ZERO-INFLATED SEMI-PARAMETRIC COX'S REGRESSION MODEL FOR LEFT-CENSORED SURVIVAL DATA.

Yves Grouwels and Roel Braekers

Institute for Biostatistics and statistical Bioinformatics
Universiteit Hasselt and Katholieke Universiteit Leuven
Agoralaan 1, 3590 Diepenbeek, Belgium
(e-mail: yves.grouwels@uhasselt.be)

ABSTRACT. In this paper, we introduce a semi-parametric regression model for left-censored data in which the response variable has a positive discrete probability at the value zero. To investigate the influence of covariates on the probability on a zero-value, a logistic regression model is used. For the strict positive part of the response variable, a Cox's regression model is given to model the influence of the covariates. The different parameters in the model are estimated using a likelihood method. Hereby, the baseline hazard function is an infinite dimensional parameter and is estimated by a step-function. As results, we show the consistency of the estimators for the different finite- and infinite-dimensional parameters in the model. We also present a simulation study and apply this model on a practical data example.

1 INTRODUCTION

In some clinical, environmental or industrial studies, the primary interest is in a positive random variable. For example, the amount of a toxic metal in a certain river system (Blackwood (1991)). Hower, due to technical limitations, there are often difficulties in measuring this positive variable. For some subjects, we only observe an upper bound for the response variable. These observations are left-censored. In several studies with left-censored data, the underlying time until an event can also become zero. For example, in an environmental study where researchers are interested in the amount of a certain toxic metal in an aquatic system, it is possible that the metal is not present in the system. As a second example, we consider a biological study on ethanol induced sleeping time in genetically selected mice (Markel *et al.* (1995)). Some mice did not fall asleep because their genetic metabolism was able to break down the alcohol fast. In such studies, it is possible to distinguish between two groups of study subjects. On the one hand there are study subjects which have a strict positive value for the time until an event (susceptible), while on the other hand there are study subjects for which the time until an event is equal to zero (non-susceptible).

In order to describe this problem mathematically, one assumes that the time until an event has a mixture distribution in which there is a continuous part for the strict positive values and a discrete probability for a zero value. Since the observed data in these studies are left-censored, it is not possible to fully discriminate between the groups of susceptible and non-susceptible subjects. The uncensored observations are susceptible subjects, but for the censored observations one cannot distinguish between unsusceptible subjects and susceptible subjects with a censored time until an event.

Moulton and Halsey (1995) developed a regression model to study the influence of covariates on the time until an event for this type of left-censored data. Hereto they assumed a parametric logistic regression model to determine the influence of the covariates on the discrete probability of a zero value for the time until an event. On the other hand they assumed that the distribution of the strict positive values for the time until an event was given by a lognormal distribution in which the influence of the covariates was described through the mean of this distribution. They also assumed that each subject in their model had the same censoring time by considering a fixed detection limit. Recently, Yang and Simpson (2010) studied computational issues regarding a general class of parametric left-inflated mixture models.
The structure of this paper is as follows. In Section 2, we introduce mathematically the zero-inflated Cox's regression model for left censored data. To estimate the different parameters in our model, we make use of maximum likelihood techniques. We can proof the consistency of the MLE's under some regularity conditions. In Section 3, a simulation study is presented. Afterwards, in Section 4, we illustrate our model on a practical data set of ethanol-induced sleep time in mice. In Section 5, we give some conclusions about our results.

2 METHODOLOGY

In this section, we introduce a zero-inflated semiparametric Cox's regression model for left-censored data. Let us denote by Y a nonnegative response variable of interest. We assume that this variable Y has a zero-inflated mixture distribution with a positive probability of having a value equal to zero and with a continuous distribution for the non-zero part. Furthermore we assume that this response variable depends on two vectors of covariates X and Z which may have covariates in common. The conditional distribution of the response Y is given by

