# Statistical inference for the semi-parametric proportional reversed hazard model for left-censored and zero-inflated data[1]

Christian Paroissin [1] & Magdalena Pereda Vivo [2]

[1] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France et christian.paroissin@univ-pau.fr*

[2] *Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France et mpvivo@univ-pau.fr*

**Résumé.** Dans ce travail, on envisage d'analyser des données censurées à gauche avec une inflation de zéros. Cependant, il n'est pas possible de différencier une valeur zéro d'une observation positive censurée à gauche. Dans la littérature, certains articles proposent un modèle de mélange pour ce type de données, mais ils assument une distribution paramétrique pour les valeurs positives strictes. Ici, on considère un modèle de régression semi-paramétrique, plus précisément le modèle de risque inverse proportionnel, pour la partie positive. On propose ensuite un modèle de mélange semi-paramétrique pour traiter les données zero-inflatées censurées à gauche et on analyse l'influence des covariables sur les variables. En plus, on présente un algorithme EM pour estimer les différents paramètres et on étudie les propriétés asymptotiques des estimateurs. Cette méthodologie a été appliquée à des données simulées.

**Mots-clés.** Censure à gauche, excès de zéro, fonction de risque inverse

**Abstract.** In this work, we aim to analyse data subject to left censoring with inflation of zeros. It is not possible to distinguish a zero value from a positive left-censored observation. In the literature, there are some articles which propose a mixture model for this type of data but they assumed a parametric distribution for the strict positive values. Here, we consider a semi-parametric regression model, more precisely the proportional reversed hazard model, for the positive part. We then propose a semi-parametric mixture model for dealing with left-censored zero-inflated data and analyse the influence of the covariates on the variables. Furthermore, we provide an EM algorithm to estimate the different parameters and we study the asymptotic properties of the estimators. This methodology has been applied to simulated data.

**Keywords.** Left-censoring, zero excess, likelihood, EM algorithm

# 1 Introduction

In many areas (like toxicology, ecotoxicology, chemistry, geosciences and more generally in environmental sciences, as few examples), studies are based on data obtained by some an-

[1] *This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945416.*

alytical methods. However, with such approaches, one may have only partial information. For instance, when dealing with concentration measurements with an analytical method, one will observe an exact measurement only if it is larger than a certain threshold, called limit of quantification (LOQ): in other words, one has only the information that the concentration lies between zero and this limit. Such a situation is called left-censoring.

Censoring is well-known phenomena in statistics. It appears, for instance, in biostatistics. When dealing with lifetime data, some duration may not be observed exactly since the event occurs later than a certain time point. A typical case is when one performs a medical study over a given period: all the lifetimes longer than this period then get censored. Such a situation is known as right censoring, and it has been investigated rather extensively in the literature. Left censoring has been less studied by statisticians.

Besides, in some situations there could be the absence of the substance under consideration, and this will lead to true zeros (see Blackwood (1991), for example). In statistical terms, we talk about data with zero excess. We can relate the zero-inflated left-censored mixture models to the mixture cure models. In survival analysis, sometimes a part of the population could not experience the event of interest at the end of the follow-up period (susceptible individuals). Those data sets have right censoring. For such a situation, some authors have proposed a mixture cure model. Kuk and Chen (1992) proposed a semi-parametric mixture cure model using a Cox's model in the susceptible group and the logistic regression model for the cure fraction. Sy and Taylor (2000) developed maximum likelihood techniques for the estimation of the parameters in this model using the non-parametric form of the likelihood and an EM algorithm.

In this work, we aim to analyse data subject to left censoring with inflation of zeros. It is not possible to distinguish a zero value from a positive left-censored observation. In the literature, there are two articles which propose a mixture model for this type of data (Moulton (1995) and Yang (2010)), but they assumed a parametric distribution for the strict positive values. Here, we consider a semi-parametric regression model, more precisely the proportional reversed hazard model, for the positive part. We can find in Grouwels (2015) a semi-parametric regression model but they propose a Cox model for the positive part. We find more natural to use the reversed hazard function instead of the hazard function when dealing with left-censoring. For that, we propose a semi-parametric mixture model for dealing with left-censored zero-inflated data and analyse the influence of the covariates on the variables. Furthermore, we provide an EM algorithm to estimate the different parameters and we study the asymptotic properties of the estimators. This methodology has been applied to simulated data.

# 2 Notations

We assume that the observations are the realisations of independent and identically distributed random variables, conditionally to the covariates. We consider the case of multiple random censoring (type III left censoring scheme). Below, we introduce some notations that we will be used in the sequel.

