# The destructive zero inflated geometric cure model

**Rodrigo Rossetto Pescim** [1], **Mariana Ragassi Urbano** [2], **Adriano K. Suzuki** [3]

## Introduction

Regression models for survival data with a cure fraction have played an important role in survival analysis in recent years. These types of survival models cover situations in which there are sample units not susceptible to the occurrence of the event of interest. Consequently, a fraction (or proportion) of these individuals are not expected to experience the event of interest, that is, these individuals are considered not susceptible or "cured" in the survival analysis context. The proportion of cured individuals is denoted by cure fraction.

The most popular type of cure rate models are the mixture cure model and the promotion time cure rate model. While the mixture cure model is based on the assumption that only a cause is responsible for the occurrence of the event of interest, that is, the unknown number of causes of the event of interest is assumed to be a Bernoulli random variable, in the promotion time cure model the number of the causes follows a Poisson distribution. In a biological context, the occurrence of the event of interest might be due to one of many competing causes with the number of causes and the distribution of survival times associated with each cause being unknown which leads to a latent competing causes structure.

However, there is an amount (or proportion) of cells that have not been initiated (normal cells), which includes repaired cells, that are not being explained properly by those cure rate models that consider the number of initiated cells related to the occurrence of a tumor being a random variable that follows the power series family of distribution which has as special cases the Poisson, Bernoulli, geometric, negative binomial, etc. In a biological context, it is noticed that there is a much larger number of cells that are not initiated (normal cells) than cells that are initiated (and consequently become malignant cells), which leads to an "excess of not initiated cells" (or "excess number of zero counts") in relation to cells which are lesioned.

Here, we introduce a new cure rate survival model so called the destructive zero inflated geometric cure model by incorporating a structure to estimate the proportion of not initiated cells. To create such structure, we use the concept of zero-inflated models by considering an extension of the discrete power series distributions by including an additional parameter $\pi$. Its interpretation is related to the proportion of repaired cells by means a repair system of the body. In this approach, we assume that the number of initiated cells follow the zero-inflated power series (ZIPS) (Gupta *et al.*, 1995) distribution, which is a suitable choice for modelling data sets that possesses excess of zeros and overdispersion. Furthermore, it provides a realistic interpretation related to the biological mechanism of the occurrence of the event of interest. Also, it includes a process of destruction of tumor cells after an initial treatment.

[1]UEL e-mail: *rrpescim@uel.br*
[2]UEL. e-mail: *mrurbano@uel.br*
[3]ICMC-USP. e-mail: *suzuki@icmc.usp.br*

# Model formulation

Let $N$ be an unobservable (latent) random variable which follows the zero-inflated power series (ZIPS) distribution, denoting the initial number of initiated cells related to the occurrence of a tumor for an individual in population, with probability mass function (pmf) given by

$$P[N=n] = \begin{cases} \pi + (1-\pi)\,\dfrac{a_0}{g(\theta)} & \text{for} \quad n=0 \\ \\ (1-\pi)\,\dfrac{a_n\,\theta^n}{g(\theta)} & \text{for} \quad n=1,2,3,\ldots, \end{cases} \tag{1}$$

where $0<\pi<1$, $a_n>0$ ($a_n$ depends only on $n$) and $g(\theta)=\sum_{n=0}^{\infty} a_n\,\theta^n$ is positive, finite and differentiable function.

The first consequence of a prolonged treatment (destructive process) is the possibly formation of precancerous lesions into the genome of the cells. These cells (with cancerous lesions) are denoted as malignant cells. Given $N=n$, let $X_j$, $j=1,2,\ldots,n$ be independent random variables (independent of $N$) following a Bernoulli distribution with success probability $p$ indicating presence of the $j$th lesion. The pgf of the Bernoulli random variable $X_j$ can be expressed as

$$\mathbb{A}_{X_j}(z) = 1-p\,(1-z), \text{ for } 0\leq z\leq 1. \tag{2}$$

The variable $D$, representing the total number of malignant cells among the $N$ initial cells (competing causes) which are not eliminated by the treatment is defined as