$$F(y|x,z) = \pi(x) + (1-\pi(x))F_{Y>0}(y|z)$$

where $F_{Y>0}(y|z)$ is a continuous conditional distribution for the non-zero part of the response Y and $\pi(x) = P(Y = 0|X = x)$ is the conditional probability on a zero response. In this paper, we assume a logistic regression function for $\pi(x)$, denoted by $\pi(\gamma,x) = \frac{e^{\gamma' x}}{1+e^{\gamma' x}}$.
For the conditional distribution of the non-zero part of the response $F_{Y>0}(y|z)$, we use a Cox's regression model (Cox (1972)). Hereby, we assume that the conditional hazard function has the following form: $\lambda_{Y>0}(t|z) = \lambda(t)e^{\beta' z}$, where λ is an unknown baseline hazard function. In most studies, it is impossible to fully observe the response variable Y. We assume that there exists a random variable C such that we only observe $T = \max(Y,C)$ and $\delta = I\{Y \geq C\}$. We call this type of data left-censored and assume that, conditionally on X and Z, Y and C are independent. To estimate the parameters γ and β and the baseline hazard function $\lambda(t)$ in this model, we construct a maximum likelihood function. Therefore, let (T_1,δ_1,X_1,Z_1), ..., (T_n,δ_n,X_n,Z_n) be a sample of the observed variables (T,δ,X,Z). Hereby X_i and Z_i are the vectors of covariate values for individual i. We find the following likelihood function:

$$\begin{aligned} L^e(\gamma,\beta,\Lambda) = \prod_{i=1}^{n} \ & \{(1-\pi(\gamma,x_i))\lambda(T_i)e^{\beta' z_i}\exp[-e^{\beta' z_i}\Lambda(T_i)]\}^{\delta_i} \\ & \{\pi(\gamma,x_i) + (1-\pi(\gamma,x_i))(1-\exp[-e^{\beta' z_i}\Lambda(T_i)])\}^{1-\delta_i}. \end{aligned}$$

In this expression, we estimate the baseline cumulative hazard function by a nonparametric step function:

$$\Lambda(t) = \sum_{k=1}^{q_n} \lambda(u_k)\mathrm{I}(u_k \leq t),$$

where $0 < u_1 < \ldots < u_{q_n}$ are the unique uncensored observations.
Remarks: In a study with left-censoring, the largest observations are often uncensored. In order to facilitate the maximum likelihood estimation procedure, we note that we can find a closed form solution for the step sizes of the nonparametric baseline cumulative hazard function in these uncensored observations. In the most extreme case, all censored observations are smaller than the smallest uncensored observation. Studies with a fixed detection limit follow this scheme. Fitting the zero-inflated Cox's regression model simplifies to fitting a logistic regression model on the censoring indicator random variables and fitting a Cox's regression model on the uncensored observations.

As a result, one can proof the consistency of the maximum likelihood estimators, following the ideas of Kim *et al.* (2010). Let $(\gamma_0, \beta_0, \Lambda_0)$ be the true values of the parameters.

Theorem 1. *Under some regularity conditions, the maximum likelihood estimators* $(\hat{\gamma}, \hat{\beta}, \hat{\Lambda})$ *are consistent. This means that,*

$$|\hat{\gamma} - \gamma_0| \to 0,\ |\hat{\beta} - \beta_0| \to 0 \text{ and } \sup_t |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0,$$

with probability 1.

3 Simulation study

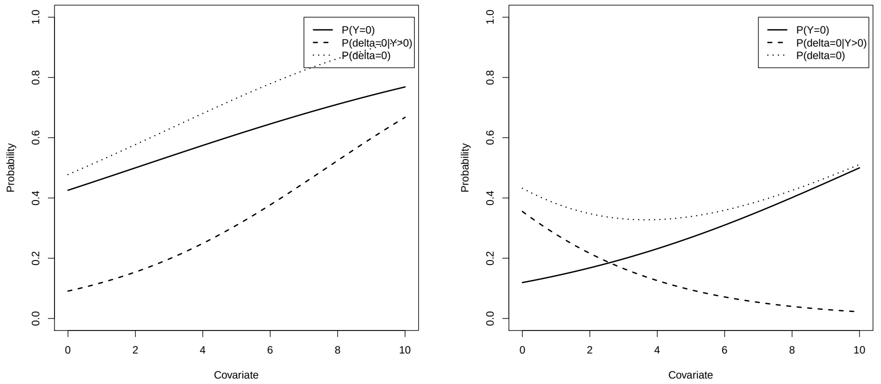
In order to show the performance of the zero-inflated semi-parametric Cox's regression model for univariate left-censored survival data, we set up a simulation study. We generate data sets from the following model:

