UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
OPEN ACCESS

Ci. e Nat., Santa Maria, v. 48, e88599, 2026 • https://doi.org/10.5902/2179460X88599
Submissão: 13/08/2024 • Aprovação: 21/10/2025 • Publicação: 01/07/2026

Estatística

Avaliação da evasão escolar de alunos de um curso de graduação via um modelo de regressão log-logística discreta

Assessment of school evasion rates among undergraduate students using a discrete log-logistic regression model

Damião Flávio dos Santos[I], **Cira Etheowalda Guevara Otiniano**[I], **Juliana Gomes**[I]

[I] Universidade de Brasília, Brasília, DF, Brasil

RESUMO

A evasão escolar, tanto no ensino fundamental e médio, como no ensino superior, é certamente um problema que afeta os resultados do sistema educacional no Brasil. Além de ser um problema que afeta diretamente os gastos públicos do país, pouco se fala sobre o assunto e muito menos existe uma política de combate à evasão. Em busca de algumas características que expliquem o tempo até que a evasão ocorra, uma das técnicas que pode ser utilizada é a análise de sobrevivência. Tradicionalmente, para modelar o tempo até a ocorrência de um evento de interesse, distribuições de probabilidade contínuas são amplamente utilizadas. Porém, quando esse tempo é observado apenas em dias, meses ou anos, distribuições de probabilidade discretas são mais apropriadas. Para modelar e explicar o tempo de evasão dos alunos do curso de Ciência da Computação da Universidade Estadual da Paraíba (UEPB) – Campus I, propomos neste trabalho o modelo de regressão Log-Logística discreta. Os parâmetros do modelo proposto foram estimados pelo método da máxima verossimilhança. Para avaliar a acurácia dos estimadores, o viés e o erro quadrático médio dos estimadores são calculados por meio de simulação de Monte Carlo, considerando diferentes tamanhos de amostra e porcentagens de censura. A partir dos coeficientes estimados, constatou-se que a idade, a instituição de ensino médio, a forma de ingresso e o tempo de estudo do aluno no curso são fatores que influenciam o tempo até a sua evasão.

Palavras-chave: Log-logística discreta; Regressão; Evasão; EMV

ABSTRACT

School evasion, both in primary and secondary education, as well as in higher education, is certainly a problem that affects the results of the educational system in Brazil. In addition to being a problem that directly affects the country's public spending, little is said about the subject and much less is there a policy to combat evasion. In search of some characteristics that explain the time until evasion occurs, one of

the techniques that can be used is survival analysis. Traditionally, to model the time until the occurrence of an event of interest, continuous probability distributions are widely used. However, when this time is observed only in days, months or years, discrete probability distributions are more appropriate. To model and explain the dropout time of students in the Computer Science course at the Universidade Estadual da Paraíba – Campus I, we propose in this work the discrete Log-Logistic regression model. The parameters of the proposed model were estimated using the maximum likelihood method. To evaluate the accuracy of the estimators, the bias and mean squared error of the estimators are calculated using Monte Carlo simulation, considering different sample sizes and censoring percentages. From the estimated coefficients, it was found that age, the secondary education institution, the form of entry and the student's study time on the course are factors that influence the time until your evasion.

Keywords: Discrete log-logistics; Regression; Evasion; EMV

1 INTRODUÇÃO

No Brasil, a evasão escolar é um dos maiores problemas enfrentados na educação, principalmente durante o período de pandemia da Covid-19. Essa evasão é generalizada tanto a nível de Ensino Fundamental e Médio, quando a Superior. De acordo com os dados do Censo da Educação Superior (INEP, 2020), a taxa de evasão do Ensino Superior em 2020 foi de 33,2%. Já de acordo com os dados do Mapa do ensino superior no Brasil 2022 (SEMESP, 2022), em 2021 a taxa de evasão de alunos do Ensino Superior, de forma presencial e de Educação a distância (EAD) foi de 36,6%.

A Paraíba tem uma população de 3.974.687 habitantes, distribuída por 223 municípios e quatro mesorregiões (IBGE, 2022), possui 41 Instituições de Ensino Superior (IES) com cursos presenciais e 61 com cursos de EAD, segundo o Mapa do Ensino Superior no Brasil de 2022 (SEMESP, 2022). De acordo com os dados apresentados, de 2019 para 2020 (SEMESP, 2022), o número de concluintes em cursos presenciais caiu 18,8% na Paraíba: 13,8% na rede privada e 25,9% na rede pública, algo que pode ser explicado pela demora das IES públicas em adotarem as aulas remotas, atrasando a conclusão dos cursos dos alunos. Em relação aos cursos de ciência da computação na Paraíba, com base nos dados do censo da educação superior (INEP, 2020), calculou-se a taxa de evasão em 2020 e obteve-se o valor de 27,4%, um pouco abaixo da taxa nacional, ao considerar todos os cursos de computação no país (32,7%), mas, ainda

assim, é uma taxa de evasão elevada. Conforme o SEMESP (2022), a taxa de evasão é calculada considerando o total de alunos que deixaram o curso — seja por trancamento, desvinculação ou falecimento — em relação ao total de vínculos existentes durante o período, incluindo tanto as matrículas ativas quanto as que foram desligadas.

Dado as altas taxas de evasão, é importante entender o motivo desse número expressivo. Por isso, o objetivo deste artigo é analisar e explicar a evasão escolar dos alunos dos cursos de Ciência da Computação da Universidade Estadual da Paraíba, na Paraíba, Brasil, no período 2010 a 2016 (dados disponibilizados). Para alcançar tal objetivo é fundamental definir o conceito de evasão que será considerado neste trabalho. É possível observar que esse conceito é distinto em alguns trabalhos (Rosa, 2007; Filho *et al*., 2007 e Fritsch *et al*., 2015). Dessa forma, este estudo considera evasão do curso de Ciência da Computação quando o estudante sai do curso antes do término (desligamento por não cumprir as regras da universidade, mudança de curso, etc) ou simplesmente abandonou o curso. Assim sendo, é necessário estudar as razões que elevam essa taxa de evasão. Uma forma de estudar a evasão no Ensino Superior é medir o tempo até a evasão, isto é, usar técnicas de análise de sobrevivência para estudar o tempo desde o ingresso na universidade até a evasão. Neste sentido, os livros de Hosmer e Lemeshow (1999), Louzada-Neto e Pereira (2000), Collet (2003), Colosimo e Giolo (2024), Lawless (2011) tornam-se importantes à esta análise.

Por outro lado, um modelo de regressão é uma técnica estatística de análise de dados que é usada para modelar o relacionamento entre duas ou mais variáveis explicativas (regressoras ou covariáveis) e uma variável de resposta. Um estudo sobre evasão escolar de um curso de Física da Universidade Federal do Rio Grande do Sul (UFRGS) foi realizado por Lima Junior, Silveira e Ostermann (2012). Nessa pesquisa, eles aplicaram técnicas não-paramétricas e semi-paramétricas de modelos de regressão. Porém, modelos de regressão paramétricos são mais eficientes na modelagem e amplamente utilizados para modelar e explicar dados de diversas áreas. As denominações dos modelos de regressão estão associadas à distribuição de

probabilidade paramétrica que ajusta os dados em estudo. Por exemplo, se os dados correspondentes a uma variável aleatória *T* são do tempo até a evasão e *T* é bem ajustada por uma distribuição Weibull, então o modelo é denominado de Modelo de Regressão Weibull. As distribuições mais utilizadas para modelar eventos de interesse são a distribuição Weibull, Logística, Log-logística, Log-Normal, Kumaraswamy, Burr, e extensões dessas distribuições (Lawless, 2011).