$$D = \begin{cases} X_1+X_2+\ldots+X_N, & \text{if} \quad N>0 \\ 0, & \text{if} \quad N=0, \end{cases} \tag{3}$$

where $D\leq N$. The idea involved in (3) considering that the initial $N$ competing causes are primary initiated malignant cells, where $X_j$ in (3) represents the number of living malignant cells that are descendants of the $j$th initiated malignant cell during some time interval. In this case, $D$ denotes the total number of living malignant cells at some specific time. The time to event for the $j$th competing cause is represented by $V_j$, $j=1,\ldots,D$. We assume that conditional on $D$, the $V_j$ are iid with cumulative distribution function $F(t)$ and survival function $S(t)=1-F(t)$. Also, we note that the total number of malignant cells $D$ and the time $V_j$ are not observable.

Here, the observed time to the event of interest (from the beginning of the treatment to tumor detection or the patient's death) is defined by the following random variable

$$Y=\min\{V_1,\ldots,V_D\} \tag{4}$$

for $D\geq 1$, and $Y=\infty$ if $D=0$, which leads to a proportion $p_0$ of the population which is called cured fraction.

Under this setup, Rodrigues *et al.* (2011) showed that the survival function for the population of the random variable $Y$ in (4) is given by

$$S_{pop}(y) = P[Y\geq y] = \mathbb{A}_D(S(y)) = \sum_{d=0}^{\infty} P[D=d]\{S(y)\}^d = \mathbb{A}_N\left(\mathbb{A}_{X_j}(S(y))\right),$$

where $S(\cdot)$ is the survival function for non-cured population and $\mathbb{A}_D(\cdot)$ is the pgf for the variable $D$. The survival function of the observable lifetime of the event of interest can be expressed as

$$S_{pop}(y) = \pi + (1-\pi)\,\frac{g\left(\theta\left[1-p\,F(y)\right]\right)}{g(\theta)}, \tag{5}$$

where $F(y) = 1 - S(y)$. Hereafter, the model in (5) is referred as the destructive zero inflated power series (DZIPS) cure rate model.

### The destructive zero-inflated geometric (DZIG) cure model

In this context, the $S_{pop}(y)$ for the destructive zero-inflated geometric (DZIG) cure rate model with parameter $\theta = \dfrac{\eta}{1+\eta}$ is given by

$$S_{pop}(y) = \pi + (1-\pi)\,\left[1+\eta\,p\,F(y)\right]^{-1}. \tag{6}$$

The cure fraction is given by $p_0 = \pi + (1-\pi)\,\left[1+\eta\,p\right]^{-1}$. The density and hazard functions of the DZIG cure model are given, respectively, by

$$f_{pop}(y) = (1-\pi)\,\eta\,p\,f(y)\,\left[1+\eta\,p\,F(y)\right]^{-2} \tag{7}$$

and

$$h_{pop}(y) = \frac{(1-\pi)\,\eta\,p\,f(y)\,\left[1+\eta\,p\,F(y)\right]^{-2}}{\pi + (1-\pi)\,\left[1+\eta\,p\,F(y)\right]^{-1}}. \tag{8}$$

# 1 Inference and estimation

Here, we consider the situation when the time to event of interest is not completely observed and is subject to right censoring. Let $C_i$ denote the censoring time. We observe $t_i = \min\{Y_i, C_i\}$ and $\delta_i = 1$ if $Y_i$ is the observed time to the event defined before and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. Let $\boldsymbol{\gamma}$ represent the parameter vector of the distribution for the unobserved lifetime in (4). Here, we note that the DZIG cure rate model are unidentifiable according to Li *et al.* (2001). So, to overcome this problem, we propose to relate the model parameters $p$ and $\theta$ (or $\eta$) to covariates $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$, respectively, without common elements and $\mathbf{x}_{i2}$ without a column of intercepts. Here, the adopted link functions are given by