(i) $X = Z \sim U[0, 10]$.
(ii) $C \sim \text{Weibull}(a_c = 1, b_c = 0.1)$.
(iii) Probability on a zero response: $\pi(X, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 * X)}{1 + \exp(\gamma_0 + \gamma_1 * X)}$.
(iv) $J \sim Bernouilli(1 - \pi(X, \gamma))$.
(v) If $J = 0$, then $T = C$ and $\delta = 0$. If $J = 1$, then $T = \max\{Y, C\}$ and $\delta = I\{Y \geq C\}$, where $Y \sim$ Cox's model (parameter: β, baseline hazard: Weibull(a_0,b_0)).

In each simulation, we generate 500 data sets with n observations. We consider two settings:

	γ_0	γ_1	β	a_0	b_0
Setting 1	-0.3	0.15	0.3	1	1
Setting 2	-2	0.20	-0.3	0.3	1

Table 1. Different settings.



(a) Setting 1

(b) Setting 2

Figure 1. Probability on a zero and censoring probability.

The corresponding probabilities on a zero response and censoring probabilities are shown in Figure 1. These probabilities depend on the value of the covariate.
We generate data sets with $n = 100$, $n = 500$ observations. For the two different settings, we calculate the mean en standard deviation of the 500 ML estimates. The results are shown in Table 2.

		Setting 1	Setting 2
n=100	$\hat{\gamma}_0$	-0.318 (0.510)	-1.445 (1.012)
	$\hat{\gamma}_1$	0.172 (0.120)	0.133 (0.132)
	$\hat{\beta}$	0.318 (0.102)	-0.304 (0.058)
	$\hat{\Lambda}(1)$	1.005 (0.410)	0.791 (0.309)
	$\hat{\Lambda}(2)$	1.815 (0.557)	1.029 (0.357)
n=500	$\hat{\gamma}_0$	-0.305 (0.244)	-1.552 (0.543)
	$\hat{\gamma}_1$	0.157 (0.056)	0.147 (0.069)
	$\hat{\beta}$	0.305 (0.039)	-0.302 (0.026)
	$\hat{\Lambda}(1)$	0.995 (0.175)	0.857 (0.176)
	$\hat{\Lambda}(2)$	2.009 (0.398)	1.091 (0.195)

Table 2. Mean (standard deviation) of ML estimates.

The means of the maximum likelihood estimates come closer to the true values as the sample size increases. The standard deviations of the maximum likelihood estimates decrease as the sample size increases. This is in line with the theoretical results. In the second setting, the bias in the estimates of the parameters in the logistic regression function is higher in comparison to setting 1. This is due to the lower probability on a zero response for small values of the covariate. The estimation of the logistic regression parameters γ_0 and γ_1 is better in settings with high probability on a zero response. The estimation of the β-parameter and of the baseline cumulative hazard function improve in settings with lower censoring probability.

4 EXAMPLE: MODELING ETHANOL-INDUCED ANESTHESIA.

In this section, we illustrate the zero-inflated Cox's regression model with a practical study of ethanol-induced anesthesia (sleep time) in genetically-selected strains of mice described by Markel *et al.* (1995). The mice were injected intraperitoneally with a $4.1g/kg$ dose of ethanol. Afterwards each mouse was placed on its back and was considered anesthetized if it did not right itself within 1 min. Therefore we use 1 min as detection limit. Due to the breeding process of the test mice it was possible that some mice were "immune" for the administered ethanol dose and would not fall asleep or slept only a very short time. In this example, we consider the influence of the following covariates on sleep time: sex, albinism, trial day, weight at trial 1, and an interaction between sex and albinism. The parameter estimates and their standard errors are given in Table 3.