Além disso, como os dados sobre o tempo de evasão utilizado estão em semestres do curso, foi necessário considerar uma distribuição discreta para o tempo de evasão. Para identificar a distribuição mais adequada de *T*, em que *T* representa o tempo até a evasão escolar dos alunos dos cursos de Ciência da Computação na Paraíba, Brasil, primeiramente realizamos um estudo de adequação não paramétrico comparando as possíveis distribuições Log-logística discreta, aqui definida, com a Weibull discreta de Nakagawa e Osaki (1975). Esse estudo apontou que a distribuição Log-logística discreta é mais adequada.

No caso contínuo, Modelos de regressão Log-logística são aplicadas a diversas áreas. A lista de trabalhos é extensa, aqui são citados alguns dos mais recentes, sendo eles: Aplicado a análise de sobrevivência (Bennett, 1983), para a medicina (Hashemian *et al.*, 2017; Tahir *et al.*, 2014), para a engenharia (Ashkar; Mahdi, 2006; Khaleeqa *et al.*, 2022), e para a física (Aldahlan, 2020). No entanto, o modelo de regressão Log-logística discreta foi inicialmente estudado por Santos (2017). Neste artigo é apresentado parte dos resultados desse trabalho, a partir dele propõe-se um modelo de regressão Log-Logística discreta para o estudo da evasão dos alunos de graduação em Ciência da Computação na Paraíba. Trabalhos decorrentes de Santos (2017), correspondem a Lima (2018), com uma aplicação do modelo para dados de políticas de zoneamento e uso do solo.

O artigo foi dividido em 4 seções, após a introdução, a Seção 2 apresenta a distribuição Log-Logística discreta com algumas medidas descritivas como momentos e função quantil. Além disso, foi realizada uma reparametrização no parâmetro de escala, originando o modelo de regressão Log-logística discreta, e em seguida é descrito o método para a construção dos

intervalos de confiança para as estimativas dos parâmetros. Na seção 3 foi realizado um estudo da qualidade das estimativas de máxima verossimilhança, tanto para o modelo sem covariáveis, quanto com uma covariável. Por fim, na Seção 4, pode-se observar a aplicação do modelo Log-logística discreta, seguido das conclusões finais.

2 MODELO DE REGRESSÃO LOG-LOGÍSTICA DISCRETA

Nesta seção se encontra descrito o modelo de regressão Log-Logística discreta que é aplicado aos dados de evasão. Inicialmente apresentamos a distribuição log-logística discreta e os resultados sobre algumas medidas de probabilidade dessa distribuição.

2.1 Distribuição Log-Logística discreta

Uma variável aleatória contínua X é Log-Logística se sua função de densidade de probabilidade (FDP) f_X, função de distribuição acumulada (FDA) $x \sim F_{\alpha,\beta}$, e função de sobrevivência (FS) S_x e a taxa de falha (TF) $h_x(x)$ são dadas, respectivamente, por:

$$f(x) = \frac{\gamma}{\alpha^\gamma} x^{\gamma-1} \left[1 + \left(\frac{x}{\alpha}\right)^\gamma\right]^{-2} \qquad x > 0 \tag{1}$$

$$F_{\alpha,\beta}(x) = 1 - \frac{1}{1 + \left(\frac{x}{\alpha}\right)^\gamma} \tag{2}$$

$$S_X(x) = \frac{1}{1 + \left(\frac{x}{\alpha}\right)^\gamma} \tag{3}$$

$$h_X(x) = \frac{\gamma(x/\alpha)^{\gamma-1}}{\alpha[1 + (x/\alpha)]^\gamma} \tag{4}$$

Em que os parâmetros $\boldsymbol{\alpha} > 0$ e $\boldsymbol{\gamma} > 0$ são de escala e forma, respectivamente.

Discretiza-se $X \sim LL_{\alpha,\beta}$ seguindo o mesmo procedimento utilizado por Nakagawa e Osaki (1975) para discretizar a distribuição Weibull. O procedimento consiste em obter uma variável discreta T com a parte inteira $[X]$ de X. Isto é, $T = [X]$ tem distribuição Log-logística discreta se sua função de probabilidade (FP), definidas por $p(t) = P(T = t) = FX(t + 1) - FX(t)$,

função de sobrevivência (FS) como sendo *S(t) = P (T > t)* e função de taxa de falha (FTF) *h(t)*, são, respectivamente, expressas por:

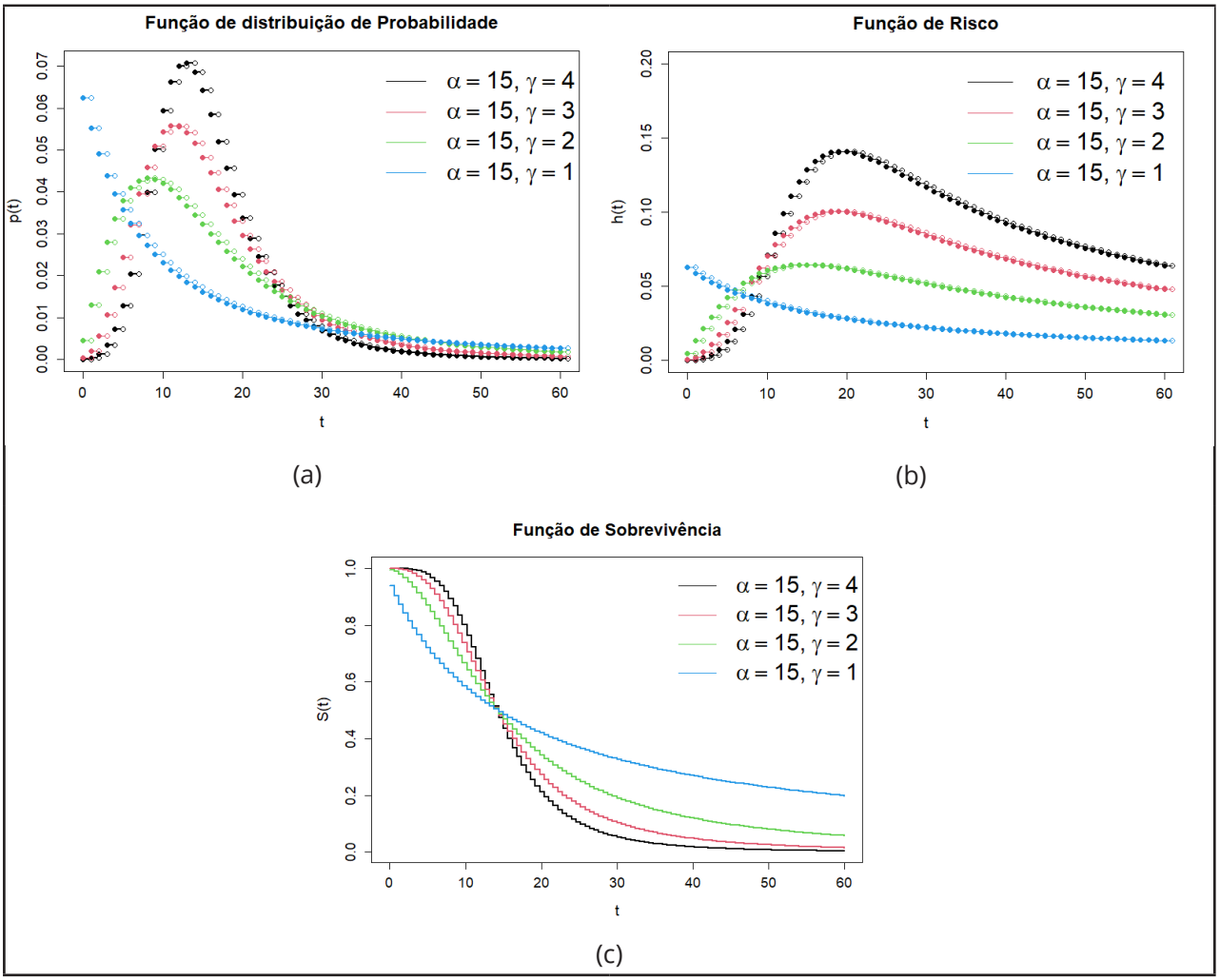
$$p(t;\alpha,\gamma) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}} \qquad t = 0,1,2,\ldots \tag{5}$$