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_1 \qquad \text{and} \qquad \log\left(\theta_i\right) = \mathbf{x}_{i2}^T\boldsymbol{\beta}_2, \tag{9}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ represent the respective parameter vectors. A critical issue is the selection of covariates to be include in the link functions in (9). From $n$ pairs of times and censoring indicators $(y_1, \delta_1), \ldots, (y_n, \delta_n)$, the observed full likelihood function under non-informative censoring reduces to

$$L(\boldsymbol{\nu}, \mathbf{D}) \propto \prod_{i=1}^{n} \{f_{pop}(t_i;\boldsymbol{\nu})\}^{\delta_i}\{S_{pop}(t_i;\boldsymbol{\nu})\}^{1-\delta_i}, \tag{10}$$

where $\boldsymbol{\nu} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\gamma}^T)^T$, $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{t} = (t_1, \ldots, t_n)$, $\mathbf{x}_1 = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n})$ and $\mathbf{x}_2 = (\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n})$ whereas $f_{pop}(\cdot; \boldsymbol{\nu})$ and $S_{pop}(\cdot; \boldsymbol{\nu})$ are defined in equations (**??**) and (5), respectively.

Now, we assume a Weibull distribution for the unobserved lifetime in (4) with cdf and pdf, respectively, given by

$$F(z; \boldsymbol{\gamma}) = 1 - \exp(-z^{\gamma_1}\, \mathrm{e}^{\gamma_2}) \qquad \text{and} \qquad \mathrm{f(z; \boldsymbol{\gamma})} = \gamma_1\, \mathrm{z}^{\gamma_1 - 1}\, \exp(\gamma_2 - \mathrm{z}^{\gamma_1}\, \mathrm{e}^{\gamma_2}),$$

for $z > 0$, $\boldsymbol{\gamma}^T = (\gamma_1, \gamma_2)^T$, $\gamma_1 > 0$ and $\gamma_2 \in \mathbb{R}$. The choice of the Weibull distribution is due to the fact that this lifetime distribution is one of the most important models used and studied in the survival analysis context.

Maximum likelihood (ML) estimation for the parameter vector $\boldsymbol{\nu}$ can be implemented by numerical maximization of the log-likelihood function $\ell(\boldsymbol{\nu}, \mathbf{D}) = \log L(\boldsymbol{\nu}, \mathbf{D})$ and it can be accomplished using statistical software R.

# Application: cutaneous melanoma data

In this section, we illustrate the usefulness of the DZIG cure rate regression model with an application to a real data set on cancer recurrence. The data are part of a study on cutaneous melanoma (a type of malignant cancer) extracted from Scheike (2009) on 205 patients observed for the evaluation of postoperative in the period from 1962 to 1977. The observed time ($Y$) refers to the time until the patient's death or the censoring time. We can note that there are 72% of censoring. The following variables involved in the study for each patient ($i = 1, \ldots, 205$) are: are: $y_i$ : observed time (in years), $x_{i1}$ : tumor thickness and $x_{i2}$ : ulceration status. As we mentioned earlier, the identifiability issue is avoided if the parameter $p$ is linked only to tumor thickness, while the parameter $\theta$ (or $\eta$) is linked to the ulceration status in the DZIG cure model.

Figure 1a shows that the Kaplan-Meier survival function estimate confirms a plateau around 0.64 and the empirical Kaplan-Meier curves stratified by ulceration status (upper: absent, lower: present) are displayed in Figure 1b and it reveals that the ulceration affects the lifetime of the patients with malignant melanoma.

For model comparison, we fitted the DZIG, Poisson, negative binomial and geometric cure models to cutaneous melanoma data. For the last three models, the destructive process is absent and consequently, the parameter $\theta$ is linked to the both variables (ulceration status and tumor thickness). In order to compare the models, we used the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The results of the DZIPS cure models and its sub-models are described in Table 1.