- $n$ is the number of observations (exact or left-censored);
- $T_1, \ldots, T_n$ are the exact measurements (not always observed);
- $C_1, \ldots, C_n$ are the censoring values (not always observed) corresponding to LOQ;
- $X_1, \ldots, X_n$ are the observed values (exact or left-censored);
- $Z_1, \ldots, Z_n$ are the $p$-dimensional covariates;
- $\delta_1, \ldots, \delta_n$ are the indicators of the observation of exact values;
- $m$ is the number of distinct (exact or not) observations;
- $x_{(1)} < \cdots < x_{(m)}$ are the ordered distinct (exact or not) observations:

$$x_{(1)} = \min\{x_i\} < x_{(2)} < \cdots < x_{(m)} = \max\{x_i\};$$

- for any $k \in \{1, \ldots, m\}$, $d_k$ is the number of exact and observed measurements equal to $x_{(k)}$:

$$d_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 1\};$$

- for any $k \in \{1, \ldots, m\}$, $q_k$ is the number of left-censored and observed measurements equal to $x_{(k)}$:

$$q_k = \#\{i \in \{1, \ldots, n\} : x_i = x_{(k)} \text{ and } \delta_i = 0\};$$

- for any $k \in \{1, \ldots, m\}$, $y_k$ is the number of observations less than or equal to $x_{(k)}$:

$$y_k = \#\{i \in \{1, \ldots, n\} : x_i \leqslant x_{(k)}\} = \sum_{j=1}^{k} (d_j + q_j);$$

We assume that $(X_1, \Delta_1, Z_1), \ldots, (X_n, \Delta_n, Z_n)$ is an i.i.d. sample of the observed variables $(X, \delta, Z)$.

# 3 Proportional reversed hazard model with inflation of zeros

We denote by $T$ a semi-continuous variable which can have the value zero or positive values (non-zero values). In addition, we assume that some covariates $Z$ may have influence both on the probability to have a zero and on the distribution for the continuous part. More precisely, we consider the following semi-parametric mixture model for the variable $T$ :

$$F_{T|Z}(t|z) = P(T \leq t|Z = z) = \pi(b, z) + (1 - \pi(b, z))F_{T>0|Z}(t|z) \tag{1}$$

where $F_{T>0}(t|z) = P(T \leq t|T > 0, Z = z)$ is the continuous conditional distribution for the non-zero values of $T$ and $\pi(b, z) = P(T = 0|Z = z)$ is the conditional probability of a zero outcome value. We assume a parametric model for $\pi(b, z)$ where $b$ is the vector of parameters.

For the conditional distribution of the non-zero values, we consider a proportional reversed hazard model. When dealing with possibly left censored data, the reversed hazard function is more natural than the hazard function which is more suitable for the situation of right censoring. As studied by Sengupta and Nanda [**?**], this regression model proposed has the following form:

$$r_{T>0|Z}(t|z) = \frac{f_{T>0|Z}(t|z)}{F_{T>0|Z}(t|z)} = r(t)g(\beta, z), \tag{2}$$

where $r(t)$ is the baseline reversed hazard function and $g(\beta, z)$ is a non-negative function. As for the so-called Cox model, a natural choice for $g$ is the exponential function and the linear relationship for $\beta$ and $z$. In such a case, we have :

$$g(\beta, z) = \exp(z'\beta) = \prod_{j=1}^{p} \beta_j z_j.$$

Let $R(t|Z)$ be the cumulative reversed hazard function of $T$ given $Z$. Taking into account the relation between the reserved hazard function and the cumulative distribution function, we have that

$$F_{T>0|Z}(t|z) = \exp\left\{-\int_t^\infty r_{T>0}(s|z)ds\right\} = \exp\{-g(\beta, z)R(t)\}$$

where $R(t) = \int_t^\infty r(s)ds$ is the baseline cumulative reversed hazard function, which is equivalent to say that

$$F_{T>0|Z}(t|z) = [F(t)]^{g(\beta,z)}$$

where $F(t)$ is the baseline cumulative distribution function.

For this semi-parametric model, the different parameters are the following ones: $b$, $\beta$ and $F$ (or equivalently $r$). The two first ones are Euclidean parameters while the third one is a functional parameter.

