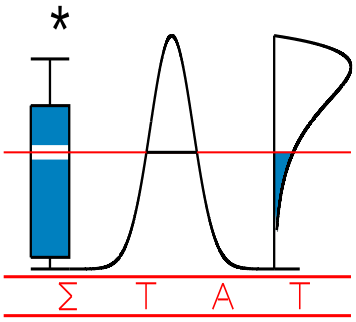
T E C H N I C A L

R E P O R T

**07034**

# ZERO-INFLATED COX REGRESSION WITH A DETECTION LIMIT

BRAEKERS, R. and PAUL D. MARKEL




I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

# Zero-inflated Cox regression with a detection limit.

Roel Braekers[1] and Paul D. Markel[2]

## Abstract

In some clinical, environmental or industrial trials, fixed-detection limits can result in a positive response variable that is type I left-censored. It is common to assume in this setting that this response variable has an absolute continuous distribution. However we notice that the number of censored observations is often larger than expected under this assumption. Furthermore we also want to investigate the influence of several covariates on this response variable. To solve these issues, we introduce in this paper a zero-inflated Cox regression model in which we assume a zero-inflated distribution for the underlying response variable of interest. We model the conditional probability of having a zero value for the response, a failure, by a logistic regression model. For the non-zero part of the response variable we use a Cox regression model. Subsequently, we present a joint model comprising a logistic regression used with the censoring indicator as a response variable plus an ordinary Cox regression model for right censored data used with the subgroup of uncensored observations. We note that this implies that the asymptotic theory for the logistic and Cox regression model remains valid in this zero-inflated Cox regression model. Afterwards we apply this model to a practical study of ethanol-induced anesthesia in genetically-selected mice and compare the results with a parametric Logistic-Weibull model.

Keywords: Cox regression, fixed detection limit, left censoring, mixture.

## 1 Introduction

In some clinical, environmental or industrial trials, the primary interest is in a positive random variable such as a CD4+ count in AIDs research, a concentration of a toxic chemical compound in a river-system or in a industrial solution, .... However we note that there are often difficulties in measuring this positive variable. For example, if we measure the concentration of a chemical compound, the measuring apparatus might not give a correct result below a certain fixed limit due to technical limitations. In this case, the positive random variable is censored below a fixed detection limit and represents a distribution of type I left-censored observations. Several methods have been proposed in the literature to handle this type of data. In parametric modeling, both a likelihood method and (multiple) imputation method are used to find estimates for the parameters of a model (Hughes (1999), Lyles et al. (2001), Thiébaut and Jacqmin-Gadda (2004)). For nonparametric models, the Kaplan-Meier estimator for right-censored data was used after the time-axe was reversed (Blackwood (1991), She (1997)).

[1]Corresponding author, E-mail: roel.braekers@uhasselt.be
Hasselt University, Center for Statistics, Campus Diepenbeek, Agoralaan, building D, 3590 Diepenbeek, Belgium
[2]Minot State University, Department of Addiction Studies, Psychology and Social Work, 500 University Avenue West, Minot (ND) 58707, U.S.A.

In this paper, we study a phenomenon often occurring with type I left-censored data; namely, we notice that the amount of censored observations below the detection limit is larger than what we expect if the positive variable has an absolute continuous distribution. A possible explanation of this phenomenon is given in a following example. In a clinical study of ethanol-induced anesthesia (sleep time) in mice, some subjects do not fall asleep. However, due to the fixed detection limit resulting from the assessment method in this study, it was not possible to distinguish between subjects who did not fall asleep and those that slept only a very short time, i.e., below the detection limit. The group of censored observations consists in this case of two different groups of individuals. There is a group of subjects that did not show any response, i.e., non responders or failures. In contrast, there is a group of subjects with a very small positive response, i.e., responders or success. To model this excess of censored observations, we assume in this paper that the underlying positive variable of interest has a zero-inflated mixture distribution as follows

$$Y \sim F_Y(y) = \pi + (1-\pi)F_c(y)$$

where $\pi$ is a positive probability of having a response equal to zero and $F_c$ is a positive absolute continuous distribution for a response strictly larger than zero. We regard the individuals which have a non-zero response for the variable as responders or success. While the individuals with a response equal to zero are referred to as non responders or failures. Due to the fixed detection limit, we note that we cannot fully discriminate between these groups since all the non-responders are censored observations but not all censored observations are non-responders.