Tabela 1: The values of max log $L(\cdot)$, the AIC and BIC statistics for the DZIG, Poisson, negative binomial and geometric cure models.

| Cure Model | max log $L(\cdot)$ | AIC | BIC |
|---|---|---|---|
| DZIG | -199.93 | **413.8** | **437.1** |
| Poisson | -207.83 | 425.6 | 442.2 |
| Negative Binomial | -201.52 | 423.0 | 439.7 |
| Geometric | -205.42 | 420.8 | 437.4 |

According to the criteria in Table 1, the DZIG cure model is considering as the best one and then, we select it as our working model. For the DZIG cure rate model, we

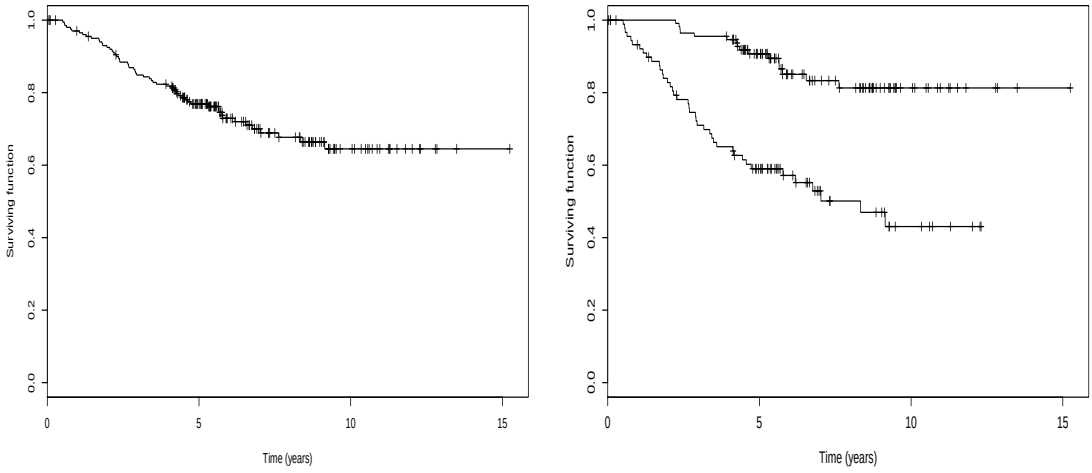


Figura 1: (a) Kaplan-Meier curve for the cutaneous melanoma data. (b) Kaplan-Meier curves stratified by ulceration status (upper: present, lower: absent).

estimate the unknown parameters using the maximum likelihood approach. Table 2 lists the MLEs of the parameters, their standard errors and $p$-values from the fitted model. We note from the fitted DZIG cure rate model that ulceration status and tumor thickness are significant 1% and there is a significant difference for the presence or absent of ulceration status and also a difference related to the thickness of the tumor. Thus, those variables have influence on the survival times of the patients. The estimate of the parameter $\pi$ is 0.3895 or (38.95%), this indicates a proportion of those cells which never experience alterations/lesions.

Tabela 2: Maximum likelihood (ML) estimates, their standard errors and $p$-value of the parameters for the DZIG cure rate regression model.

| Parameter | Estimate | Standard Error | $p$-value |
|---|---|---|---|
| $\gamma_1$ | 2.4153 | 0.2849 | (-) |
| $\gamma_2$ | -5.0041 | 0.6145 | (-) |
| $\pi$ | 0.3895 | 0.2450 | (-) |
| $\beta_{1,intercept}$ | -4.4190 | 0.9348 | $< 0.001$ |
| $\beta_{1,thickness}$ | 0.8656 | 0.2677 | 0.001 |
| $\beta_{2,ulc:present}$ | 2.5968 | 0.8822 | 0.003 |
| $\beta_{2,ulc:ausent}$ | 3.7651 | 0.7496 | $< 0.001$ |

Figure 2 shows the estimated survival function of the DZIG cure rate model for patients with 0.320 mm, 1.940 mm and 4.254 mm tumor thickness. We can note that these represent 5%, 50% and 80% tumor thickness quantiles. The survival rate decreases more rapidly for patients with thicker tumors in presence of ulceration. On the other hand, for patients with less thicker tumor in presence of ulceration, the survival rate does not fall bellow 75% as we can observe in Figure 2a.

