stats

MDPI

Article

A Mixture Model for Survival Data with Both Latent and Non-Latent Cure Fractions

Eduardo Yoshio Nakano [1,*], Frederico Machado Almeida [1] and Marcílio Ramos Pereira Cardial [2]

[1] Department of Statistics, University of Brasilia, Campus Darcy Ribeiro, Asa Norte, Brasilia 70910-900, Brazil; frederico.almeida@unb.br
[2] Institute of Mathematics and Statistics, Federal University of Goias, Goiania 74001-970, Brazil; marcilio.cardial@ufg.br
* Correspondence: nakano@unb.br

Abstract

One of the most popular cure rate models in the literature is the Berkson and Gage mixture model. A characteristic of this model is that it considers the cure to be a latent event. However, there are situations in which the cure is well known, and this information must be considered in the analysis. In this context, this paper proposes a mixture model that accommodates both latent and non-latent cure fractions. More specifically, the proposal is to extend the Berkson and Gage mixture model to include the knowledge of the cure. A simulation study was conducted to investigate the asymptotic properties of maximum likelihood estimators. Finally, the proposed model is illustrated through an application to credit risk modeling.

Keywords: cure rate models; long-term survival; survival analysis; mixture model; credit scoring

MSC: 62N01

Academic Editor: Wei Zhu

Received: 26 July 2025
Revised: 10 September 2025
Accepted: 11 September 2025
Published: 13 September 2025

Citation: Nakano, E.Y.; Almeida, F.M.; Cardial, M.R.P. A Mixture Model for Survival Data with Both Latent and Non-Latent Cure Fractions. *Stats* **2025**, *8*, 82. https://doi.org/10.3390/stats8030082

1. Introduction

Statistical techniques for censored data have been extensively studied in the literature. A common assumption underlying these models is that each individual in the study will eventually experience the event of interest if followed long enough. However, this assumption does not hold in many real-world scenarios, including biomedical, financial, demographic, criminological, and engineering research. Such individuals are typically referred to as cured, non-susceptible, immune, or long-term survivors, and their survival times are considered infinite. The remaining individuals are classified as susceptible.

Models for analyzing data with a proportion of cured individuals are often called cure models or cure rate models. According to [1], such a model, in practical terms can be used, for example, to model data related to various types of cancer for which a significant proportion of patients are cured.

Since the seminal mixture model proposed by [2], which assumes that the population under study is a combination of cured and susceptible individuals, several approaches have been developed to accommodate the presence of a cured fraction. These include promotion time and frailty models. Promotion time models treat the time-to-event as a result of the first occurrence among a set of latent failures. These models are useful in contexts where multiple failure mechanisms are present [3]. Extensions of this modeling approach have been studied in [4–6]. Frailty models incorporate unobserved heterogeneity

between individuals by modeling vulnerability to the event of interest as a latent variable (frailty). These models can be extended to accommodate cure fractions, as presented in [7,8]. Models that incorporate long-term survivors offer an advantage over standard survival techniques, as they allow for the simultaneous estimation of parameters associated with both the susceptible and cured subpopulations [9,10].

Among the aforementioned models, the most popular cure fraction model is the Berkson and Gage mixture model [2]. This model assumes the existence of heterogeneity in the population under study. Consequently, the modeling is based on the mixture of two distributions: one representing the distribution of failure or survival times of susceptible individuals and the other corresponding to a degenerate distribution (which allows for, in principle, infinite survival times) for cured individuals.

Let T be a non-negative random variable denoting the survival time of the entire population. Under the mixture assumption, the Berkson and Gage [2] model has the form

$$S_T(t) = (1 - \phi)S_Y(t) + \phi, \tag{1}$$

where $0 \leq \phi \leq 1$ is the proportion of cured individuals and $S_Y(\cdot)$ denotes the survival function of the susceptible group.

However, the model (1) may not be suitable for some problems since it treats the cure as a latent event. In fact, there are situations in which the cure is known, i.e., situations in which it is known that the censoring occurred due to the individual's cure. One example arises in credit risk modeling, where the variable of interest is the time to default (i.e., delay in loan repayment): a customer who fully repays the loan is a known cured case, and this information should be incorporated into the analysis.

In this context, the main contribution of this work is to develop a model that accommodates both latent and non-latent cure fractions. More specifically, the proposal is to extend the Berkson and Gage mixture model (1) to incorporate known cure information. The model proposed in this paper is illustrated using artificial data of customers who have taken out loans from a financial institution. A credit risk score derived from the proposed model is then used to classify customers according to their risk of default.

This manuscript is organized as follows: Section 2 introduces the model formulation and the procedures for estimating the model parameters using the maximum likelihood method. Section 3 presents a simulation study conducted to investigate whether the usual asymptotic properties of maximum likelihood estimators hold and illustrates the proposed model using artificial data. Finally, concluding remarks are provided in Section 4.

2. Materials and Methods

2.1. Model Formulation

The model proposed in this paper is formulated considering that an individual observed in the sample can be part of one of three distinct subpopulations, consisting of susceptible (non-cured) individuals, non-susceptible individuals who are known to be cured (non-latent cure), and non-susceptible individuals whose status as cued is unknown (latent cure).

The knowledge of the cure of an individual can be represented by a random variable K following a Bernoulli distribution with a success probability ρ. In addition, given $K = 0$, the latent cure variable, C, follows a Bernoulli distribution with success probability ϕ. Thus,

$$K = \begin{cases} 0, & \text{if it is unknown that the individual is cured} \\ 1, & \text{if it is known that the individual is cured,} \end{cases} \tag{2}$$

and

$$C = \begin{cases} 0, \text{ if the individual is not cured} \\ 1, \text{ if the individual is cured.} \end{cases} \tag{3}$$

Since $P(C = 1|K = 1) = 1$, $P(C = 1|K = 0) = \phi$, $P(C = 0|K = 1) = 0$ and $P(C = 0|K = 0) = 1 - \phi$, the probability of an individual being cured is given by

$$\begin{aligned} \mu &= P(C = 1) = P(C = 1, K = 1) + P(C = 1, K = 0) \\ &= P(K = 1)P(C = 1|K = 1) + P(K = 0)P(C = 1|K = 0) \\ &= \rho + (1 - \rho)\phi. \end{aligned} \tag{4}$$

This model proposes a mixture of distributions for individuals who are not known to be cured, i.e., given $K = 0$, we have $f_{T|K=0}(t) = (1 - \phi)f_Y(t) + \phi f_Z(t)$. Here, T is a random variable that represents the time to failure, Y is a continuous random variable that represents the time to failure of non-cured individuals, and Z is a degenerate variable at infinity (i.e., $P(Z > z) = 1, \forall t > 0$) that represents the time to failure of cured individuals.

Given the value of K, the survival function in mixture form is given by

$$\begin{aligned} S_{T|K=0}(t) &= P(T > t|K = 0) \\ &= P(T > t, C = 0|K = 0) + P(T > t, C = 1|K = 0) \\ &= P(C = 0|K = 0)P(T > t|C = 0, K = 0) \\ &+ P(C = 1|K = 0)P(T > t|C = 1, K = 0) \\ &= (1 - \phi)S_Y(t) + \phi, \end{aligned} \tag{5}$$

and

$$S_{T|K=1}(t) = P(T > t|K = 1) = 1. \tag{6}$$

Thus, from (5) and (6), we have that the survival function and the probability density function of the entire population are given, respectively, by

$$\begin{aligned} S_T(t) &= P(T > t, K = 1) + P(T > t, K = 1) \\ &= P(K = 0)P(T > t|K = 0) + P(K = 1)P(T > t|K = 1) \\ &= (1 - \rho)[(1 - \phi)S_Y(t) + \phi] + \rho, \end{aligned} \tag{7}$$

and

$$f_T(t) = -\frac{d}{dt}S_T(t) = (1 - \rho)(1 - \phi)f_Y(t), \tag{8}$$

where ρ is the proportion of individuals who are known to be cured, ϕ is the proportion of cured individuals, and $S_Y(\cdot)$ and $f_Y(\cdot)$ are, respectively, the survival and probability density function of susceptible individuals.