The structure of this article is as follows: In Section 2, we introduce mathematically a zero-inflated Cox regression model with a zero-inflated distribution of the response variable with covariates. Furthermore we show in this section how this zero-inflated Cox model is equivalent to a joint model of a logistic model on the censoring indicator and an ordinary Cox regression model on the subgroup of uncensored observations. Afterwards, in Section 3, we illustrate this model on a practical data set of ethanol-induced sleep time in mice. In Section 4 we give some conclusions about our results and some indications about future research.

# 2 Methodology

In this section, we introduce a zero-inflated semiparametric Cox regression model for left-censored data under a fixed detection limit. Let us denote by $Y$ a positive response variable of interest. We assume that this variable 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 group of covariates $\mathbf{Z}$ and $\mathbf{X}$ which may have covariates in common. The conditional distribution of the response $Y$ is given by

$$F_{Y|\mathbf{Z},\mathbf{X}}(y|\mathbf{z},\mathbf{x}) = P(Y \leq y|\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}) = \pi(\mathbf{z}) + (1-\pi(\mathbf{z}))F_{c|\mathbf{X}}(y|\mathbf{x}) \tag{1}$$

where $F_{c|\mathbf{X}}$ is a continuous conditional distribution for the non-zero part of the response $Y$ and $\pi(\mathbf{z})$ is the probability for a failure.

However, in a variety of studies, the response variable Y is not observed fully due to a fixed detection limit $d > 0$. This variable is only observed when it exceeds the detection limit. For observed data, we have the following variables

$$T = \max(Y, d) \text{ and } \delta = I(Y \geq d)$$

where $T$ is the maximum of the response variable and the detection limit, and $\delta$ indicates whether the response is greater than the detection limit. Using (1), we find that the conditional distribution for the observed variable $T$ is given by

$$F_{T|\mathbf{Z},\mathbf{X}}(t|\mathbf{z},\mathbf{x}) = \begin{cases} 0 & ,t < d \\ \pi(\mathbf{z}) + (1-\pi(\mathbf{z}))F_{c|\mathbf{X}}(t|\mathbf{x}) & ,t \geq d \end{cases}.$$

Sofar, we did not specify how the probability $\pi(\mathbf{z})$ of having a zero-response depends on the covariates $\mathbf{Z}$ or how the conditional distribution $F_{c|\mathbf{X}}$ for the success depends on the covariates $\mathbf{X}$. In the literature, mostly parametric models are considered for each of these quantities (Moulton and Halsey (1995), Taylor et al.(2001)). However in this paper we only partly follow this trend. For the probability $\pi(\mathbf{z})$ for a failure, we assume a parametric logistic regression model

$$\text{Logit } \pi(\mathbf{z}) = \gamma^{\mathbf{t}}\mathbf{Z} \tag{2}$$

where $\gamma$ is the vector of coefficients for the different covariates $\mathbf{Z}$. On the other hand, for the conditional distribution function $F_{c|\mathbf{X}}$ of the non-zero part of the response, we use a Cox regression model (Cox (1972)) to model this distribution and assume that the conditional hazard function has the following form

$$\lambda_c(t|\mathbf{x}) = \frac{f_{c|\mathbf{X}}(t|\mathbf{x})}{\bar{F}_{c|\mathbf{X}}(t|\mathbf{x})} = \lambda_0(t)e^{\beta^t\mathbf{X}} \tag{3}$$

where $\lambda_0(t)$ is an unknown baseline hazard function and the vector $\beta$ is a vector of coefficients for the covariates $\mathbf{X}$. In this expression $f_{c|\mathbf{X}}$ is the conditional density function of the non-zero part of the response $Y$. We can rewrite the conditional distribution function in this model as

$$F_{c|\mathbf{X}}(t|\mathbf{x}) = 1 - e^{-e^{\beta^t\mathbf{x}}\int_0^t \lambda_0(s)ds}.$$