$$S(t;\alpha,\gamma) = \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}} \qquad t = 0,1,2,\ldots \tag{6}$$

$$h(t;\alpha,\gamma) = 1 - \frac{1+\left(\frac{t}{\alpha}\right)^{\gamma}}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}} \qquad t = 0,1,2,\ldots \tag{7}$$

Sendo $\boldsymbol{\alpha} > 0$ o parâmetro de escala e $\boldsymbol{\gamma} > 0$ o parâmetro de forma.

Figura 1 – Ilustração da forma da função de distribuição de probabilidade (a), de risco (b) e de sobrevivência (c) da distribuição Log-Logística discreta utilizando alguns valores $\boldsymbol{\alpha}$ para e $\boldsymbol{\gamma}$



Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

2.1.1 Momentos

Se *T~LLD(α,Y)*, seu r-ésimo momento é dado por:

$$E(T^r) = \alpha^\gamma \sum_{k=0}^{\infty} \left(\frac{k^r}{\alpha^\gamma + k^\gamma} \right) - \alpha^\gamma \sum_{k=0}^{\infty} \left(\frac{k^r}{\alpha^\gamma + (k+1)^\gamma} \right) \quad (8)$$

Para verificar se há alguma restrição para o momento ***r*** das séries, utiliza-se a seguinte comparação. A partir do primeiro argumento da série definida em (6), tem-se que:

$$\frac{k^\gamma}{\alpha^\gamma + k^\gamma} < \frac{k^r}{k^\gamma} = k^{r-\gamma}$$

Além disso, como $\sum_{k=0}^{\infty} k^{r-\gamma}$ é uma série ***p***, com $p = Y - r$, então ela é convergente para $p > 1$. Assim, as séries definidas em (8), são convergentes para *c*. Portanto, o r-ésimo momento dado por (8) é finito para um número real *r*, em que $r \leq Y - 1$. Para $r > Y - 1.$, o r-ésimo momento não existe, o que comprova que a densidade de *T* tem cauda pesada.

Em particular, a esperança e variância de *T*, válidas apenas para $r \leq Y - 1$ e obtidas quando $r = 1$ e 2, são dados, respectivamente por:

$$E(T) = \alpha^\gamma \sum_{k=0}^{\infty} \left(\frac{k}{\alpha^\gamma + k^\gamma} \right) - \alpha^\gamma \sum_{k=0}^{\infty} \left(\frac{k}{\alpha^\gamma + (k+1)^\gamma} \right) \quad (9)$$

$$Var(T) = \alpha^\gamma \left[\sum_{k=0}^{\infty} \left(\frac{k^2}{\alpha^\gamma + k^\gamma} \right) - \sum_{k=0}^{\infty} \left(\frac{k^2}{\alpha^\gamma + (k+1)^\gamma} \right) \right] - [E(T)]^2 \quad (10)$$

2.1.2 Função quantil

A função inversa generalizada de $T \sim F_{LLD}$, dada por:

$$F_{LLD}^{[-1]} = inf\{t : m \leq F_{LLD}(t)\} \text{ para } 0 < m \leq 1 \quad (11)$$

Define a função quantil $q_m = F_{LLD}^{[-1]}(m)$, que satisfaz $F_{LLD}(T) = U \sim U[0,1]$ e

$$q(F_{LLD}(T)) = T \quad (12)$$

O resultado (12) permite utilizar a função quantil para gerar amostras de T. A função quantil da variável Log-Logística discreta, de $T \sim F_{LLD}$, é dada por:

$$q_m(\alpha,\gamma) = inf\left\{t:\alpha\left[\frac{m}{1-m}\right]^{1/\gamma} - 1 \leq t\right\} \qquad \alpha \geq 1 \tag{13}$$

Em particular, a mediana é dada por:

$$q_{0,5}(\alpha,\gamma) = \begin{cases} \inf\{t:\alpha - 1 \leq t\} & ,\alpha \geq 1 \\ 0 & ,\alpha < 1 \end{cases} \tag{14}$$

No intuito de verificar a convergência das séries definidas em (9) e (10), simularam-se amostras de tamanho *n* de $T \sim F_{LLD}$ para seis conjuntos de parâmetros, através dos seguintes passos:

Na Tabela 1 estão os resultados de *E(T)* e *Var(T)*, conforme (9) e (10), respectivamente. Comparam-se esses resultados com a respectiva média $\underline{t}$ amostral e variância amostral S^2. Por esses valores têm-se que, tanto os resultados da média quanto da variância são satisfatórios para valores de $Y = 4$ e $Y = 5$. Isto se deve ao fato de que a condição de convergência é satisfeita, pois $Y - 1 > 2$. Verificou-se que quando a condição não é satisfeita, as equações (9) e (10) não calculam a média e variância, respectivamente.

Além disso, calculou-se a estimativa da mediana de acordo com a equação (14) e nota-se que o valor da mediana está de acordo com o valor do parâmetro α menos uma unidade.