For now, we will consider the estimation of the parameters by considering the case of type III left censoring (random censoring). We assume that there exists a random variable $C$ such that we only observe $X = \max(T, C)$ and $\Delta = I(T \geq C)$. Conditionally on the covariate $Z$, we assume that $T$ and $C$ are independent. The observations will be thus $(X_i, \Delta_i, Z_i)$ for $i = 1, \ldots, n$.

## 3.1 Estimation with EM algorithm

In this section, we estimate the parameters $\beta$ and $\gamma$ which maximize $L(\gamma, \beta, F)$. To do so, we will use the Expectation-Maximization (EM) algorithm proposed in Dempster (1997). This is an approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. The missing part is due to situation that, if an observation is smaller the limit of quantification (LOQ), we have not the information if this is a zero or a positive value between zero and the LOQ.

The observed full likelihood for the previous model is

$$L(b,\beta,F)=\prod_{i=1}^{n}[(1-\pi(b,z_i))f_{T>0}(X_i|Z_i)]^{\delta_i}[\pi(b,Z_i)+(1-\pi(b,z_i))F_{T>0}(X_i|Z_i)]^{1-\delta_i}.$$

We define $\widetilde{\Delta}_i=1$ if $T_i>0$ and $\widetilde{\Delta}_i=0$ if $T_i=0$, which is not always observed. Denote the complete data by $(X_i,\Delta_i,Z_i,\widetilde{\Delta}_i)$, which includes the observed data and unobserved $\widetilde{\Delta}_i$. If $\delta_i=1$ we know the exact value of $T_i>C_i>0$, then $\widetilde{\delta}_i=1$. Otherwise, $\widetilde{\delta}_i$ is unobserved since $T_i$ is lower than the censored threshold and we do not know if it is positive. The complete-data full likelihood is

$$L_C(b,\beta,F;\widetilde{\delta}) \;=\; \prod_{i=1}^{n}[\pi(b,Z_i)]^{1-\widetilde{\delta}_i}[(1-\pi(b,Z_i))]^{\widetilde{\delta}_i}r_{T>0}(X_i|Z_i)^{\delta_i\widetilde{\delta}_i}F_{T>0}(X_i|Z_i)^{\widetilde{\delta}_i}.$$

Then, the complete data full log-likelihood is

$$\ell_C(b,\beta,F;\widetilde{\delta})=\sum_{i=1}^{n}(1-\widetilde{\delta}_i)\log\pi(b,Z_i)+\widetilde{\delta}_i\log(1-\pi(b,Z_i))+\widetilde{\delta}_i\left[\delta_i\log r_{T>0}(X_i|Z_i)+\log F_{T>0}(X_i|Z_i)\right].$$

We choose an estimation approach based on the Expectation-Maximization algorithm because of the presence of the latent variable $\widetilde{\delta}$. The EM algorithm is based on iterative maximizations of the expectation of the log-likelihood relative to the complete data.

The E-step takes the expectation of $\ell_C(\gamma,\beta,F;\widetilde{\delta})$ with respect to the unobserved $\widetilde{\delta}_i$, which has the following form

$$\begin{aligned}\tilde{\ell}_C(b,\beta,R;\gamma(\widetilde{\delta})) &= \sum_{i=1}^{n}\Big\{[1-\gamma(\widetilde{\delta}_i)]\log\pi(b,Z_i)+\gamma(\widetilde{\delta}_i)\log(1-\pi(b,Z_i))\Big\}\\ &\quad+\sum_{i=1}^{n}\gamma(\widetilde{\delta}_i)[\delta_i\log r_{T>0}(X_i|Z_i)+\log F_{T>0}(X_i|Z_i)]\\ &= \widetilde{\ell}_1(b;\gamma(\widetilde{\delta}))+\widetilde{\ell}_2(\beta,R;\gamma(\widetilde{\delta})).\end{aligned}\tag{3}$$

where

$$\gamma(\widetilde{\delta}_i)=E[\widetilde{\delta}_i|X_i,\Delta_i,Z_i].$$

- If $\delta_i=0$, then $\widetilde{\delta}_i$ is unobserved and

$$\begin{aligned}\mathbb{E}(\widetilde{\delta}_i|X_i,\Delta_i,Z_i) &= \frac{P(\widetilde{\delta}_i=1,\delta_i=0,|X_i,Z_i)}{P(\delta_i=0,|X_i,Z_i)}=\frac{P(\widetilde{\delta}_i=1,|X_i,Z_i)P(\delta_i=0|\widetilde{\delta}_i=1)}{P(\delta_i=0,|X_i,Z_i)}\\ &= \frac{P(T_i>0)P(\delta_i=0|T_i>0)}{P(\delta_i=0)}=\frac{(1-\pi(\gamma,Z_i))F_{T>0}(X_i|Z_i)}{\pi(\gamma,Z_i)+(1-\pi(\gamma,Z_i))F_{T>0}(X_i|Z_i)}\end{aligned}\tag{4}$$

- If $\delta_i=1$, as $\widetilde{\delta}_i=1$, we get $\mathbb{E}(\widetilde{\delta}_i|X_i,\Delta_i,Z_i)=1$ .