Finally, we turn our attention to the role of the ulceration status and thickness tumor covariates on the estimation of the surviving fraction ($p_0$). To estimate the proportion of cured individuals, we use equation (9) and the MLEs of the parameters. So, for the DZIG cure model, the estimative of cure fraction $\hat{p}_0 = \hat{\pi} + (1 - \hat{\pi})\,[1 + \hat{\eta}\,\hat{p}]^{-1}$ is 0.6450.

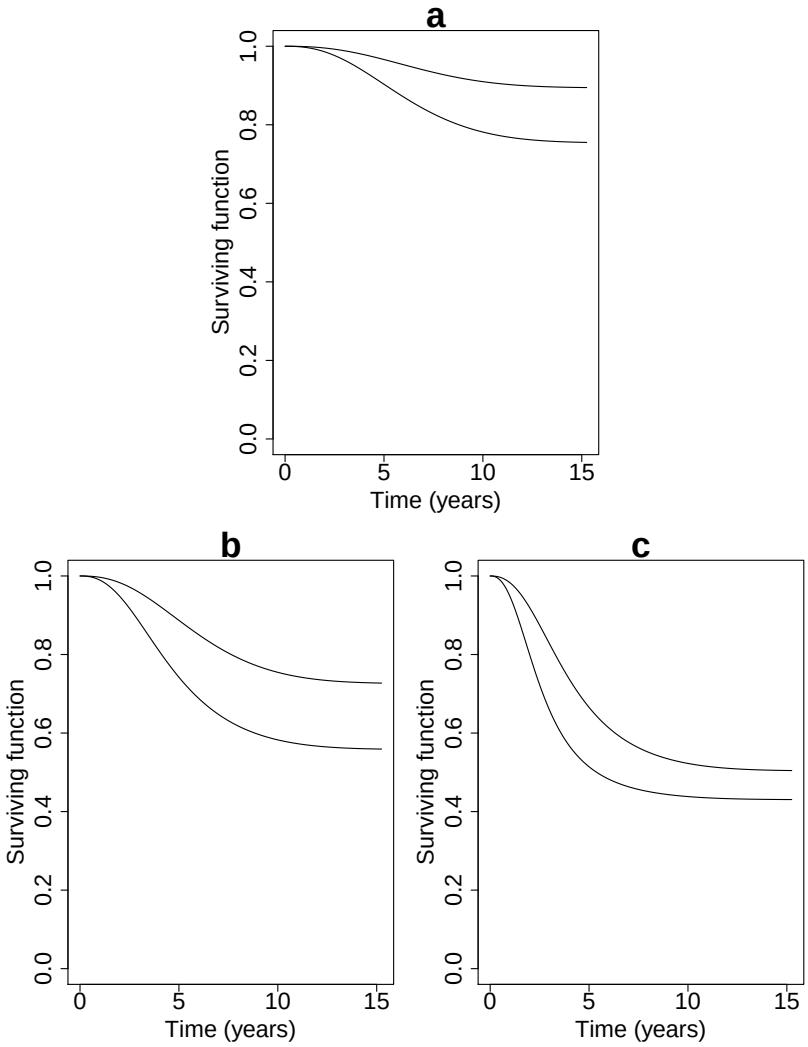


Figura 2: Estimated survival function under the DZIG cure rate model stratified by ulceration status (upper: absent, lower: present) for patients with tumor thickness equal to (a) 0.320 mm, (b) 1.940 mm, and (c) 4.254 mm.

This fact is confirmed in Figure 1a. Also, we noted that the cure rate decreases as tumor thickness size increases and it is smaller for patients with presence of ulceration.

# References

Gupta P.L.; Gupta, R.L. and Tripathi, R.C. (1995). Inflated Modified Power Series Distributions with Applications. *Communications in Statistics - Theory and Methods*, **24**, 2355-2374.

Li, C.S.; Taylor, J. and Sy, J. (2001). Identifiability of cure models. *Statistics and Probability Letters*, **54**, 389-395.

Scheike,T. (2009). Timereg Package. R Package Version 3.4.0. With contributions from T. Martinussen and J. Silver. R package version 1.1-6.