Note that if $\rho = 0$, the model (7) reduces to the Berkson and Gage [2] mixture model (1). In addition, this model can also be characterized as a cure rate mixture model with competing risks [11], where the non-latent cure is considered a competing cause, and its time is assumed to be a degenerate variable at infinity.

2.2. Likelihood Function

Assuming a non-informative right-censoring mechanism, the response of the individual i observed in the sample can be represented by the term $(t_i, \delta_{1i}, \delta_{2i})$, where t_i is

the observed time of the i-th individual in the sample and δ_{1i} and δ_{2i} are their respective indicators of censorship and knowledge of the cure, $i = 1, 2, \ldots, n$. Here,

$$\delta_{1i} = \begin{cases} 0, & \text{if the observation } t_i \text{ is censored} \\ 1, & \text{if the observation } t_i \text{ is not censored,} \end{cases} \tag{9}$$

and

$$\delta_{2i} = \begin{cases} 0, & \text{if it is unknown that individual } i \text{ is cured in } t_i \\ 1, & \text{if it is known that individual } i \text{ is cured in } t_i. \end{cases} \tag{10}$$

The contribution of the i-th individual to the likelihood function is given by $P(K = 1) = \rho$, if it is known that individual i is cured in t_i; $f_T(t_i) = (1-\rho)(1-\phi)f_Y(t_i;\boldsymbol{\theta})$, if it is unknown that individual i is cured and the observation t_i is not censored; and $P(T > t_i, K = 0) = (1-\phi)S_Y(t_i;\boldsymbol{\theta}) + \phi]$, if it is unknown that individual i is cured and the observation t_i is censored. That is, for an observed value of $(t_i, \delta_{1i}, \delta_{2i})$, the contribution to the likelihood function is

$$[\rho]^{\delta_{2i}}[(1-\rho)(1-\phi)f_Y(t_i;\boldsymbol{\theta})]^{(1-\delta_{2i})\delta_{1i}}[(1-\rho)[(1-\phi)S_Y(t_i;\boldsymbol{\theta}) + \phi]]^{(1-\delta_{2i})(1-\delta_{1i})}.$$

Thus, the likelihood function is given by

$$\begin{aligned} L(\rho, \phi, \boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2}) \quad \propto \quad & \prod_{i=1}^{n} \Big\{ [\rho]^{\delta_{2i}}[(1-\rho)(1-\phi)f_Y(t_i;\boldsymbol{\theta})]^{(1-\delta_{2i})\delta_{1i}} \\ \times \quad & [(1-\rho)[(1-\phi)S_Y(t_i;\boldsymbol{\theta}) + \phi]]^{(1-\delta_{2i})(1-\delta_{1i})} \Big\}, \end{aligned} \tag{11}$$

where $0 \leq \rho \leq 1$, $0 \leq \phi \leq 1$ and $\boldsymbol{\theta}$ are the parameters to be estimated. Here, $S_Y(\cdot;\boldsymbol{\theta})$ and $f_Y(\cdot;\boldsymbol{\theta})$ are, respectively, the survival and probability density functions of susceptible (non-cured) individuals, and $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)'$ is the vector of times observed in the sample with their respective indicators of censorship $\boldsymbol{\delta_1} = (\delta_{11}, \delta_{12}, \ldots, \delta_{1n})'$ and knowledge of cure $\boldsymbol{\delta_2} = (\delta_{21}, \delta_{22}, \ldots, \delta_{2n})'$.

Note that when the distribution of susceptible individuals, $f_Y(\cdot;\theta)$, is identifiable, the likelihood function (11) is identifiable as well, except in cases where no observations with known cure status are present in the sample. In fact, when $\sum_{i=1}^{n} \delta_{2i} = 0$, the likelihood function (11) results in $(1-\rho)(1-\phi)\prod_{i=1}^{n} f_Y(t_i;\boldsymbol{\theta})$, which is not identifiable (the non-identifiability arises due to the permutation of the values of ϕ and ρ). In such scenarios, where no known cures are observed, the proposed model is not recommended, and it is more appropriate to consider the Berkson and Gage mixture model (1).

2.3. Regression Model

In this work, the proposal is to incorporate the covariates into the model in the probability of cure μ, as presented in (4). Thus, considering the logit link function, the regression model will be expressed by $\log\left(\frac{\mu}{1-\mu}\right) = \boldsymbol{X}'\boldsymbol{\beta}$, which results in

$$\mu = \frac{\exp\{\boldsymbol{X}'\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{X}'\boldsymbol{\beta}\}}. \tag{12}$$

In (12), $\boldsymbol{X} = (1, X_1, \ldots, X_k)'$ is a vector of k explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is its respective vector of regression coefficients.

The distribution of the random variable T was reparametrized in order to ensure that one of its new parameters represents the probability of cure. The proposed reparameterization is given by

$$\begin{cases} \mu = \rho + (1-\rho)\phi \\ \sigma = \dfrac{\rho}{(1-\rho)\phi}, \end{cases} \tag{13}$$

which results in

$$\begin{cases} \phi = \dfrac{\mu}{\sigma(1-\mu)+1} \\ \rho = \dfrac{\mu\sigma}{\sigma+1}. \end{cases} \tag{14}$$

In the new parameterization proposed in (13), μ is a parameter that corresponds to the probability of cure and σ represents the ratio between the odds of the known cure and the proportion of the latent cure. Large values of σ indicate a greater known cure in relation to latent cure.

From (11), (12) and (14), the likelihood function can be rewritten as

$$L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2}) \propto \prod_{i=1}^{n} \left[\frac{\mu_i \sigma}{\sigma+1}\right]^{\delta_{2i}} [(1-\mu_i) f_Y(t_i;\boldsymbol{\theta})]^{(1-\delta_{2i})\delta_{1i}} \left[(1-\mu_i) S_Y(t_i;\boldsymbol{\theta}) + \frac{\mu_i}{\sigma+1}\right]^{(1-\delta_{2i})(1-\delta_{1i})}. \tag{15}$$

In (15), $\sigma > 0$, $\mu_i = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}$ and $S_Y(\cdot;\boldsymbol{\theta})$ and $f_Y(\cdot;\boldsymbol{\theta})$ are the survival and probability density functions of susceptible individuals, respectively. Here, $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)'$ is the vector of times observed in the sample with their respective indicators of censorship $\boldsymbol{\delta_1} = (\delta_{11}, \delta_{12}, \ldots, \delta_{1n})'$ and knowledge of the cure $\boldsymbol{\delta_2} = (\delta_{21}, \delta_{22}, \ldots, \delta_{2n})'$, $\boldsymbol{\theta}$ is the vector of parameters of the distribution of non-susceptible individuals, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ are the regression coefficients associated with the observed explanatory variables $\boldsymbol{x_i} = (1, x_{i1}, \ldots, x_{ik})'$, $i = 1, 2, \ldots, n$.