To estimate the different parameters in this model, we construct a maximum likelihood function. Therefore, let $(T_1,\delta_1,\mathbf{Z}_1,\mathbf{X}_1),\ldots,(T_n,\delta_n,\mathbf{Z}_n,\mathbf{X}_n)$ be a sample of the observed variables $(T,\delta,\mathbf{Z},\mathbf{X})$. Hereby $\mathbf{Z}_i$ and $\mathbf{X}_i$ are the vectors of covariate values for individual $i$. Before we assemble the likelihood function, we first derive the conditional sub-distributions for the (un)censored observations. We have that

$$\begin{aligned} P(T \leq t, \delta = 1|\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}) &= P(\max(Y,d) \leq t, Y \geq d|\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}) \\ &= \begin{cases} 0 & ,t < d \\ P(d \leq Y \leq t|\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}) & ,t \geq d \end{cases} \\ &= \begin{cases} 0 & ,t < d \\ (1-\pi(\mathbf{z}))(F_{c|\mathbf{X}}(t|\mathbf{x}) - F_{c|\mathbf{X}}(d|\mathbf{x})) & ,t \geq d \end{cases} \end{aligned}$$

and

$$P(T \leq t, \delta = 0|\mathbf{Z}=\mathbf{z},\mathbf{X}=\mathbf{x}) = \begin{cases} 0 & ,t < d \\ \pi(\mathbf{z}) + (1-\pi(\mathbf{z}))F_{c|\mathbf{X}}(d|\mathbf{x}) & ,t \geq d \end{cases}$$

We note that the conditional sub-distribution of the uncensored observations is a continuous distribution and the conditional sub-distribution of the censored observations is degenerate at the detection limit $d$. From these quantities, we construct the likelihood function of (1) which is given by

$$L = \prod_{i=1}^{n}\left[\frac{\partial}{\partial t}P(T \leq t, \delta = 1|\mathbf{Z}=\mathbf{z_i},\mathbf{X}=\mathbf{x_i})\Big|_{t=t_i}\right]^{\delta_i}\left[P\left(T \leq d, \delta_i = 0|\mathbf{Z}=\mathbf{z_i}\mathbf{X}=\mathbf{x_i}\right)\right]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[(1-\pi(\mathbf{z}_i))f_{c|\mathbf{X}}(t_i|\mathbf{x}_i)\right]^{\delta_i} \left[\pi(\mathbf{z}_i) + (1-\pi(\mathbf{z}_i))F_{c|\mathbf{X}}(d|\mathbf{x}_i)\right]^{1-\delta_i}$$

After filling the models (2) and (3) in the likelihood, we get

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[(1-\pi(\mathbf{z}_i))\lambda_c(t_i|\mathbf{x}_i)\bar{F}_{c|\mathbf{X}}(t_i|\mathbf{x}_i)\right]^{\delta_i} \left[\pi(\mathbf{z}_i) + (1-\pi(\mathbf{z}_i))F_{c|\mathbf{X}}(d|\mathbf{x}_i)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{\lambda_0(t_i)e^{\beta^t\mathbf{x}_i} e^{-e^{\beta^t\mathbf{x}_i}\int_0^{t_i}\lambda_0(s)ds}}{1+e^{\gamma^t\mathbf{z}_i}}\right]^{\delta_i} \left[\frac{e^{\gamma^t\mathbf{z}_i}+1-e^{-e^{\beta^t\mathbf{x}_i}\int_0^{d}\lambda_0(s)ds}}{1+e^{\gamma^t\mathbf{z}_i}}\right]^{1-\delta_i}. \end{aligned}$$

Taking the logarithm, we get the loglikelihood

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \left[\log(\lambda_0(t_i)) + \beta^t\mathbf{x}_i - e^{\beta^t\mathbf{x}_i}\int_0^{t_i}\lambda_0(s)ds - \log(1+e^{\gamma^t\mathbf{z}_i})\right] \\ &+ \sum_{i=1}^{n}(1-\delta_i)\left[\log\left(e^{\gamma^t\mathbf{z}_i}+1-e^{-e^{\beta^t\mathbf{x}_i}\int_0^{d}\lambda_0(s)ds}\right) - \log(1+e^{\gamma^t\mathbf{z}_i})\right]. \end{aligned} \tag{4}$$

In this expression we estimate the cumulative baseline hazard by a nonparametric step function. Let $0 = U_0 < U_1 < \ldots < U_k$ denote the distinct uncensored lifetimes, we assume that the baseline hazard is constant between these values and define an estimator for the cumulative baseline hazard as

$$\int_0^t \hat{\lambda}_0(s)ds = \sum_{j=1}^{k} \lambda_j I(U_j \leq t).$$

By the fixed detection limit, we note that the smallest uncensored observation has a lifetime greater or equal than the detection limit. Since the sub-distribution of the uncensored observations is a continuous distribution, we assume that all observed uncensored lifetimes are strictly greater than the detection limit. This is the case for most of the data sets. However this greatly simplifies the loglikelihood function because the integral in the contribution of a censored observation is zero. Therefore the log likelihood follows:

$$\begin{aligned} l &= \sum_{i=1}^{n} \delta_i \left[\log\left(\sum_{j=1}^{k}\lambda_j I(U_j = t_i)\right) + \beta^t\mathbf{x}_i - e^{\beta^t\mathbf{x}_i}\sum_{j=1}^{k}\lambda_j I(U_j \leq t_i) - \log(1+e^{\gamma^t\mathbf{z}_i})\right] \\ &+ \sum_{i=1}^{n}(1-\delta_i)\left[\gamma^t\mathbf{z}_i - \log(1+e^{\gamma^t\mathbf{z}_i})\right]. \end{aligned}$$

To find the $\lambda_j$'s in this expression, we differentiate this loglikelihood and solve the score equations.

$$\frac{\partial l}{\partial \lambda_j} = \sum_{i=1}^{n} \delta_i \left[\frac{I(U_j = t_i)}{\sum_{j=1}^{k}\lambda_j I(U_j = t_i)} - e^{\beta^t\mathbf{x}_i} I(U_j \leq t_i)\right] = 0, \quad j = 1, \ldots, k$$

$$\Leftrightarrow \quad \lambda_j = \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)}, \quad j = 1, \ldots, k$$

In the loglikelihood function, this gives the following profile loglikelihood, depending only on the parameters $\gamma$ and $\beta$

$$\begin{aligned} l(\gamma, \beta) &= \sum_{i=1}^{n} \delta_i \left[ \log \left( \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)} \right) + \beta^t \mathbf{x}_i - e^{\beta^t \mathbf{x}_i} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)} I(U_j \leq t_i) \right] \\ &+ \log \prod_{i=1}^{n} \left[ \frac{e^{\gamma^t \mathbf{z}_i}}{1 + e^{\gamma^t \mathbf{z}_i}} \right]^{1-\delta_i} \left[ \frac{1}{1 + e^{\gamma^t \mathbf{z}_i}} \right]^{\delta_i}. \end{aligned}$$

We note that this profile loglikelihood splits in two parts $l(\gamma, \beta) = l_1(\gamma) + l_2(\beta)$ where

$$l_1(\gamma) = \log \prod_{i=1}^{n} \left( \frac{e^{\gamma^t \mathbf{z}_i}}{1 + e^{\gamma^t \mathbf{z}_i}} \right)^{1-\delta_i} \left( \frac{1}{1 + e^{\gamma^t \mathbf{z}_i}} \right)^{\delta_i}$$

is the loglikelihood of an ordinary logistic regression model with the indicator $1 - \delta_i$ as response variable. On the other hand, the second part of the loglikelihood function is equal to

$$\begin{aligned} l_2(\beta) &= \sum_{i=1}^{n} \delta_i \left[ \log \left( \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)} \right) + \beta^t \mathbf{x}_i - e^{\beta^t \mathbf{x}_i} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)} I(U_j \leq t_i) \right] \\ &= \log \prod_{i=1}^{n} \left( e^{\beta^t \mathbf{x}_i} \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \delta_i I(U_j = t_i)}{\sum_{i=1}^{n} \delta_i e^{\beta^t \mathbf{x}_i} I(t_i \geq U_j)} \right)^{\delta_i} - \sum_{i=1}^{n} \delta_i \end{aligned}$$

which is, up to a constant, the partial likelihood of an ordinary Cox regression model for right-censored data when applied on a subgroup of uncensored observations.

This result has several appealing implications. Firstly, the asymptotic theory of the logistic regression model and the Cox regression model apply in this zero-inflated Cox regression model. Therefore we have that the parameter estimates in either part of the zero-inflated Cox regression model are consistent estimates and are asymptotically normal-distributed. Because the profile likelihood comprises two separate parts, the parameter estimates in the logistic regression model are independent of the parameter estimates in the Cox regression model. A second consequence of this result is that we can use existing software programs to find the different parameter estimates in this zero-inflated Cox regression model.

Before we end this section we notice that we observe in cure rate models for right-censored data a similar situation in which an large amount of censored observations have a long observed survival time (Kuk and Chen (1992), Sy and Taylor (2000), Peng and Dear (2000)). However in these models as well as in the model presented here, the interest lies on modeling the conditional hazard function. Therefore it is not possible to find the present model by reversing the time-axes in a cure rate model.

