}$ and $\hat{\beta}$, we differentiate the function $S(\alpha,\beta)$ with respect to α and β, getting the following system of equations

$$\begin{cases}\sum_{i=1}^{n}\hat{t}_i(\hat{\alpha},\hat{\beta})(y_i-\hat{\alpha}-\hat{\beta}x_i)=0\\ \sum_{i=1}^{n}\hat{t}_i(\hat{\alpha},\hat{\beta})x_i(y_i-\hat{\alpha}-\hat{\beta}x_i)=0\end{cases}$$

where

$$\hat{t}_i(\hat{\alpha},\hat{\beta}) = \frac{\hat{w}_i^*(\hat{\beta})\,\psi\left(\dfrac{y_i-\hat{\alpha}-\hat{\beta}x_i}{\sigma(\hat{\alpha},\hat{\beta})}\right)}{y_i-\hat{\alpha}-\hat{\beta}x_i}$$

and to solve it, we need a search procedure. So, if we consider $\hat{\alpha}_0$ and $\hat{\beta}_0$ as initial values given by the regression line under the uncensored observations respectively, we found

$$\hat{\alpha}_1 = \frac{\sum_u \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)\,y_i - \hat{\beta}^0\sum \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)\,x_i}{\sum \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)}$$

$$\hat{\beta}_1 = \frac{\sum_u \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)\,x_i\,y_i - \hat{\alpha}^0\sum_u \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)\,x_i}{\sum \hat{t}_i^*(\hat{\alpha}^0,\hat{\beta}^0)\,x_i^2}$$

The process continues until convergence, which, as in Miller's procedure, may not take place.

3. Results

In order to compare the proposed procedure and compare it with Miller and Cox's methods, a simulation was performed considering the Weibull distribution with parameter α and γ for the survival time and uniform $[0,\theta]$ distribution for the censored time. It was taken $\theta = 4$, representing 35% of censored time in the dataset. The explanatory variable was generated from as uniform [0,1] distribution. The α parameter was considered as $\exp(x\,\beta)$, where $\beta = -0{,}5$ and $\gamma = 0{,}25$ and 50 replications. Comparisons were made calculating the standard deviation of the slope, the mean square error and the residual sum squares. As the proposed procedure take in account robustness, data with 2% of outliers were also generate in order to verify the fitness of the linear model compared with other methods. The results are shown in Table 1 and 2. The ψ-Huber function depends on the value of k as we can see in the definition of this function, given in the previous section. Thus, it was considered $k = 1{,}345$ and $k=1{,}585$. The first value is suggested when normality of the data is presented, that is not the case and the second value was tested among several values of k providing better fit.

Table 1. Comparison among Miller, Cox and the Robust Method for simulated data from Weibull distribution for n=50,200 and 500, without outliers

n	Method	Intercept	Slope	s.d.(slope)	RSS	MSE
50	Miller Modified	0,8012	-0,2836	0,2688	8,9913	0,1191
	Robust k=1,345	0,7432	-0,4297	0,2983	7,5412	0,0927
	Robust k=1,585	0,7325	-0,4316	0,2855	7,4992	0,0862
	Cox	-	-0,4342	0,3361	20,7377	0,1173
200	Miller Modified	0.9208	-0,5881	0,1037	27,4933	0,0154
	Robust k=1,345	0,9320	-0,5261	0,0941	24,9312	0,0095
	Robust k=1,585	0,9314	-0,5222	0,0913	24,5614	0,0088
	Cox	-	-0,5170	0,1025	96,2317	0,0108
500	Miller Modified	0,8542	-0,1741	0,0740	73,2578	0,1117
	Robust k=1,345	0,8614	-0,4391	0,0630	72,2136	0,0077
	Robust k=1,585	0,8598	-0,4425	0,0619	72,1465	0,0071
	Cox	-	-0,5420	0,1870	254,8389	0,0367