Applying the logarithm to the likelihood function (15), we get

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}; \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2};) &= \sum_{i=1}^{n} \left\{ \delta_{2i} \log\left[\frac{\mu_i\sigma}{\sigma+1}\right] + (1-\delta_{2i})\delta_{1i} \log\left[(1-\mu_i) f_Y(t_i;\boldsymbol{\theta})\right] \right. \\ &+ \left. (1-\delta_{2i})(1-\delta_{1i}) \log\left[(1-\mu_i) S_Y(t_i;\boldsymbol{\theta}) + \frac{\mu_i}{\sigma+1}\right] \right\} + c, \end{aligned} \tag{16}$$

where c is a constant that does not depend on σ, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

The likelihood equation is given by

$$\boldsymbol{U}(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \mathbf{0}. \tag{17}$$

Thus, the value $\widehat{\boldsymbol{\vartheta}} = (\widehat{\sigma}, \widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\beta}}')'$, which satisfies Equation (17), is the maximum likelihood estimator of the model parameters, which under appropriate regularity conditions has asymptotically a multivariate normal distribution with mean $\boldsymbol{\vartheta}$ and variance and covariance matrix given by

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}}) = \left[-\frac{\partial^2 \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} \bigg|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}} \right]^{-1} = \left[J(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}} \right]^{-1}. \tag{18}$$

The value of $\widehat{\boldsymbol{\vartheta}} = (\widehat{\sigma}, \widehat{\boldsymbol{\theta}}', \widehat{\boldsymbol{\beta}}')'$ and the observed information matrix $J(\boldsymbol{\vartheta})$ can be obtained numerically using computational optimization methods using the Newton–Raphson-type algorithm, which provides an accurate numerical approximation for this matrix. From

these results, it is possible to construct confidence intervals for the parameters and carry out significance tests on the model covariates.

3. Results and Discussions

3.1. Simulation Study

This section describes a simulation study conducted to investigate whether the usual asymptotic properties of maximum likelihood estimators are present. The performance of the proposed model was also evaluated in the presence of censored data.

The study of the proposed model was conducted considering simulated data in R software, version 4.3.3 [12]. The simulations performed in this work considered in the model a dichotomous covariate, x_1, generated from a Bernoulli distribution with a probability of success $p = 0.5$ and a numerical covariate, x_2, with a standard normal distribution. These covariates were included in the the probability of cure (4) taking into account a logit link function, i.e., $\mu = \frac{\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}}$, where $\boldsymbol{x} = (1, x_1, x_2)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. For a fixed value of σ, the values of ρ and ϕ are given by (14).

This simulation study considered that the time to failure (U) of susceptible individuals follows a Weibull distribution with shape parameter α and scale parameters λ. The Weibull distribution was chosen due to its popularity in modeling survival data. In addition, the time to censoring (V) follows a Weibull distribution with the same shape parameter α and scale parameters $\lambda_2 = \lambda\left(\frac{1-q}{q}\right)^{1/\alpha}$, where q denotes the proportion of censorship. This result is based on the fact that $q = P(V < U) = \frac{\lambda^\alpha}{\lambda^\alpha+\lambda_2^\alpha}$ when $U \sim Weibull(\alpha, \lambda)$ and $V \sim Weibull(\alpha, \lambda_2)$. The survival times and their respective censoring indicators can be obtained following the steps of Algorithm 1.

Algorithm 1: Obtaining the survival time of the proposed model

1. Define the values of β_0, β_1, β_2, σ, α, λ and q (censoring proportion) and set $i = 1$;
2. Generate the covariates:
   $x_{1i} \sim Bernoulli(0.5)$ and
   $x_{2i} \sim Normal(0, 1)$;
3. Calculate ρ and ϕ_i from Equation (14);
4. Generate $K_i \sim Bernoulli(\rho)$ and $C_i \sim Bernoulli(\phi_i)$;
5. Generate the time to failure $u_i \sim Weibull(\alpha, \lambda)$;
6. Generate the time to censoring $v_i \sim Weibull(\alpha, \lambda_2)$, where $\lambda_2 = \lambda\left(\frac{1-q}{q}\right)^{1/\alpha}$;
7. Define the indicator of the failure of susceptible individuals, R_i:
   If $u_i \leq v_i$, set $R_i = 1$
   Else, set $R_i = 0$;
8. Define the censoring indicators:
   $\delta_{1i} = (1 - K_i)(1 - C_i)R_i$ and
   $\delta_{2i} = K_i$;
9. Calculate the observed time $t_i = \delta_{1i}u_i + (1 - \delta_{1i})v_i$;
10. If $i < n$, set $i = i + 1$ and return to Step 2. Else end algorithm.

Note 1: In the case of no censoring ($q = 0$), the value of λ_2 in Step 6 can be specified in a way that $\lambda_2 >> \lambda$. For example, $\lambda_2 = 10^5\lambda$.

Note 2: The Step 9 implies that the censored times of non-susceptible individuals in which the cure is not known will be equal to v_i.

The survival time samples were simulated considering the values $\beta_1 = 1$, $\beta_2 = 0.5$, $\alpha = 2$, $\lambda = 3$ and several values of σ and β_0. The values of α and β_0 were defined in order to vary the proportions of known and latent cures (ρ and ϕ, respectively). A total of 5 simulation scenarios were considered, as shown in Table 1.

Table 1. Scenarios used in the simulations.

Scenario	σ	β_0	β_1	β_2	α	λ	$\widetilde{\phi}$	$\widetilde{\rho}$
S_1	1.0	−3.0	1.0	0.5	2.0	5.0	5%	5%
S_2	1.0	−2.0	1.0	0.5	2.0	5.0	12%	10%
S_3	1.5	−0.5	1.0	0.5	2.0	5.0	31%	30%
S_4	3.0	−1.0	1.0	0.5	2.0	5.0	15%	30%
S_5	0.5	−1.0	1.0	0.5	2.0	5.0	30%	13%

Note: $\widetilde{\phi}$ and $\widetilde{\rho}$ are, respectively, the expected values for the percentage of latent and non-latent cures based on the scenarios presented.

The mean estimates, the mean square error (MSE), and the coverage probability (CP) of the estimators were obtained from 1000 Monte Carlo replicates, considering sample sizes n = 50, 100, 200, and 500 and censoring percentages equal to 0%, 10%, and 30% (i.e., censoring susceptible individuals).

For the construction of confidence intervals (CI) for the calculation of the CP, a confidence level of 95% was considered. In addition, since σ and the parameters α and λ of the Weibull distribution are positive, a logarithmic transformation was applied for constructing the CIs of these parameters.

Analysis of the results shown in Figures 1–6, referring to Scenario 2—in which data histograms were plotted with normal distribution curves superimposed, along with the values of the average, mean square error (MSE) and coverage probability (CP)—provides evidence of the asymptotic normality of the estimators, regardless of the censoring percentage.

When the censoring percentage is equal to zero, the estimators have excellent statistical properties. The parameter estimates exhibit small bias and low variance, indicating consistency. In addition, the CPs remain close to the nominal confidence level, regardless of the parameter and sample size. The largest difference observed was 0.0440 (0.9940–0.9500), for the σ parameter with a sample size of n = 50.

In the presence of censored observations, there is an increase in the deviations of the estimates from the true parameter values, and this effect is more pronounced as the percentage of censoring increases. Consequently, the probability of coverage tends to move away from the nominal confidence level. However, this behavior is to be expected in censored contexts. It is important to note that as the sample size increases, even under high censoring percentages, the estimators once again show desirable properties. For example, for n = 500 and 30% censoring, the biggest difference observed was 0.0660 (1.066–1.000) in the average, 0.1035 in the MSE, and 0.0150 (0.9650–0.9500) in the CP, all three associated with the parameter σ.

Finally, the results showed evidence of the asymptotic normality of the estimators for all parameters and censorship percentages, as evidenced by the fit of the normal curves superimposed on the histograms, which improves with increasing sample size. This behavior justifies the use of normal approximations for constructing confidence intervals and testing hypotheses about the model parameters, even in scenarios with censoring. The results of the other scenarios were similar to those observed in Scenario 2. The numerical results for these scenarios are shown in Table 2.