# 3 Example: Modeling ethanol-induced anesthesia (sleep time).

In this section, we illustrate the zero-inflated Cox 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 original study includes two parental inbred strains, their isogenic $F_1$ population and a genetically-segregating $F_2$ population derived by crosses of $F_1$ mice, and had as primary goal to study the genetic influence on sleep time. In addition, this data set has a repeated measurement design where mice are tested at two different times. In this example we only consider the observations of the first test session for the segregating $F_2$ mouse population. From Markel et al. (1995), we learn that the mice were injected intraperitoneally with a 4.1 g/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. After a mouse awoke from an ethanol-induced challenge, it's duration of sleep time was recorded in minutes. Due to the breeding process of the test mice it is possible that some mice were "immune" for ethanol 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 (which is a binary variable indicating whether the mouse was albino), trial day, weight at trial 1, and an interaction between sex and albinism. We selected these covariates by a forward selection criteria and found that they were significant in the logistic part or the hazard part of the zero-inflated Cox regression model. The parameter estimates and their standard error are given in Table 1.

| | Semi-parametric Zero-inflated Cox model | Parametric Logistic-Weibull model |
|---|---|---|
| | Logistic part | |
| Intercept | -4.1675 (1.6995) | -1.1232 (1.6377) |
| Sex | 0.8651 (0.5202) | 0.4708 (0.5201) |
| Albinism | 1.5661 (0.4741) | 1.8197 (0.4736) |
| Sex*Albinism | -1.7203 (0.7413) | -2.1802 (0.7814) |
| Trial day | -0.0015 (0.0011) | -0.0020 (0.0011) |
| Weight | 0.0473 (0.0708) | -0.0769 (0.0706) |
| | Hazard part | |
| Intercept | | -8.5053 (0.3826) |
| Sex | -0.0139 (0.0910) | -0.0181 (0.0908) |
| Albinism | 0.0927 (0.1043) | 0.0503 (0.1042) |
| Sex*Albinism | -0.0108 (0.1480) | 0.0406 (0.1476) |
| Trial day | 0.0019 (0.0002) | 0.0018 (0.0002) |
| Weight | -0.0426 (0.0135) | -0.0428 (0.0135) |
| Scale | | 1.8709 (0.0455) |

Table 1: The estimates for the different covariates in the zero-inflated Cox model and the parametric logistic-weibull model. Standard errors are given in brackets

In the same table we also give a parametric Logistic-Weibull model to compare with the zero-inflated Cox

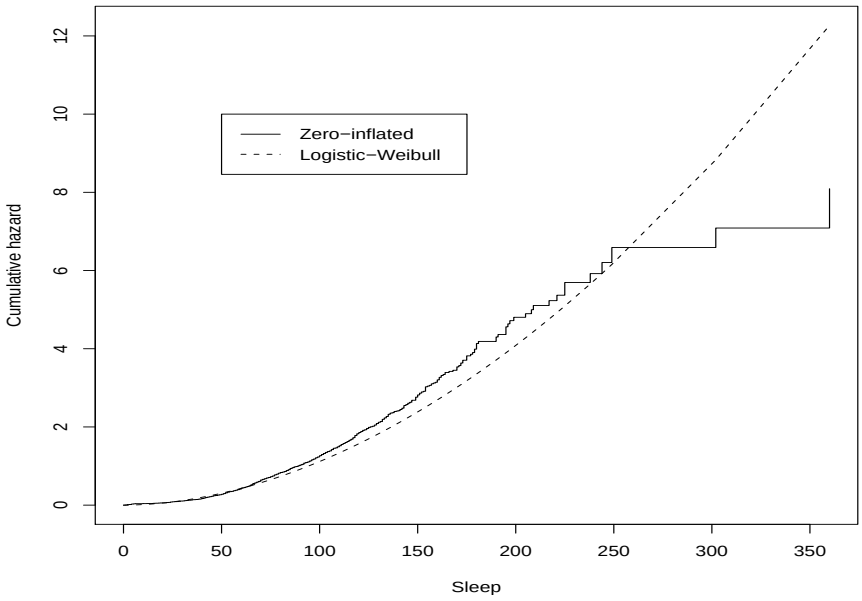



Figure 1: The baseline cumulative hazard function for the Zero-inflated Cox and the Logistic-Weibull model

regression model. As we saw in the methodology, the zero-inflated Cox regression model assumes that the baseline hazard is zero before the smallest uncensored observation. Consequentially, the probability for a value of sleep time between zero and the detection limit is also zero. All censored observations are considered in this model as a result of a failure for the sleep time. In the parametric Logistic-Weibull model, we assume that the baseline hazard comes from a Weibull distribution. The probability for a non-zero censored value of the sleep time has, in this case, an expression which depends on the parameters of the Weibull baseline hazard and is non-zero. Unlike in the zero-inflated Cox regression model the loglikelihood of the parametric model does not simplify and we have to estimate the parameters by maximizing expression (4).