Tabela 1 – Estatísticas para simulação dos dados que seguem uma distribuição Log-Logística discreta, com *n* = 100.000

(α ; Y)	E(T)	$\underline{t}$	Var(T)	S^2	Md(T)
(35;4)	38,37523	38,34735	413,0257	412,9074	34,00010
(25;4)	27,26802	27,24884	210,7682	210,6832	24,00005
(15;5)	15,53439	15,52563	40,27561	40,19533	14,00007
(7;3)	7,96437	7,95503	46,93826	47,50118	6,00008
(3;2)	4,21237	4,20489	**197,7791**	**146,6396**	2,00011
(3;1)	**34,67820**	**55,77207**	**2.998.546**	**66.831.513**	2,00021

Fonte: Elaborado pelos autores com base nos dados da pesquisa (2025)

2.2 Modelo de regressão Log-Logística discreta

A análise de sobrevivência consiste em uma classe de técnicas e procedimentos estatísticos para estudos com dados em que a variável resposta é, geralmente, o tempo até a ocorrência de um determinado evento de interesse, denominado tempo de falha. Além disso, uma das características principais em análise de sobrevivência é a presença de informações incompletas ou parciais, intituladas como censuras. Diante de dados em que as informações não estão completas, há presença de censura, sendo assim, se faz necessária a inclusão de uma variável indicadora de censura (ou falha) definida da seguinte forma:

δ_i =0, se o i-ésimo dado foi censurado
δ_i = 1, se o i-ésimo dado não foi censurado (falhou) (15)

Nesta pesquisa considera-se a censura à direita, que é quando o tempo de interesse está à direita do tempo registrado.

Além disso, na maioria dos estudos de análise de sobrevivência é possível verificar que algumas covariáveis observadas influenciam o tempo de ocorrência do evento de interesse. O uso dessas covariáveis em um modelo de regressão é uma maneira importante de representar a heterogeneidade em uma população (Lawless, 2011).

O modelo de regressão Log-Logística discreta (MRLLD) é definido pela distribuição (5) da variável resposta $T \sim F_{LLD}$, onde o parâmetro α está relacionado a um vetor conhecido de covariáveis $x^T = (1, x_1, \ldots x_p)$, através da função de ligação:

$$\alpha = exp\ (x^T \beta) \tag{16}$$

Em que $\beta = (\beta_0, \beta_1, \ldots \beta_p)$ é o vetor dos coeficientes de regressão. Assim, ao utilizar as funções (5), (6) e (7), tem-se que a função de probabilidade, de sobrevivência e de risco do MRLLD são dadas, respectivamente, por:

$$p\left(\frac{t}{x}\right) = \frac{1}{1+\left[\frac{t}{exp\ (x^T\beta)}\right]^{\gamma}} - \frac{1}{1+\left[\frac{(t+1)}{exp\ (x^T\beta)}\right]^{\gamma}} \qquad t = 0,1,2,\cdots \tag{17}$$

$$S\left(\frac{t}{x}\right) = \frac{1}{1+\left[\frac{(t+1)}{exp\ (x^T\beta)}\right]^{\gamma}} \qquad t = 0,1,2,\cdots \tag{18}$$

$$h\left(\frac{t}{x}\right) = 1 - \frac{1+\left[\frac{t}{exp\ (x^T\beta)}\right]^{\gamma}}{1+\left[\frac{(t+1)}{exp\ (x^T\beta)}\right]^{\gamma}} \qquad t = 0,1,2,\cdots \tag{19}$$

Os parâmetros $\theta = (\beta,\gamma)$ do novo MRLLD são estimados utilizando o método da máxima verossimilhança. Para os valores amostrais ($t_1, \ldots t_n$) de $T \sim F_{MRLLD}$, o logaritmo da função de verossimilhança, definida a partir de (17) e (18), é dada por:

$$log\ (L(\theta)) = \sum_{i=1}^{n}\left\{\delta_i log\left[\frac{1}{1+\left[\frac{t_i}{exp\ (x_i{}^T\beta)}\right]^{\gamma}} - \frac{1}{1+\left[\frac{(t_i+1)}{exp\ (x_i{}^T\beta)}\right]^{\gamma}}\right]\right\} + (1-\delta_i)\, log\left[\frac{1}{1+\left[\frac{(t_i+1)}{exp\ (x_i{}^T\beta)}\right]^{\gamma}}\right] + C \tag{20}$$

Em C é uma constante que não depende de θ.

Os estimadores de máxima verossimilhança (EMV) são os valores de θ que maximizam $L(\theta)$ ou, equivalentemente o logaritmo de $L(\theta)$. Eles são encontrados resolvendo-se o sistema de equações:

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} = 0 \tag{21}$$

A solução deste sistema de equações para um conjunto de dados particular deve ser obtida por meio de um método numérico, o que, usualmente, utiliza-se o método de Newton-Raphson e a utilização de um pacote estatístico. Neste trabalho, o *software* R (TEAM, 2025) foi utilizado para obter $\hat{\theta}$ por meio da função *optim.*

2.3 Intervalo de confiança para os parâmetros

Após a estimação dos parâmetros, é importante a construção dos seus intervalos de confiança. O intervalo de confiança é construído a partir da matriz de informação de Fisher definida por:

$$I_f(\theta) = E\left[\left(\frac{\partial l(t,\delta|\theta)}{\partial \theta}\right)^2\right] = -E\left[\left(\frac{\partial^2 l(t,\delta|\theta)}{\partial \theta^2}\right)\right]$$

Em que *l(θ) = logL (θ).*

Utiliza-se, então, a distribuição assintótica do estimador de máxima verossimilhança. Para grandes amostras, sob condições de regularidade, essa propriedade estabelece que a distribuição de $\hat{\theta}$ converge assintoticamente para uma distribuição normal multivariada de média θ e matriz de variância e covariância Var($\hat{\theta}$) ou seja,

$$\hat{\theta} \sim^{a} N_k\left(\theta, Var\left(\hat{\theta}\right)\right)$$

Em que *k* é a dimensão de *θ*. Além disso, tem-se que:

$$Var\left(\hat{\theta}\right) \approx \left[I_f\left(\hat{\theta}\right)\right]^{-1}$$

Em que $I_f(\hat{\theta})$ é a informação de Fisher observada da amostra.

Desta forma, um intervalo aproximado de (1 – *α*)100% de confiança para *θ* é dado por:

$$\hat{\theta} \pm Z_{1-\alpha/2}\sqrt{\widehat{Var(\hat{\theta})}} \qquad (22)$$

Em que $Z_{1-\alpha/2}$ é o quantil (*1 – α/2*)de uma distribuição normal padrão.

Em alguns casos, torna-se necessário estimar uma função dos parâmetros e uma das propriedades dos estimadores de máxima verossimilhança é a propriedade de invariância, ou seja, se $\hat{\theta}$ é um EMV de *θ* e *g(θ)* é uma função bijetora, então *g(θ̂)* é um EMV de *g(θ)*.

Para a construção de intervalo de confiança para *t = g(θ)*, é necessário obter uma estimativa para o erro-padrão de $\hat{\tau} = g(\hat{\theta})$ e isso é feito utilizando o método delta. Além disso, para grandes amostras, tem-se que:

$$g(\hat{\theta}) \sim^{a} N_k\left(g(\theta), Var\left(\hat{\theta}\right)[g'(\theta)]^2\right) \quad (23)$$

Ou seja, $\hat{\tau} = g(\hat{\theta})$ converge assintoticamente para uma distribuição normal multivariada com média *t = g(θ)* e matriz de variância e covariância $Var\left(\hat{\theta}\right)[g'(\theta)]^2$, em que *k* é a dimensão de $g(\hat{\theta})$ e *g' (θ)* é derivada de primeira ordem de *g(θ)*.