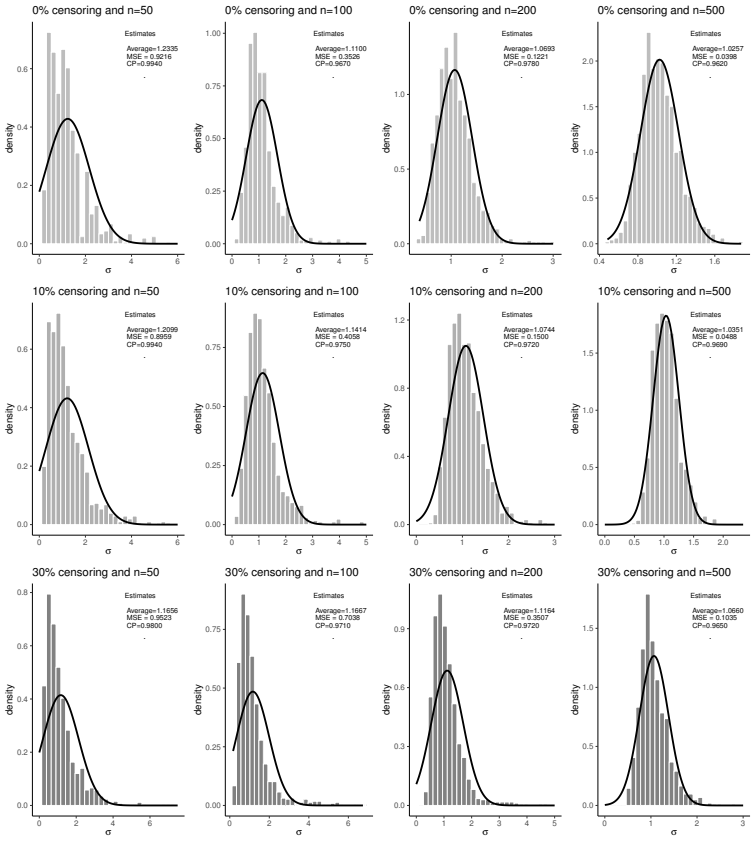

Figure 1. Empirical distribution of the estimates of σ from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

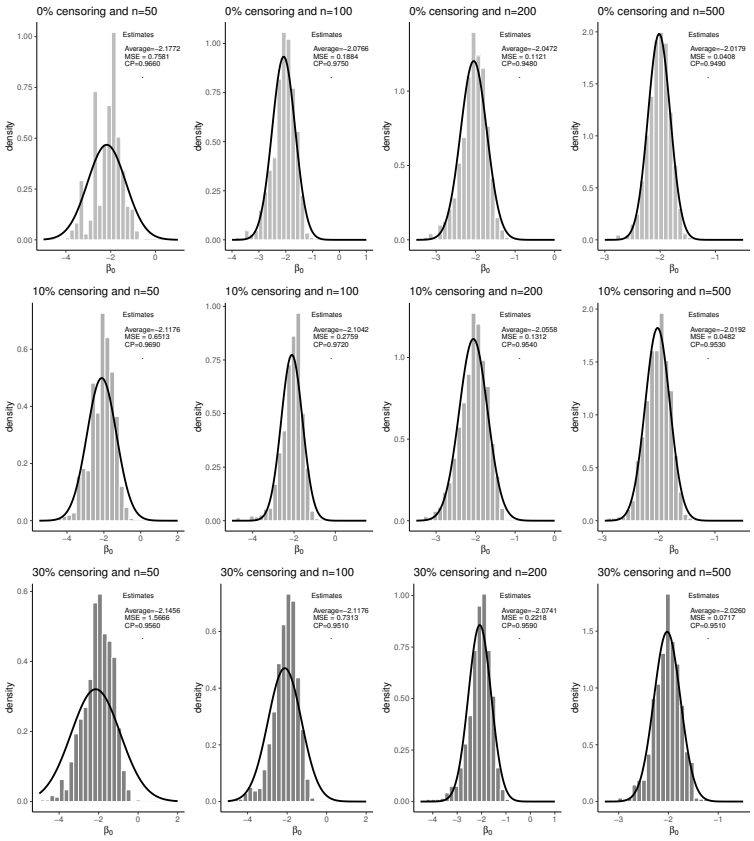

Figure 2. Empirical distribution of the estimates of β_0 from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

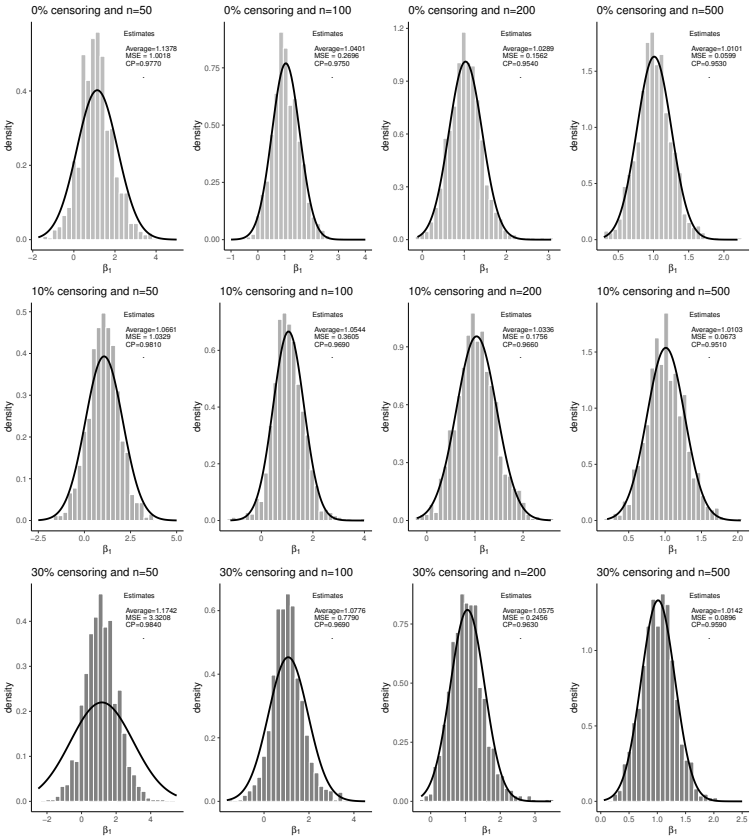

Figure 3. Empirical distribution of the estimates of β_1 from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

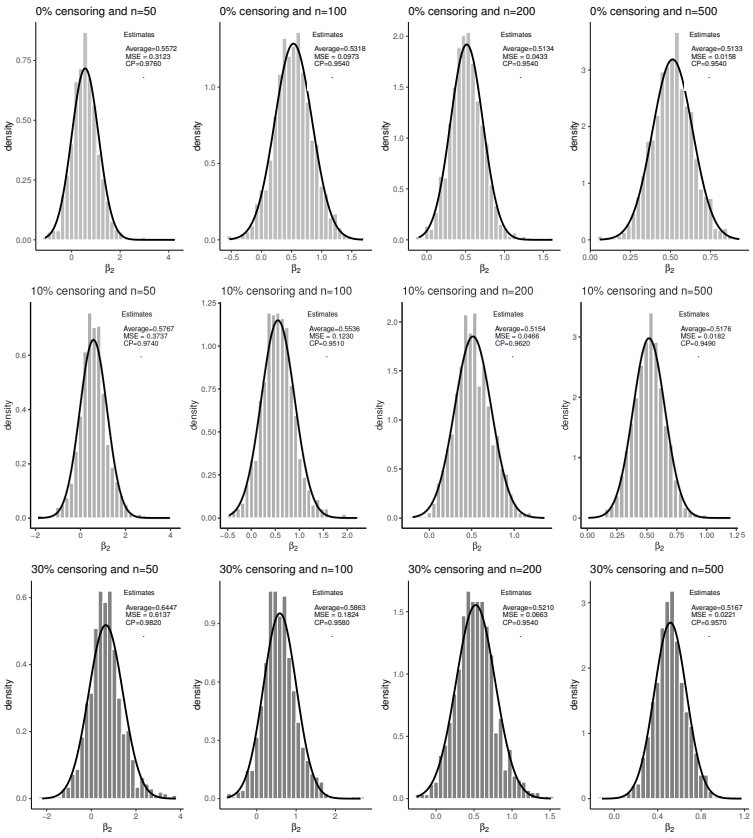

Figure 4. Empirical distribution of the estimates of β_2 from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