Table 2. Comparison among Miller, Cox and the Robust Method for simulated data from a Weibull distribution for n=50,200 and 500, with 2% of outliers

n	Method	Intercept	Slope	s.d.(slope)	RSS	MSE
50	Miller Modified	0,9499	-0,7951	0,5198	31,2908	0,3573
	Robust k=1,345	0,4125	-0,3453	0,2257	15,9018	0,0749
	Robust k=1,585	0,7325	-0,3498	0,2234	15,8721	0,0725
	Cox	-	-0,4972	0,5312	65,2314	0,2822
200	Miller Modified	0,1496	-0,5696	0,2904	36,7275	0,0892
	Robust k=1,345	0,1311	-0,3444	0,2082	12,6764	0,0676
	Robust k=1,585	0,1298	-0,3401	0,2035	12,5878	0,0670
	Cox	-	-0,6027	0,2745	188,6970	0,0859
500	Miller Modified	0,1053	-0,3789	0,1583	104,1732	0,0397
	Robust k=1,345	0,0945	-0,4646	0,0687	95,3146	0,0060
	Robust k=1,585	0,0926	-0,4552	0,0620	94,8718	0,0059
	Cox	-	-0,5622	0,1596	425,7184	0,0293

As a numerical example, we apply the proposed procedure to the Stanford Heart Transplant data, found in Miller (1981). Here the dependent variable is taken as logarithm to base 10 of the survival time associated to the mismatch score covariate (T5), that measure the degree of dissimilarity between donor and recipient tissue with respect to HL-A antigens, and it is therefore potentially related to the phenomenon of rejection of the donor heart by the recipient's mechanism. The results are shown on Table 3.

Table 3. Comparison among Miller, Cox and the Robust Method for Stanford Heart Transplant data.

Method	Intercept	Slope	s.d.(slope)
Miller Modified	2,092	-0,01299	0,04022
Robust k=1,345	2,2061	-0,02677	0,008031
Robust k=1,585	2,1925	-0,02583	0,008089
Cox	-	0,0645	0,047

4. Discussion

According to the results shown in Tables 1 and 2, the behavior of the robust method, proposed as another modification of Miller's method, is better providing a lower mean square error than Miller Modified and Cox, doesn't mattering whether outlier is present or not. The value of the estimated slope is quite near of the real one and, of course, this become more evident when there is an increasing of the sample size. On the other hand, there is an increasing in the residual sum of squares in the presence of outliers for any sample size, as it was expected.

It was also noted so many convergence problems with Miller Modified method. Most of the estimated values for the slope were obtained by mean of two or three values provided for the method, that is, no convergence was verified. This sort of problem was not found when robust function was used.

For the Stanford Heart Transplant data, we can see that the robust fit provide estimators a bit different from Miller Modified with lower standard deviation for the slope, given in Table 3. It is important to note again that, in Miller Modified, the value of the slope is the mean of two values, once it did not get convergence in this method. On the other hand, the robust method converges all the time, which represents an advantage related to Miller Modified.

In order to get the fitted model for Miller Modified, a routine in S-Plus language was developed, while for the robust proposed method, a robust function developed by S-Plus was used and it was much more convenient for this sort of methodology. All the programs are available by the authors.

In this way, we can conclude that the proposed modification in Miller's method might be an alternative by the Cox's method when Weibull distribution in the dataset is verified.

References

1. KAPLAN, E.L. & MEIER, P. Nonparametric estimation from incomplete observation, *J.Am. Statist. Assoc.*, 53, 475-481, 1958.
2. BUSTOS, O. *Estimação Robusta no Modelo de Posição*. 13°. Colóquio Brasileiro de Matemática, Poços de Caldas – MG, 135p. 1981.
3. CORRENTE, J.E. *Regressão com Dados Censorados*. Dissertação de Mestrado, Instituto de Matemática Pura e Aplicada/CNPq – Rio de Janeiro, 53p, 1984.
4. CHALITA, L.V.A.S. *Inferência Paramétrica e Não Paramétrica em Análise de Sobreviência.* Dissertação de Mestrado, Escola Superior de Agricultura "Luiz de Queiroz". Universidade de São Paulo – Piracicaba – SP, 110p, 1992.
5. MILLER, R.G. *Survival Analysis*, J.Wiley, New York, 238p, 1981.
6. MILLER, R.G. Least squares regression with censored data. *Biometrika*, 63: 3, 449-464, 1976.
7. S-Plus for Windows v.3.2 – *User's Manual*, vol. 1 and 2. Statistics Science Inc, Seattle, Washington, 1993.

ACKNOWLEDGMENTS

The authors would like to thank to the referees for the relevant observation and for the many contributions to improve the final version of this paper.