2.4 Interpretação dos parâmetros de regressão

A interpretação dos coeficientes de regressão é definida a partir da função quantil do modelo Log-Logística discreta. Uma proposta de interpretação é a de se fazer uso da razão de tempos medianos (Hosmer; Lemeshow, 1999). Ao considerar o modelo LLD, tem-se a seguinte relação:

$$t_{0,5}(\hat{\alpha},\hat{\gamma}) = inf\left\{t;\hat{\alpha}\left[\frac{0,5}{(1-0,5)}\right]^{\frac{1}{\hat{\gamma}}} - 1 \leq t\right\} \cong \hat{\alpha} - 1$$

Considerando $\hat{\alpha} = exp\ (\hat{\beta}_0 + \hat{\beta}_1 x)$, tem-se que $t_{0,5} + 1 = exp\ (\hat{\beta}_0 + \hat{\beta}_1 x)$. Sendo assim, a razão de tempo mediano de um modelo de regressão Log-Logística discreta composto por uma covariável dicotômica é dada pela seguinte expressão:

$$\frac{1 + t_{0,5}(x = 1,\hat{\gamma},\hat{\beta})}{1 + t_{0,5}(x = 0,\hat{\gamma},\hat{\beta})} = \frac{exp\ (\hat{\beta}_0 + \hat{\beta}_1 x)}{exp\ (\hat{\beta}_0)} = e^{\hat{\beta}_1}$$

Logo, se $\hat{\beta}_1$ for positivo, interpreta-se que o tempo mediano mais uma unidade ($t_{0,5}$ + 1) do indivíduo pertencente ao grupo *x* = 1 é $e^{\hat{\beta}_1}$ vezes o tempo mediano mais uma unidade de um indivíduo do grupo x = 0. No entanto, caso $\hat{\beta}_1$ for negativo, conclui-se que o tempo ($t_{0,5}$ + 1) de um indivíduo do grupo x = 0 é $e^{-\hat{\beta}_1}$ ou $\frac{1}{e^{\hat{\beta}_1}}$ vezes o tempo ($t_{0,5}$ + 1) de um indivíduo do grupo *x* = 1.

3 ESTUDO DE SIMULAÇÕES

3.1 Simulação de LLD

O comportamento das EMV para os parâmetros da distribuição log-logística discreta, definido na seção 2.1, é estudado via simulação de Monte Carlo, realizada por meio do *software* R. Para a simulação de valores do tempo de sobrevivência *T* foi utilizado o método da inversão e para as censuras considerou-se uma variável δ_i com distribuição Binomial; $\delta_i \sim Bin(n,p)$, em que *n* é o tamanho da amostra e *p* é a proporção de censura desejada, conforme algoritmo à seguir:

1. Gerar $U \sim U(n,0,1)$
2. Retorne $T=F^{[-1]}(U)$
3. Gerar $\delta_i \sim Bin(n,p)$

Além disso, destaca-se que o mecanismo de censura utilizado foi o de censura à direita e considerada como independente do tempo de sobrevivência. Com diversas amostras buscou-se pesquisar a influência do percentual de censuras nas EMV. Desta forma, utilizou-se de quatro percentuais específicos de censura, sendo esses 0%, 10%, 20% e 30%.

Sendo a taxa de falha do modelo LLD decrescente e unimodal, a escolha do vetor de parâmetros contempla esses comportamentos. Na Tabela 2 estão os vetores de parâmetros que definem dois cenários. No Cenário 1 os valores do parâmetro de escala, α, é alto e no Cenário 2 α é pequeno.

Para examinar a precisão do EMV, as estimativas médias de θ denotadas por $\hat{\theta}$, o viés de θ denotado por EQM(θ) e o erro quadrático médio de denotado por foram calculados com 2.000 réplicas de Monte Carlo para amostras de tamanho *n* = *50*, 100, 200 e 500. Os resultados dessas estatísticas estão nas Tabelas 3 e 4.

Na Tabela 3 foram estudados quatro vetores de parâmetros. A relação entre o parâmetro de escala α e de forma Y fica evidente nas estimativas médias, viés e EQM. Para os dois primeiros vetores de parâmetros, α próximo de Y, os resultados mostram que o EMV tem um bom comportamento, pois o EQM decresce à medida que o tamanho da

amostra aumenta. Além disso, as estimativas sofrem influência do percentual de censuras, ou seja, à medida em que diminui o percentual de censura, o EQM também decresce.

Tabela 2 – Cenários utilizados na simulação da distribuição LLD

Cenários	α	Y	Função de risco
1	7,4	7	Unimodal
	7,4	5,2	Unimodal
	7,4	3,2	Unimodal
	7,4	1	Decrescente
2	0,9	7	Unimodal
	0,9	5,2	Unimodal
	0,9	3,2	Unimodal
	0,9	1	Decrescente

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Para os vetores θ_3 e θ_4, conforme a diferença entre os valores dos parâmetros α e Y aumenta, ocorre um aumento significativo relacionado ao $EQM(\hat{\alpha})$ e um decréscimo no $EQM(\hat{\gamma})$.

No θ_4 percebe-se que, com uma amostra de tamanho 50 e com percentual de censura 30%, tem-se que o $EQM(\hat{\alpha})$ é alto. No entanto, ocorre uma queda significativa quando o tamanho da amostra aumenta para n=500 e percentual de censura igual a 0% o $EQM(\hat{\alpha})$ é pequeno. Observa-se, também que, quando há algum percentual de censura, o vício do estimador $\hat{\alpha}$ em todos os casos é positivo.

De maneira geral, percebe-se que as estimativas dos parâmetros sofrem influência principalmente no que se refere à distância entre os valores de α e Y. Quanto mais próximos forem os valores de α e Y, menores serão os EQM's. Além disso, valores altos dos parâmetros, mesmo sendo próximos, contribuem para o aumento do EQM. Na maioria dos casos, há a influência do percentual de censura sobre as estimativas. No entanto, em casos em que há uma diferença significativa entre os valores dos parâmetros, essa relação não se satisfaz, como pode ser visto no Cenário 2 com α = 0,9 e Y = 7, em que os menores EQM's ocorreram com percentual de censura igual a 10% e com amostra de tamanho 500.