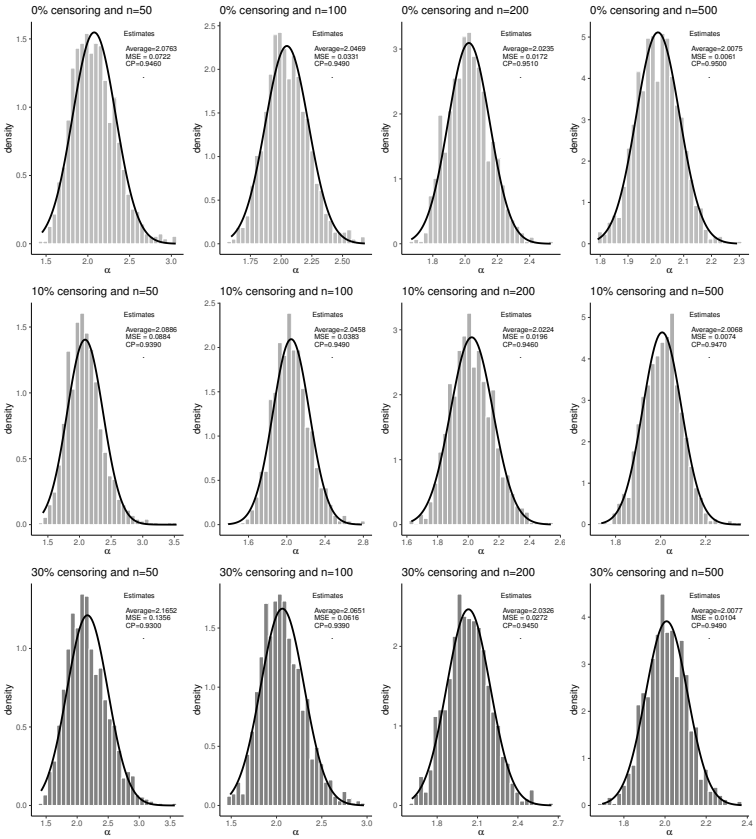

Figure 5. Empirical distribution of the estimates of α from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

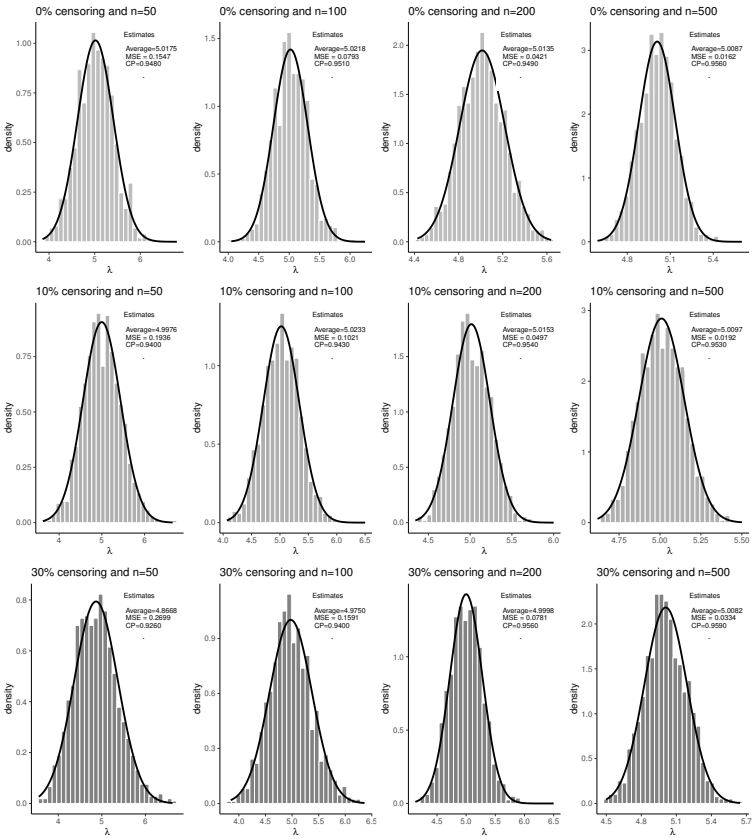

Figure 6. Empirical distribution of the estimates of λ from 1000 Monte Carlo replications with the fitted normal curve (Scenario 2).

Table 2. Average of estimates, MSE, and probability of coverage (CP) of parameters considering the simulation scenarios and different sample sizes and censoring percentages.