We notice in Table 1 that in both the zero-inflated Cox 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 it's 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 that indicates the

hazard decreases for heavier animals which means a longer sleep time for these animals. This conclusion is expected since the effective dose of ethanol that is administered to each animal, was based on the weight that the animal had the day before the test session. Therefore heavier mice received a larger effective dose.

In Table 1, we also see that the estimates for the different covariates are almost the same in the zero-inflated Cox model and in the parametric Logistic-Weibull model. Initially, we would not expect this because, as stated before, we assumed that the baseline cumulative hazard is zero before the first uncensored observation in the zero-inflated Cox model; this is not the case for the parametric Logistic-Weibull model. In Figure 1, we plotted the baseline cumulative hazard functions for both models and see that they are almost the same for small values of sleep time. So, from the data we notice that the cumulative baseline hazard in the parametric model is almost zero for small times which explains why there is not much difference here between the zero-inflated Cox model and the parametric Logistic-Weibull model. This also explains why the estimates for the covariates do not differ greatly.

# 4 Conclusion

In this paper we investigated why in some clinical, environmental or industrial studies with type I left-censored data, the number of censored observations is larger than what we expect if we assume an absolute continuous distribution for the underlying positive variable of interest. To accommodate for this problem and to study the influence of covariates on the response variable, we introduced a zero-inflated Cox 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 regression model. In constructing this model, the likelihood splits into one part containing the parameters describing the logistic regression model and into another part containing the parameters of the Cox regression model. This results in attractive implications, both in the theoretical and practical sense. Separating the likelihood into two parts, the asymptotic theory of both the logistic regression model and the Cox regression model apply such that the parameter estimates in the zero-inflated Cox regression model are consistent estimates and are asymptotically normally-distributed. A practical implication of the split in the likelihood is that we can use existing software to attain the different parameter estimates in the zero-inflated Cox regression model. In the example, we have applied this regression model in a practical data analysis on a ethanol-induced sleep time study in mice. We note that albinism is the main covariate in the logistic part of the model while Trial Day and Weight have a large influence on the hazard part of the zero-inflated Cox regression model.

# Acknowledgements

Financial support from IAP research network and P6/03 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The data set is this paper was generated under the grants AA03527 and AA08940 from ADAMHA, the Department of Veterans Affairs Alcohol Research Center, and under a training grant of the NIMH (MH-16880).

# 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 B*, **34**, 187-200.

Hughes, J.P. (1999), Mixed effects models with censored data with application to HIV RNA levels, *Biometrics*, **55**, 625-629.

Kuk, A.Y.C. and Chen, C-H (1992), A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, **79**, 531-541.

Lyles, R.H., Fan D. and Chuachoowong, R (2001), Correlation coefficient estimation involving a left censored laboratory assay variable, *Statistics in medicine*, **20**, 2921-2933.

Markel, P.D., DeFries, J.C. and 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. and Halsey, N.A. (1995), A mixture model with detection limits for regression analysis of antibody response to vaccine, *Biometrics*, **51**, 1570-1578.

Peng, Y and Dear, K.B.G. (2000), A nonparametric mixture model for cure rate estimation, *Biometrics*, **56**, 237-243.

She, N. (1997), Analyzing censored water quality data using a non-parametric approach, *Journal of the American water resources association*, **33**, 615-624.

Sy, J.P. and Taylor, J.M.G. (2000), Estimation in a Cox proportional hazards cure model, *Biometrics*, **56**, 227-236.

Taylor, D.J., Kupper, L.L., Rappaport, S.M. and Lyles, R.H. (2001), A mixture model for occupational exposure mean testing with a limit of detection, *Biometrics*, **57**, 681-688.

Thiébaut, R. and Jacqmin-Gadda, H. (2004), Mixed models for longitudinal left-censored repeated measures, *Computer Methods and Programs in Biomedicine*, **74**, 255-260.