Tabela 3 – Resultados de $\hat{\alpha}$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 1**
(Continua)

N		θ_1(α=7,4; Y =7)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	8,1709	6,1589	7,8660	6,4912	7,6161	6,8344	7,4029	7,1726
	b($\hat{\theta}$)	0,7709	-0,8411	0,4660	-0,5088	0,2161	-0,1656	0,0029	0,1726
	EQM($\hat{\theta}$)	0,7361	1,4841	0,3245	1,0308	0,1336	0,7878	0,0663	0,7975
100	$\hat{\theta}$	8,1615	6,0342	7,8578	6,3754	7,6081	6,7284	7,4043	7,0863
	b($\hat{\theta}$)	0,7615	-0,9658	0,4578	-0,6246	0,2081	-0,2716	0,0043	0,0863
	EQM($\hat{\theta}$)	0,6481	1,3186	0,2603	0,7793	0,0831	0,4598	0,0342	0,3756
200	$\hat{\theta}$	8,1548	5,9633	7,8548	6,3088	7,6070	6,6714	7,4021	7,0467
	b($\hat{\theta}$)	0,7548	-1,0367	0,4548	-0,6912	0,2070	-0,3286	0,0021	0,0467
	EQM($\hat{\theta}$)	0,6037	1,2509	0,2320	0,6541	0,0627	0,2888	0,0167	0,1952
500	$\hat{\theta}$	8,1489	5,9561	7,8500	6,3030	7,6036	6,6585	7,3996	7,0107
	b($\hat{\theta}$)	0,7489	-1,0439	0,4500	-0,6970	0,2036	-0,3415	-0,0004	0,0107
	EQM($\hat{\theta}$)	0,5740	1,1588	0,2125	0,5541	0,0495	0,1859	0,0067	0,0740
N		θ_2(α=7,4; Y =5,2)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	8,4078	4,6422	8,0069	4,8678	7,6812	5,1002	7,4033	5,3273
	b($\hat{\theta}$)	1,0078	-0,5578	0,6069	-0,3322	0,2812	-0,0998	0,0033	0,1273
	EQM($\hat{\theta}$)	1,2814	0,7491	0,5670	0,5323	0,2374	0,4213	0,1194	0,4276
100	$\hat{\theta}$	8,3926	4,5529	7,9944	4,7882	7,6696	5,0248	7,4064	5,2593
	b($\hat{\theta}$)	0,9926	-0,6471	0,5944	-0,4118	0,2696	-0,1752	0,0064	0,0593
	EQM($\hat{\theta}$)	1,1138	0,6347	0,4481	0,3868	0,1469	0,2442	0,0629	0,2023
200	$\hat{\theta}$	8,3829	4,5035	7,9897	4,7413	7,6678	4,9855	7,4036	5,2314
	b($\hat{\theta}$)	0,9829	-0,6965	0,5897	-0,4587	0,2678	-0,2145	0,0036	0,0314
	EQM($\hat{\theta}$)	1,0287	0,5834	0,3937	0,3083	0,1075	0,1452	0,0306	0,1038
500	$\hat{\theta}$	8,3748	4,4955	7,9832	4,7344	7,6631	4,9757	7,3998	5,2077
	b($\hat{\theta}$)	0,9748	-0,7045	0,5832	-0,4656	0,2631	-0,2243	-0,0002	0,0077
	EQM($\hat{\theta}$)	0,9743	0,5357	0,3580	0,2555	0,0837	0,0889	0,0119	0,0395
N		θ_3(α =7,4; Y =3,2)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	8,8839	3,1662	8,2859	3,3062	7,8097	3,4470	7,4123	3,5857
	b($\hat{\theta}$)	1,4839	-0,3338	0,8859	-0,1938	0,4097	-0,0530	0,0123	0,0857
	EQM($\hat{\theta}$)	2,8391	0,3132	1,2428	0,2311	0,5218	0,1879	0,2624	0,1931
100	$\hat{\theta}$	8,8579	3,1071	8,2658	3,2533	7,7914	3,3992	7,4117	3,5397
	b($\hat{\theta}$)	1,4579	-0,3929	0,8658	-0,2467	0,3914	-0,1008	0,0117	0,0397
	EQM($\hat{\theta}$)	2,4300	0,2558	0,9657	0,1609	0,3176	0,1073	0,1367	0,0909
200	$\hat{\theta}$	8,8395	3,0724	8,2557	3,2213	7,7866	3,3724	7,4065	3,5223
	b($\hat{\theta}$)	1,4395	-0,4276	0,8557	-0,2787	0,3866	-0,1276	0,0065	0,0223
	EQM($\hat{\theta}$)	2,2216	0,2292	0,8378	0,1232	0,2300	0,0618	0,0660	0,0474
500	$\hat{\theta}$	8,8261	3,0662	8,2453	3,2157	7,7788	3,3641	7,4004	3,5052
	b($\hat{\theta}$)	1,4261	-0,4338	0,8453	-0,2843	0,3788	-0,1359	0,0004	0,0052
	EQM($\hat{\theta}$)	2,0911	0,2063	0,7559	0,0986	0,1757	0,0358	0,0259	0,0179

Tabela 3 – Resultados de $\hat{\alpha}$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 1**
(Conclusão)

N		θ_4(α =7,4; Y =1)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	14,1334	0,9114	11,1217	0,9488	9,0890	0,9864	7,5943	1,0221
	b($\hat{\theta}$)	6,7334	-0,0886	3,7217	-0,0512	1,6890	-0,0136	0,1943	0,0221
	EQM($\hat{\theta}$)	66,3408	0,0270	24,3114	0,0206	8,9046	0,0173	3,5264	0,0170
100	$\hat{\theta}$	13,7312	0,8947	10,8590	0,9346	8,8934	0,9733	7,5206	1,0099
	b($\hat{\theta}$)	6,3312	-0,1053	3,4590	-0,0654	1,4934	-0,0267	0,1206	0,0099
	EQM($\hat{\theta}$)	49,0883	0,0208	16,5671	0,0136	4,8981	0,0096	1,7821	0,0083
200	$\hat{\theta}$	13,5200	0,8844	10,7413	0,9249	8,8231	0,9658	7,4611	1,0058
	b($\hat{\theta}$)	6,1200	-0,1156	3,3413	-0,0751	1,4231	-0,0342	0,0611	0,0058
	EQM($\hat{\theta}$)	41,6915	0,0179	13,3293	0,0100	3,2710	0,0054	0,8303	0,0043
500	$\hat{\theta}$	13,3813	0,8830	10,6490	0,9238	8,7639	0,9637	7,4165	1,0013
	b($\hat{\theta}$)	5,9813	-0,1170	3,2490	-0,0762	1,3639	-0,0363	0,0165	0,0013
	EQM($\hat{\theta}$)	37,3586	0,0155	11,3924	0,0075	2,3565	0,0029	0,3209	0,0016

Fonte: Elaborado pelos autores conforme dados da pesquisa (2023)

Na Tabela 4, estão os resultados das estatísticas correspondentes ao Cenário 2. Neste cenário foi considerado um valor pequeno de α, com α = 0,9.

Assim como no cenário anterior, conforme diminui a distância entre os parâmetros, ocorre uma maior acurácia nos estimadores. Desta forma, utilizando α = 0,9 e Y = 1, percebe-se que o viés e EQM para ambos os parâmetros são pequenos para *n* grande e à medida que o percentual de censura diminui combinado com um tamanho maior de amostra os resultados são melhores.