Cens. Per.	Par.	Scen.	*n*											
			50			100			200			500		
			Average	MSE	CP	Average	MSE	CP	Average	MSE	CP	Average	MSE	CP
0%	σ	S_1	1.3710	1.8469	0.9880	1.2491	1.2125	0.9960	1.1354	0.3518	0.9710	1.0540	0.1083	0.9580
		S_2	1.2335	0.9216	0.9940	1.1100	0.3526	0.9670	1.0693	0.1221	0.9780	1.0257	0.0398	0.9620
		S_3	1.6732	0.5973	0.9580	1.5933	0.2629	0.9470	1.5620	0.1094	0.9690	1.5260	0.0394	0.9580
		S_4	3.4201	5.4567	0.9790	3.2404	1.7302	0.9680	3.1478	0.7116	0.9760	3.0662	0.2258	0.9740
		S_5	0.5795	0.1021	0.9620	0.5305	0.0383	0.9520	0.5102	0.0158	0.9660	0.5045	0.0062	0.9440
	β_0	S_1	−3.3681	6.5650	0.9340	−3.1403	0.5361	0.9750	−3.0952	0.2522	0.9730	−3.0424	0.0866	0.9550
		S_2	−2.1772	0.7581	0.9660	−2.0766	0.1884	0.9750	−2.0472	0.1121	0.9480	−2.0179	0.0408	0.9490
		S_3	−0.5182	0.1624	0.9680	−0.5022	0.0797	0.9580	−0.5099	0.0403	0.9600	−0.5064	0.0159	0.9650
		S_4	−1.0742	0.2131	0.9720	−1.025	0.0960	0.9610	−1.0296	0.0515	0.9560	−1.0145	0.0212	0.9530
		S_5	−1.0555	0.1929	0.9820	−1.0109	0.0906	0.9660	−1.0209	0.0480	0.9560	−1.0013	0.0198	0.9590
	β_1	S_1	0.9545	4.2504	0.9920	1.0208	0.6388	0.9860	1.0219	0.3278	0.9700	1.0152	0.1103	0.9700
		S_2	1.1378	1.0018	0.9770	1.0401	0.2696	0.9750	1.0289	0.1562	0.9540	1.0101	0.0599	0.9530
		S_3	1.0426	0.3891	0.9630	1.0202	0.1912	0.9580	1.0273	0.0954	0.9490	1.0164	0.0368	0.9420
		S_4	1.0947	0.4374	0.9590	1.0325	0.2171	0.9500	1.0354	0.0997	0.9470	1.0191	0.0407	0.9400
		S_5	1.0430	0.3964	0.9660	1.0093	0.1981	0.9520	1.0390	0.0924	0.9540	1.0056	0.0379	0.9470
	β_2	S_1	0.5596	0.5837	0.9810	0.5609	0.2174	0.9600	0.5144	0.0737	0.9580	0.5186	0.0304	0.9480
		S_2	0.5572	0.3123	0.9760	0.5318	0.0973	0.9540	0.5134	0.0433	0.9540	0.5133	0.0158	0.9540
		S_3	0.5473	0.1935	0.9650	0.5380	0.0622	0.9530	0.5261	0.0295	0.9490	0.5107	0.0103	0.9510
		S_4	0.5666	0.2074	0.9710	0.5275	0.0672	0.9450	0.5227	0.0299	0.9550	0.5107	0.0110	0.9530
		S_5	0.5556	0.1980	0.9710	0.5400	0.0607	0.9660	0.5205	0.0313	0.9480	0.5107	0.0105	0.9530
	α	S_1	2.0725	0.0653	0.9470	2.0363	0.0283	0.9540	2.0172	0.0156	0.9370	2.0061	0.0053	0.9450
		S_2	2.0763	0.0722	0.9460	2.0469	0.0331	0.9490	2.0235	0.0172	0.9510	2.0075	0.0061	0.9500
		S_3	2.1217	0.1284	0.9320	2.0664	0.0568	0.9410	2.0349	0.0267	0.9370	2.0172	0.0093	0.9500
		S_4	2.1047	0.0975	0.9350	2.0543	0.0459	0.9420	2.0312	0.021	0.9500	2.0121	0.0080	0.9500
		S_5	2.1186	0.1076	0.9340	2.0651	0.0447	0.9450	2.0308	0.0203	0.9470	2.0126	0.0078	0.9610
	λ	S_1	5.0184	0.1564	0.9400	5.0157	0.0728	0.9490	5.0084	0.0391	0.9450	5.0048	0.0146	0.9580
		S_2	5.0175	0.1547	0.9480	5.0218	0.0793	0.9510	5.0135	0.0421	0.9490	5.0087	0.0162	0.9560
		S_3	5.0137	0.2481	0.9410	5.0318	0.1307	0.9530	5.0133	0.0607	0.9510	5.0088	0.0274	0.9500
		S_4	5.0272	0.2071	0.9480	5.0150	0.1125	0.9460	5.0173	0.0513	0.9550	5.0040	0.0225	0.9470
		S_5	5.0245	0.1903	0.9500	5.0297	0.0989	0.9520	5.0147	0.0522	0.9460	5.0084	0.0218	0.9540
10%	σ	S_1	1.2154	1.5844	0.9970	1.1788	0.9440	0.9940	1.1843	0.6120	0.9820	1.0666	0.1404	0.9630
		S_2	1.2099	0.8959	0.9940	1.1414	0.4058	0.9750	1.0744	0.1500	0.9720	1.0351	0.0488	0.9690
		S_3	1.6926	0.6867	0.9810	1.6120	0.2889	0.9540	1.5597	0.1193	0.9660	1.5353	0.0459	0.9550
		S_4	3.3084	4.3520	0.9740	3.2556	2.1547	0.9740	3.1729	0.8767	0.9750	3.0969	0.2928	0.9710
		S_5	0.5911	0.1344	0.9640	0.5345	0.0414	0.9580	0.5121	0.0162	0.9580	0.5062	0.0068	0.9450
	β_0	S_1	−3.2250	6.5393	0.9450	−3.1851	1.3014	0.9700	−3.1135	0.3746	0.9690	−3.0493	0.1088	0.9590
		S_2	−2.1176	0.6513	0.9690	−2.1042	0.2759	0.9720	−2.0558	0.1312	0.9540	−2.0192	0.0482	0.9530
		S_3	−0.5378	0.1959	0.9640	−0.5133	0.0925	0.9580	−0.5138	0.0455	0.9620	−0.5096	0.0181	0.9640
		S_4	−1.0401	0.2475	0.9650	−1.0198	0.1115	0.9490	−1.0231	0.0568	0.9540	−1.0089	0.0215	0.9570
		S_5	−1.0707	0.2742	0.9810	−1.0300	0.1091	0.9630	−1.017	0.0563	0.9630	−1.0091	0.0235	0.9570
	β_1	S_1	0.9581	4.7612	0.9870	1.0900	1.4194	0.9860	1.0435	0.4529	0.9730	1.0066	0.1391	0.9720
		S_2	1.0661	1.0329	0.9810	1.0544	0.3605	0.9690	1.0336	0.1756	0.9660	1.0103	0.0673	0.9510
		S_3	1.0818	0.4512	0.9510	1.0378	0.2198	0.9520	1.0309	0.1042	0.9440	1.0199	0.0403	0.9540
		S_4	1.0739	0.4803	0.9560	1.0396	0.2389	0.9520	1.0231	0.1100	0.9480	1.0146	0.0425	0.9560
		S_5	1.0751	0.5271	0.9660	1.0301	0.2334	0.9590	1.0290	0.1079	0.9550	1.0101	0.0432	0.9560
	β_2	S_1	0.5797	0.9074	0.9760	0.5615	0.2597	0.9630	0.5217	0.0918	0.9540	0.5121	0.0371	0.9560
		S_2	0.5767	0.3737	0.9740	0.5536	0.1230	0.9510	0.5154	0.0466	0.9620	0.5176	0.0182	0.9490
		S_3	0.5850	0.2265	0.9670	0.5439	0.0717	0.9600	0.5257	0.0328	0.9510	0.5115	0.0117	0.9590
		S_4	0.5957	0.2584	0.9620	0.5392	0.0746	0.9560	0.5286	0.0330	0.9590	0.5087	0.0119	0.9570
		S_5	0.6122	0.2545	0.9730	0.5385	0.0751	0.9620	0.5273	0.0347	0.9560	0.5143	0.0124	0.9590
	α	S_1	2.0870	0.0716	0.9510	2.0410	0.0330	0.9560	2.0185	0.0168	0.9380	2.0052	0.0062	0.9490
		S_2	2.0886	0.0884	0.9390	2.0458	0.0383	0.9490	2.0224	0.0196	0.9460	2.0068	0.0074	0.9470
		S_3	2.1321	0.1642	0.9200	2.0728	0.0671	0.9400	2.0414	0.0309	0.9420	2.0170	0.0112	0.9550
		S_4	2.1300	0.1407	0.9190	2.0577	0.0526	0.9430	2.0280	0.0248	0.9420	2.0141	0.0100	0.9520
		S_5	2.1333	0.1357	0.9240	2.0682	0.0529	0.9390	2.0377	0.0252	0.9500	2.0145	0.0093	0.9490
	λ	S_1	4.9806	0.1624	0.9400	5.0018	0.0836	0.9490	5.0068	0.0422	0.9580	5.0029	0.0165	0.9550
		S_2	4.9976	0.1936	0.9400	5.0233	0.1021	0.9430	5.0153	0.0497	0.9540	5.0097	0.0192	0.9530
		S_3	5.0235	0.2918	0.9430	5.0168	0.1504	0.9580	5.0196	0.0782	0.9400	5.0178	0.0337	0.9470
		S_4	5.0132	0.2472	0.9340	5.0125	0.1197	0.9550	5.0122	0.0648	0.9500	5.0043	0.0254	0.9530
		S_5	5.0442	0.2466	0.9450	5.0295	0.1313	0.9400	5.0155	0.0628	0.9510	5.0100	0.0275	0.9480
30%	σ	S_1	1.1062	2.2105	0.9710	1.0792	1.1345	0.9690	1.1302	0.7752	0.9680	1.1337	0.3702	0.9680
		S_2	1.1656	0.9523	0.9800	1.1667	0.7038	0.9710	1.1164	0.3507	0.9720	1.0660	0.1035	0.9650
		S_3	1.7216	0.9721	0.9750	1.6745	0.6094	0.9630	1.5967	0.1942	0.9670	1.5572	0.0723	0.9600
		S_4	3.0242	4.1212	0.9450	3.2559	3.8182	0.9670	3.2553	1.9640	0.9730	3.1657	0.7049	0.9730
		S_5	0.6328	0.2201	0.9770	0.5271	0.0877	0.9660	0.5232	0.0240	0.9700	0.5100	0.0093	0.9410
	β_0	S_1	−3.5985	16.9135	0.9000	−3.2110	3.6533	0.9430	−3.0845	0.5922	0.9590	−3.0518	0.1701	0.9600
		S_2	−2.1456	1.5666	0.9560	−2.1176	0.7313	0.9510	−2.0741	0.2218	0.9590	−2.0260	0.0717	0.9510
		S_3	−0.4919	0.3014	0.9500	−0.5270	0.1350	0.9540	−0.5131	0.0617	0.9610	−0.5139	0.0243	0.9650
		S_4	−0.9959	0.3337	0.9550	−1.0086	0.1298	0.9620	−1.0243	0.0709	0.9600	−1.0120	0.0273	0.9600
		S_5	−1.0756	0.4171	0.9670	−1.0774	0.2440	0.9520	−1.0309	0.1026	0.9650	−1.0173	0.0414	0.9490
	β_1	S_1	0.8974	11.2328	0.9760	1.0723	2.8033	0.9890	1.0637	0.6524	0.9720	1.0227	0.1880	0.9630
		S_2	1.1742	3.3208	0.9840	1.0766	0.7790	0.9690	1.0575	0.2456	0.9630	1.0142	0.0896	0.9590
		S_3	1.1063	0.6261	0.9580	1.0884	0.3041	0.9590	1.0245	0.1325	0.9580	1.0196	0.0486	0.9600
		S_4	1.1503	0.6570	0.9560	1.0643	0.2656	0.9560	1.0447	0.1301	0.9550	1.0216	0.0485	0.9560
		S_5	1.0921	0.6976	0.9810	1.0690	0.3869	0.9460	1.0390	0.1588	0.9560	1.0073	0.0582	0.9600
	β_2	S_1	0.5918	1.5797	0.9240	0.6365	0.4378	0.9690	0.5353	0.1197	0.9640	0.5201	0.0481	0.9540
		S_2	0.6477	0.6137	0.9820	0.5863	0.1824	0.9580	0.5210	0.0663	0.9540	0.5167	0.0221	0.9570
		S_3	0.6290	0.3468	0.9690	0.5613	0.0981	0.9530	0.5260	0.0434	0.9460	0.5132	0.0146	0.9460
		S_4	0.6152	0.3375	0.9680	0.5438	0.0950	0.9420	0.5253	0.0355	0.9640	0.5120	0.0139	0.9510
		S_5	0.6265	0.4351	0.9830	0.5748	0.1286	0.9610	0.5335	0.0522	0.9490	0.5213	0.0176	0.9590
	α	S_1	2.1724	0.1398	0.9240	2.0832	0.0569	0.9320	2.0336	0.0216	0.9530	2.0070	0.0086	0.9600
		S_2	2.1652	0.1356	0.9300	2.0651	0.0616	0.9390	2.0326	0.0272	0.9450	2.0077	0.0104	0.9490
		S_3	2.2008	0.2610	0.9090	2.0991	0.1030	0.9310	2.0518	0.0449	0.9470	2.0185	0.0179	0.9350
		S_4	2.2064	0.2160	0.9220	2.0892	0.0823	0.9360	2.0391	0.0333	0.9580	2.0099	0.0142	0.9430
		S_5	2.1610	0.1840	0.9240	2.0801	0.0813	0.9320	2.0427	0.0367	0.9480	2.0146	0.0134	0.9530
	λ	S_1	4.8033	0.2512	0.9100	4.9119	0.1367	0.9300	4.9648	0.0612	0.9510	4.9956	0.0264	0.9580
		S_2	4.8668	0.2699	0.9260	4.9750	0.1591	0.9400	4.9998	0.0781	0.9560	5.0082	0.0334	0.9590
		S_3	4.9333	0.4564	0.9290	5.0227	0.2816	0.9440	5.0163	0.1388	0.9450	5.0163	0.0650	0.9410
		S_4	4.8761	0.3553	0.9280	4.9811	0.2026	0.9400	5.0069	0.1046	0.9480	5.0062	0.0436	0.9530
		S_5	4.9788	0.4640	0.9290	5.0235	0.2677	0.9420	5.0125	0.1282	0.9510	5.0170	0.0575	0.9450