In the M-step, we have to maximize $\tilde{\ell}_C(b, \beta, R; \gamma(\widetilde{\delta}))$ with respect to $b$, $\beta$ and $R$, given $\gamma(\widetilde{\delta})$. Since

$$\tilde{\ell}_C(b, \beta, R; \gamma(\widetilde{\delta})) = \widetilde{\ell}_1(b; \gamma(\widetilde{\delta})) + \widetilde{\ell}_2(\beta, R; \gamma(\widetilde{\delta}))$$

we can estimate $b$ and $\{\beta, R\}$ separately. We will combine EM algorithm and profile likelihood approach, since we need an estimate of the distribution function $F_{T>0}(X_i|Z_i)$ for calculate $\gamma(\widetilde{\delta}_i) = \mathbb{E}[\widetilde{\Delta}_i|X_i, \Delta_i, Z_i]$ in (4) and start using the EM algorithm.

First, we have to make an initial guess of the values of the parameters. We set $\widehat{b}^{(0)} = 0$ and we estimate $\widehat{R}_0^{(0)}$ assuming that $\widehat{\beta}^{(0)} = 0$ and using the non-parametric estimator of the cumulative reversed hazard function

$$\forall t \geqslant 0, \quad \widehat{R}^{(0)}(t) = \sum_{j; x_{(j)} > t} \frac{d_j}{y_j - q_j}.$$

In the $k$-th iteration of the EM algorithm, we proceed as follows.

1. In the E-step, we compute the expected values of $\widetilde{\delta}_i$ using the estimations $\widehat{b}^{(k-1)}$ and $\widehat{\beta}^{(k-1)}$ at the previous iteration:

$$\gamma(\widetilde{\delta}_i)^{(k)} = \delta_i + (1 - \delta_i) \frac{(1 - \pi(\widehat{b}^{(k-1)}, Z_i) \widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)})}{\pi(\widehat{b}^{(k-1)}, Z_i) + (1 - \pi(\widehat{b}^{(k-1)}, Z_i)) \widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)})},$$

   taking into account that

$$\widehat{F}_{T>0}^{(k-1)}(X_i|Z_i; \widehat{\beta}^{(k-1)}) = \exp\{-g(\widehat{\beta}^{(k-1)}, Z_i) \widehat{R}^{(k-1)}(X_i; \widehat{\beta}^{(k-1)})\}$$

   where

$$\widehat{R}^{(k-1)}(X_i; \widehat{\beta}^{(k-1)}) = \sum_{j=1}^{m} \left( \frac{d_j I(X_{(j)} > X_i)}{\sum_{l=1}^{n} \gamma(\widetilde{\delta}_l)^{(k-1)} I(X_{(j)} > X_l) g(\widehat{\beta}^{(k-1)}, Z_l)} \right), \ i = 1, \ldots, n.$$

2. The M-step consists in maximizing the expected likelihood with respect to the parameters of the model.

# Bibliographie


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

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

Sy, Judy P. and Taylor, Jeremy M. G. (2000), Estimation in a Cox Proportional Hazards Cure Model, *Biometric*, 56, pp. 227-236.

Lawrence H. Moulton and Neal A. Halsey (1995), A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine, *Wiley, International Biometric Society*, 51, pp. 1570-1578.

Yang, Yan and Simpson, Douglas (2010), Unified computational methods for regression analysis of zero-inflated and bound-inflated data, *Computational Statistics & Data Analysis*, 54, pp. 1525-1534.

Braekers, Roel and Grouwels, Yves (2015), A semi-parametric Cox's regression model for zero-inflated left-censored time to event data, *Communications in Statistics - Theory and Methods*, 45.

Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1997), Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, pp. 1-22.

Hirose, Yuichi and Liu, Ivy (2020), Statistical Generalized Derivative Applied to the Profile Likelihood Estimation in a Mixture of Semiparametric Models, *Entropy*, 22.