Tabela 4 – Resultados de $\hat{\alpha}$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 2**
(Continua)

N		θ_1(α =0,9; Y =7)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	1,0454	2,7457	0,9485	3,7591	0,9023	6,7457	0,9475	19,0572
	b($\hat{\theta}$)	0,1454	-4,2543	0,0485	-3,2409	0,0023	-0,2543	0,0475	12,0572
	EQM($\hat{\theta}$)	0,0367	20,0401	0,0128	22,5745	0,0072	46,6584	0,0048	**186,7429**
100	$\hat{\theta}$	1,0447	2,5956	0,9467	3,1256	0,8939	4,6216	0,9382	16,8482
	b($\hat{\theta}$)	0,1447	-4,4044	0,0467	-3,8744	-0,0061	-2,3784	0,0382	9,8482
	EQM($\hat{\theta}$)	0,0279	19,6606	0,0067	15,6978	0,0033	17,4951	0,0038	**155,7016**
200	$\hat{\theta}$	1,0425	2,5380	0,9458	3,0214	0,8906	3,9245	0,9256	13,6306
	b($\hat{\theta}$)	0,1425	-4,4620	0,0458	-3,9786	-0,0094	-3,0755	0,0256	6,6306
	EQM($\hat{\theta}$)	0,0237	20,0284	0,0043	15,9981	0,0016	10,0166	0,0025	**104,9254**

Tabela 4 – Resultados de $\hat{\alpha}$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 2**
(Continua)

N		θ_1(α =0,9; Y =7)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
500	$\hat{\theta}$	1,0415	2,5373	0,9457	3,0227	0,8906	3,8823	0,9094	9,4137
	b($\hat{\theta}$)	0,1415	-4,4627	0,0457	-3,9773	-0,0094	-3,1177	0,0094	2,4137
	EQM($\hat{\theta}$)	0,0213	19,9636	0,0030	15,8869	**0,0007**	**9,8448**	0,0010	**37,2721**
N		θ_2(α =0,9; Y =5,2)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	1,1001	2,6439	0,9930	3,2840	0,9301	4,9439	0,9258	12,3952
	b($\hat{\theta}$)	0,2001	-2,5561	0,0930	-1,9160	0,0301	-0,2561	0,0258	7,1952
	EQM($\hat{\theta}$)	0,0563	7,2490	0,0190	8,1531	0,0076	21,5745	0,0051	**128,8790**
100	$\hat{\theta}$	1,0991	2,5508	0,9919	3,0037	0,9258	3,8482	0,9134	8,6768
	b($\hat{\theta}$)	0,1991	-2,6492	0,0919	-2,1963	0,0258	-1,3518	0,0134	3,4768
	EQM($\hat{\theta}$)	0,0471	7,2245	0,0131	5,3025	0,0037	3,9932	0,0026	**61,5490**
200	$\hat{\theta}$	1,0964	2,5020	0,9912	2,9297	0,9244	3,6476	0,9045	6,0664
	b($\hat{\theta}$)	0,1964	-2,6980	0,0912	-2,2703	0,0244	-1,5524	0,0045	0,8664
	EQM($\hat{\theta}$)	0,0422	7,3742	0,0106	5,2810	0,0021	2,6291	0,0011	**12,5291**
500	$\hat{\theta}$	1,0946	2,5001	0,9902	2,9284	0,9239	3,6256	0,9010	5,3167
	b($\hat{\theta}$)	0,1946	-2,6999	0,0902	-2,2716	0,0239	-1,5744	0,0010	0,1167
	EQM($\hat{\theta}$)	0,0393	7,3286	0,0091	5,2117	0,0012	2,5609	0,0004	0,3323
N		θ_3(α =0,9; Y =3,2)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	1,1986	2,2199	1,0614	2,4942	0,9664	2,9075	0,9032	3,8419
	b($\hat{\theta}$)	0,2986	-0,9801	0,1614	-0,7058	0,0664	-0,2925	0,0032	0,6419
	EQM($\hat{\theta}$)	0,1146	1,2339	0,0427	0,8285	0,0158	1,4968	0,0078	**10,1046**
100	$\hat{\theta}$	1,1959	2,1599	1,0603	2,4276	0,9653	2,7908	0,9026	3,3303
	b($\hat{\theta}$)	0,2959	-1,0401	0,1603	-0,7724	0,0653	-0,4092	0,0026	0,1303
	EQM($\hat{\theta}$)	0,0995	1,1985	0,0333	0,7340	0,0095	0,3544	0,0038	0,7584
200	$\hat{\theta}$	1,1920	2,1258	1,0586	2,3842	0,9641	2,7392	0,9016	3,2602
	b($\hat{\theta}$)	0,2920	-1,0742	0,1586	-0,8158	0,0641	-0,4608	0,0016	0,0602
	EQM($\hat{\theta}$)	0,0911	1,2122	0,0288	0,7323	0,0066	0,2998	0,0018	0,1339
500	$\hat{\theta}$	1,1901	2,1281	1,0578	2,3890	0,9638	2,7331	0,9003	3,2203
	b($\hat{\theta}$)	0,2901	-1,0719	0,1578	-0,8110	0,0638	-0,4669	0,0003	0,0203
	EQM($\hat{\theta}$)	0,0863	1,1718	0,0263	0,6841	0,0051	0,2507	**0,0007**	**0,0497**
N		θ_4(α =0,9; Y =1)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
50	$\hat{\theta}$	1,9625	0,8823	1,4740	0,9310	1,1539	0,9828	0,9271	1,0381
	b($\hat{\theta}$)	1,0625	-0,1177	0,5740	-0,0690	0,2539	-0,0172	0,0271	0,0381
	EQM($\hat{\theta}$)	1,5596	0,0511	0,5305	0,0417	0,1776	0,0366	0,0670	0,0374
100	$\hat{\theta}$	1,9118	0,8619	1,4433	0,9123	1,1316	0,9654	0,9166	1,0187
	b($\hat{\theta}$)	1,0118	-0,1381	0,5433	-0,0877	0,2316	-0,0346	0,0166	0,0187
	EQM($\hat{\theta}$)	1,2108	0,0364	0,3873	0,0249	0,1070	0,0188	0,0342	0,0181

Tabela 4 – Resultados de $\hat{\alpha}$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 2**
(Conclusão)

N		θ_4(α =0,9; Y =1)							
		30%		20%		10%		0%	
		α	Y	α	Y	α	Y	α	Y
200	$\hat{\theta}$	1,8815	0,8483	1,4272	0,8991	1,1223	0,9535	0,9092	1,0105
	b($\hat{\theta}$)	0,9815	-0,1517	0,5272	-0,1009	0,2223	-0,0465	0,0092	0,0105
	EQM($\hat{\theta}$)	1,0491	0,0316	0,3198	0,0187	0,0734	0,0107	0,0164	0,0087
500	$\hat{\theta}$	1,8629	0,8459	1,4163	0,8974	1,1145	0,9500	0,9024	1,0031
	b($\hat{\theta}$)	0,9629	-0,1541	0,5163	-0,1026	0,2145	-0,0500	0,0024	0,0031
	EQM($\hat{\theta}$)	0,9598	0,0270	0,2833	0,0137	0,0558	0,0056	**0,0064**	**0,0033**

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

3.2 Simulação de MRLLD

O comportamento das EMV para os parâmetros do modelo de regressão log-logística discreta, definido na seção 2.2, é estudado via simulação de Monte Carlo, realizada por meio do *software* R. Para a simulação de valores do tempo de sobrevivência *T* foi utilizado o método da inversão e para as censuras considerou-se uma variável δ_i com distribuição Bernoulli; δ_i*~Bin(n,p)*, em que *n* é o tamanho da amostra e *p* é a proporção de censura desejada e uma variável *X~Bin(n,p = 0.4)*, sendo *p* estabelecido em 0,4.