3.2. An Illustrative Example

This section presents an illustration of the application of the proposed model to an artificial dataset of clients who take a loan at a financial institution. The decision to use artificial data for this illustration is due to the high commercial value associated with credit risk data and also due to legal factors that prevent the disclosure of sensitive customer information. Given these limitations, it has been decided to consult the existing literature in order to identify the variables most frequently used to classify low- and high-risk applicants in the context of credit risk analysis. Based on this investigation, the following explanatory variables were selected: (1) credit limit: is the maximum amount of credit a lender authorizes to each customer (log scale); (2) gender (female and male); (3) social class, coded into 5 levels (class A to class E); (4) marital status, coded into 3 levels (single, married, and widowed/separated) and; (5) age in years.

Based on these variables, a simulated database was created with 1000 observations, also including the time (in months) until the occurrence of default. In these studies, it has been observed that censorship rates are generally high, resulting in a low number of registered defaults. In this context, this study adopted a censorship rate of 50% in order to approximate the real data reported in the literature in studies such as [13]. This strategy allowed the simulated base to more faithfully represent the characteristics observed in real credit risk analysis scenarios.

In this study, the explanatory variables were generated as follows: credit limit by a lognormal distribution, gender by a Bernoulli distribution, social classification by a multinomial distribution with five categories, marital status by a multinomial distribution with three categories, and age by a Poisson distribution. Survival times were generated using Algorithm 1 described in Section 3. In addition, the time values were truncated at $t = 60$. In these cases, the indicators of censorship and knowledge cure were defined as $\delta_1 = 0$ and $\delta_2 = 1$, respectively. This procedure was adopted to represent loans with terms up to 60 months. Here, $\delta_2 = 1$ indicates that default will not occur (an individual who is known to be cured) either because they have paid off the loan early or because they have completed the entire loan period without missing a payment.

The simulated dataset is provided in Table S1 of the Supplementary Materials. The sample exhibited a censoring rate of 54.4%, of which 44.0% correspond to individuals known to be cured ($\delta_2 = 1$).

The application data was adjusted using the proposed model considering that the time to default of susceptible individuals follows a Weibull distribution with a shape parameter α and a scale parameter λ, and the covariates were included in the probability of cure (4), taking into account a logit link function. The maximum likelihood point estimates of the parameters, with their respective confidence intervals, are shown in Table 3.

The results in Table 3 indicate that, except for age, all covariates significantly influence the cure at the 5% significance level. The probability of cure for the clients was calculated using (12), based on the obtained estimates and considering all covariates. This probability of cure was used to classify clients as having a low or high risk of default. The cutoff point considered was 0.456, that is, a customer is classified as having a high risk of default if their probability of cure, $\mu < 0.456$ (or classified as low risk if $\mu \geq 0.456$).

The definition of the cut-off point took into account the proportion of defaulting customers in the sample. In the observed sample, a customer is considered to be in default when $\delta_1 = 0$. This cutoff point resulted in correct predictions, sensitivity (probability of the model classifying a good customer as low risk) and specificity (probability of the model classifying a defaulting customer as having a high risk) of 70.3%, 56.4% , and 82.0%, respectively, indicating a good classification of customers, particularly the defaulting ones.

In fact, the cut-off point can be adjusted to increase or decrease sensitivity or specificity in order to control the error that the financial institution considers to be the most critical.

Table 3. Point and interval estimates of the parameters of the proposed model.

Covariates	Parameter	Estimate	Standard Error	95% CI
Intercept	β_0	−4.040	1.002	(−6.004; −2.076)
Credit limit	β_1	0.209	0.104	(0.005; 0.413)
Gender [1]				
Male	β_2	1.027	0.142	(0.749; 1.305)
Social class [2]				
B	β_{31}	0.433	0.221	(−0.001; 0.867)
C	β_{32}	0.863	0.226	(0.419; 1.306)
D	β_{33}	0.765	0.224	(0.326; 1.204)
E	β_{34}	1.423	0.224	(0.985; 1.861)
Marital status [3]				
Married	β_{41}	1.325	0.174	(0.984; 1.666)
widowed/separated	β_{42}	0.959	0.174	(0.618; 1.301)
Age	β_5	0.013	0.011	(−0.009; 0.035)
	σ	3.386	0.312	(2.827; 4.056) [4]
	α	1.283	0.043	(1.202; 1.367) [4]
	λ	9.956	0.378	(9.243; 10.725) [4]

Reference level of the covariates: [1] Female; [2] Class A; [3] Single; [4] A logarithmic transformation was applied for constructing the CIs.