	Semi-parametric Zero-inflated Cox model	Parametric Logistic-Weibull model
	Logistic part	
Intercept	-4.0384 (1.6995)	-4.0601 (1.7110)
Sex	0.7316 (0.4925)	0.7384 (0.4969)
Albinism	1.3077 (0.4464)	1.3140 (0.4488)
Sex*Albinism	-1.2499 (0.6890)	-1.2570 (0.6937)
Trial day	-0.0006 (0.0004)	-0.0006 (0.0004)
Weight	0.0682 (0.0691)	0.0694 (0.0696)
	Hazard part	
Sex	0.0062 (0.0909)	0.0001 (0.0902)
Albinism	0.1187 (0.1045)	0.0783 (0.1043)
Sex*Albinism	-0.0280 (0.1483)	0.0204 (0.1479)
Trial day	0.0005 (0.0001)	0.0005 (0.0001)
Weight	-0.0341 (0.0134)	-0.0354 (0.0134)

Table 3. Estimates (standard errors) for the different covariates.

In the same table we also give a parametric Logistic-Weibull model to compare with the zero-inflated Cox's regression model. We notice in Table 3 that in both the zero-inflated Cox's model and the parametric Logistic-Weibull model, the same covariates have a significant effect in the logistic and the hazard part of each model. In the logistic part of the models, an albino mouse has a significant higher probability on having a zero value for the sleep time than a non-albino mouse. Furthermore we note that the gender of a mouse also has a significant effect in this part, through its interaction with albinism. We see that a female albino mouse has a lower probability on non-sleep than a male mouse. The other covariates do not have a significant effect in the logistic part of both models. For the hazard part of each model, we see that only the covariates Trial day and Weight before the first test session have a significant influence on the hazard. The estimate for the parameter of Trial day is positive which indicates that the hazard increases when the study progresses. This data set was collected over a period of 3 years and such an increasing hazard likely indicates that the investigators became more skilled and were better able to assess sleep time in these mice. Therefore, the observations for sleep time became shorter as the studied progressed. For the other significant variable Weight, we have in both models a negative sign which indicates that the hazard decreases for heavier animals. This means a longer sleep time for these animals. In Table 3, we also see that the estimates for the different covariates are almost the same in the zero-inflated Cox's model and in the parametric Logistic-Weibull model. Finally, we note that, for small values of sleep time, there is not much difference between the estimates of the cumulative baseline hazard in the parametric model and in the zero-inflated Cox's model.

5 Conclusion

In several studies with left-censored data, the underlying time until an event can also become zero. To accommodate for this problem and to study the influence of covariates on the response variable, we introduced a zero-inflated Cox's regression model. In this model, we assumed that the probability of having a zero response is modeled through a logistic regression. Furthermore we assumed that the hazard of the non-zero part of the response follows a Cox's regression model. We estimated the baseline cumulative hazard function by a nonparametric step function. The different parameters in the model are estimated by maximum likelihood techniques. The consistency of the maximum likelihood estimators was stated as an important result. The simulation results showed that the model performs well. Finally, we applied the regression model on a practical data set of ethanol-induced sleep time in mice.

References

BLACKWOOD, L. G. (1991): Analyzing censored environmental data using survival analysis: single sample techniques. *Environmental Monitoring and Assessment*, *18*, 25-40.

COX, D.R. (1972): Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, *34*, 187-220.

KIM, Y., KIM, B, JANG, W. (2010): Asymptotic properties of the maximum likelihood estimator for the proportional hazards model with doubly censored data. *Journal of Multivariate Analysis*, *101*, 1339-1351.

MARKEL, P.D., DEFRIES, J.C., JOHNSON, T.E. (1995): Ethanol-induced anesthesia in inbred strains of long-sleep and short-sleep mice: A genetic analysis of repeated measures using censored data. *Behavior Genetics*, *25*, 67-73.

MOULTON, L.H., HALSEY, N.A. (1995): A mixture model with detection limits for regression analysis of antibody response to vaccine. *Biometrics*, *51*, 1570-1578.

PARNER, E. (1998): Asymptotic theory for the correlated gamma frailty model. *The Annals of Statistics*, *26*, 183-214.

YANG, Y., SIMPSON, D. (2010): Unified computational methods for regression analysis of zero-inflated and bound-inflated data. *Computational Statistics and Data Analysis*, *54*, 1525-1534.