For benchmarking purposes, a logistic regression model and the Weibull Berkson and Gage mixture model were used to classify customers in this data set. The results of the logistic regression (correct predictions of 71.7%, sensitivity of 56.4%, and specificity of 82.4%) and the Berkson and Gage model (correct predictions of 70.2%, sensitivity of 54.4% and specificity of 83.5%) were similar to those of the proposed model. Figure 7 shows the probabilities of cure estimated by the proposed model, logistic regression, and Berkson and Gage mixture model, indicating agreement among these three methodologies.

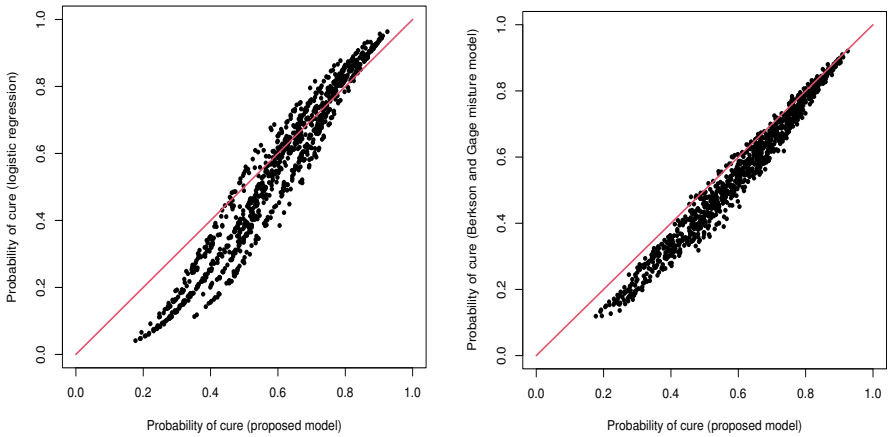

Figure 7. Probabilities of cure estimated by the proposed model, logistic regression model, and Berkson and Gage mixture model.

The results show that the model proposed in this study achieved performance similar to that of the benchmark models. Under a naive comparison, the overall accuracy rate of the proposed model was 70.3%, while the logistic regression and the Berkson and Gage mixture models achieved rates of 71.7% and 70.2%, respectively.

However, it is important to note that, despite the similarity in results to the logistic regression model, the proposed model also accounts for the time until default for susceptible

clients. Moreover, the logistic model is not fully appropriate in this context, as there are latent cured (non-defaulting) clients for whom it is not known with certainty whether they are truly cured. In addition, as expected, the Weibull mixture model underestimates the cure probabilities. This occurs because the Berkson and Gage model ignores the presence of known cures, potentially considering some clients who are clearly cured as non-cured individuals.

4. Conclusions

This work presents an extension of the Berkson and Gage mixture model [2] that accommodates both latent and non-latent cure fractions. This model can be viewed as a cure rate mixture model with competing risks, considering the non-latent cure as a competing cause.

The proposed model was reparameterized in terms of the cure probability, with a regression structure attached to this parameter. Furthermore, a simulation study was conducted considering a Weibull distribution for the time to the event of susceptible (non-cured) individuals. The results of these simulated data provide evidence of the asymptotic properties of the estimators.

The proposed model is illustrated using a synthetic dataset of customers who have taken out loans from a financial institution, and its performance is compared with that of logistic regression and the standard Berkson and Gage mixture model. The results show that the proposed model not only achieves competitive accuracy but also offers important conceptual and practical advantages, since the logistic model ignores the time-to-event distribution and the Berkson and Gage model neglects known cure information when such data are available.

It is important to note that any other probability distribution could be used to model the time to event of susceptible individuals. Moreover, a regression structure, such as a proportional hazards or accelerated failure time model, can also be employed to incorporate covariates into the modeling of this time. In addition, future works could consider the implementation of informative censoring mechanisms, which are commonly encountered in credit risk modeling.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/stats8030082/s1, Table S1. Artificial dataset of 1000 clients who take a loan in a financial institution.

Author Contributions: Conceptualization, E.Y.N., F.M.A. and M.R.P.C.; methodology, E.Y.N., F.M.A. and M.R.P.C.; software, E.Y.N., F.M.A. and M.R.P.C.; validation, E.Y.N., F.M.A. and M.R.P.C.; formal analysis, E.Y.N., F.M.A. and M.R.P.C.; investigation, E.Y.N., F.M.A. and M.R.P.C.; resources, E.Y.N., F.M.A. and M.R.P.C.; data curation, E.Y.N. and M.R.P.C.; writing—original draft preparation, E.Y.N., F.M.A. and M.R.P.C.; writing—review and editing, E.Y.N., F.M.A. and M.R.P.C.; visualization, E.Y.N., F.M.A. and M.R.P.C.; supervision, E.Y.N.; project administration, E.Y.N.; funding acquisition, E.Y.N. and F.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES)—Finance Code 001, National Council for Scientific and Technological Development (CNPq), Editais de Auxílio Financeiro DPI/DPG/UnB, DPI/DPG/BCE/UnB, and PPGEST/UnB.

Data Availability Statement: The data presented in this study are available in the Supplementary Materials. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, M.H.; Ibrahim, J.G.; Sinha, D. A new Bayesian model for survival data with a surviving fraction. *J. Am. Stat. Assoc.* **1999**, *94*, 909–919. [CrossRef] [PubMed]
2. Berkson, J.; Gage, R.P. Survival curve for cancer patients following treatment. *J. Am. Stat. Assoc.* **1952**, *47*, 501–515. [CrossRef]
3. Yakovlev, A.Y.; Tsodikov, A.D. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*; World Scientific: Singapore, 1996.
4. Oliveira, M.R.; Moreira, F.; Louzada, F. The zero-inflated promotion cure rate model applied to financial data on time-to-default. *Cogent Econ. Financ.* **2017**, *5*, 1395950. [CrossRef]
5. Chen, T.; Du, P. Promotion time cure rate model with nonparametric form of covariate effects. *Stat. Med.* **2018**, *37*, 1625–1635. [CrossRef] [PubMed]
6. Gómez, Y.M.; Gallardo, D.I.; Bourguignon, M.; Bertolli, E.; Calsavara, V.F. A general class of promotion time cure rate models with a new biological interpretation. *Lifetime Data Anal.* **2023**, *29*, 66–86. [CrossRef] [PubMed]
7. Leão, J.; Leiva, V.; Saulo, H.; Tomazella, V. Incorporation of frailties into a cure rate regression model and its diagnostics and application to melanoma data. *Stat. Med.* **2018**, *37*, 4421–4440. [CrossRef] [PubMed]
8. Cancho, V.G.; Barriga, G.; Leão, J.; Saulo, H. Survival model induced by discrete frailty for modeling of lifetime data with long-term survivors and change-point. *Commun. Stat.-Theory Methods* **2021**, *50*, 1161–1172. [CrossRef]
9. Maller, R.A.; Zhou, X. *Survival Analysis with Long-Term Survivors*; Wiley: Hoboken, NJ, USA, 1996.
10. Rodrigues, J.; Cancho, V.G.; de Castro, M.; Louzada-Neto, F. On the unification of long-term survival models. *Stat. Probab. Lett.* **2009**, *79*, 753–759. [CrossRef]
11. Larson, M.G.; Dinse, G.E. A Mixture Model for the Regression Analysis of Competing Risks Data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1985**, *34*, 201–2011. [CrossRef]
12. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: https://www.R-project.org/ (accessed on 1 September 2025).
13. Dirick, G.C.L.; Baesens, B. Time to default in credit scoring using survival analysis: A benchmark study. *J. Oper. Res. Soc.* **2017**, *68*, 652–665. [CrossRef]