O mesmo mecanismo de censura e tamanho de amostra foi utilizado nessa seção, assim como em 3.1.

Tabela 5 – Cenários utilizados na simulação da distribuição LLD

Cenários	β_0	β_1	*Y*	Função de risco
1	1,1	0,5	7	Unimodal
	1,1	0,5	1	Decrescente
2	1,5	-2,0	7	Unimodal
	1,5	-2,0	1	Decrescente

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Os resultados das simulações para o Cenário 1 são apresentados na Tabela 6. Nota-se que, em geral, os estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\gamma}$ apresentam vícios reduzidos e erros quadráticos médios (EQM) que decrescem conforme o tamanho da amostra aumenta, assim como nas simulações da distribuição LLD, na Seção 3.1.

Além disso, para Y = 7, nota-se uma leve superestimação do parâmetro em amostras pequenas, seguido de EQMs elevados, especialmente sob 30% de censura. Por outro lado, à medida que a censura diminui e o tamanho da amostra aumenta, os EQMs tornam-se menores, indicando maior precisão dos estimadores.

No caso de Y = 1, os EQM de $\hat{\beta}_0$ e $\hat{\beta}_1$ são mais elevados sob censura de 30%, mas reduzem-se significativamente com *N* = 500. Destaca-se, portanto, que esse comportamento evidencia, mais uma vez, a consistência dos estimadores.

Tabela 6 – Resultados de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\gamma}$, e , e seus respectivos EQM e viés para o **cenário 1**

N		$\theta_1(\beta_0=1,1;\ \beta_1=0,5;\ Y=7)$											
		30%			20%			10%			0%		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$
50	$\hat{\theta}$	1,137	0,502	7,304	1,121	0,502	7,325	1,109	0,501	7,316	1,102	0,497	7,351
	b($\hat{\theta}$)	0,037	0,002	0,304	0,021	0,002	0,325	0,009	0,001	0,316	0,002	-0,003	0,351
	EQM($\hat{\theta}$)	0,005	0,005	1,911	0,004	0,004	1,589	0,003	0,003	1,357	0,003	**0,003**	**1,143**
100	$\hat{\theta}$	1,135	0,500	7,051	1,120	0,501	7,111	1,109	0,501	7,128	1,100	0,500	7,159
	b($\hat{\theta}$)	0,035	0,000	0,051	0,020	0,001	0,111	0,009	0,001	0,128	0,000	0,000	0,159
	EQM($\hat{\theta}$)	0,003	0,005	0,732	0,002	0,004	0,648	0,001	0,003	0,566	0,001	**0,003**	**0,523**
200	$\hat{\theta}$	1,135	0,499	6,962	1,120	0,499	7,031	1,108	0,500	7,069	1,100	0,502	7,069
	b($\hat{\theta}$)	0,035	-0,001	-0,038	0,020	-0,001	0,031	0,008	0,000	0,069	0,000	0,002	0,062
	EQM($\hat{\theta}$)	0,002	0,002	0,327	0,001	0,002	0,288	0,001	0,002	0,263	0,001	**0,001**	**0,225**
500	$\hat{\theta}$	1,135	0,499	6,858	1,120	0,499	6,938	1,108	0,500	6,988	1,100	0,500	7,035
	b($\hat{\theta}$)	0,035	-0,001	-0,142	0,020	-0,001	-0,062	0,008	0,000	-0,012	0,000	0,000	0,035
	EQM($\hat{\theta}$)	0,002	0,001	0,141	0,001	0,001	0,111	**0,000**	**0,001**	**0,096**	0,000	**0,001**	**0,085**
N		$\theta_2(\beta_0=1,1;\ \beta_1=0,5;\ Y=1)$											
		30%			20%			10%			0%		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$
50	$\hat{\theta}$	1,356	0,507	1,059	1,249	0,508	1,052	1,170	0,504	1,044	1,108	0,485	1,040
	b($\hat{\theta}$)	0,256	0,007	0,059	0,149	0,008	0,052	0,070	0,004	0,044	0,008	-0,015	0,040
	EQM($\hat{\theta}$)	0,226	0,226	0,038	0,162	0,162	0,031	0,128	0,128	0,026	0,111	**0,111**	**0,021**
100	$\hat{\theta}$	1,350	0,496	1,027	1,247	0,500	1,026	1,168	0,501	1,021	1,099	0,503	1,019
	b($\hat{\theta}$)	0,250	-0,004	0,027	0,147	0,000	0,026	0,068	0,001	0,021	-0,001	0,003	0,019
	EQM($\hat{\theta}$)	0,139	0,190	0,016	0,087	0,160	0,014	0,062	0,144	0,012	0,051	**0,128**	**0,010**
200	$\hat{\theta}$	1,343	0,487	1,020	1,243	0,492	1,020	1,162	0,497	1,017	1,094	0,515	1,017
	b($\hat{\theta}$)	0,243	-0,013	0,020	0,143	-0,008	0,020	0,062	-0,003	0,017	-0,006	0,015	0,008
	EQM($\hat{\theta}$)	0,098	0,097	0,008	0,054	0,085	0,007	0,033	0,073	0,006	0,026	**0,066**	**0,005**
500	$\hat{\theta}$	1,345	0,487	1,004	1,241	0,494	1,005	1,162	0,496	1,004	1,098	0,503	1,005
	b($\hat{\theta}$)	0,245	-0,013	0,004	0,141	-0,006	0,005	0,062	-0,004	0,004	-0,002	0,003	0,005
	EQM($\hat{\theta}$)	0,075	0,039	0,003	0,033	0,033	0,002	0,015	0,030	0,002	0,010	0,026	0,002

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

A Tabela 7 exibe os resultados para o Cenário 2, com o caso particular β_1 negativo e duas variações parâmetro de forma (Y), alterando a função de risco para unimodal ou decrescente. Nota-se um padrão consistente com as simulações anteriores: tanto o viés quanto os EQMs dos estimadores diminuem com o aumento do tamanho amostral e a redução do percentual de censura.

Para Y = 7, os EQM de $\hat{\gamma}$ são altos inicialmente, mas caem para valores inferiores a 0,2 com N = 500, mesmo com 30% de censura. Por outro lado, para Y = 1, os EQMs são baixos, mesmo com amostras menores e com percentuais de censuras altos, o que demonstra a robustez das estimativas dos parâmetros do modelo de regressão.

Tabela 7 – Resultados de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\gamma}$, e seus respectivos EQM e viés para o **cenário 2**

		$\theta_1(\beta_0$=1,5; β_1=-2,00; Y =7)											
N		**30%**			**20%**			**10%**			**0%**		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$
50	$\hat{\theta}$	1,536	-3,173	7,387	1,521	-3,157	7,357	1,510	-3,085	7,351	1,502	-3,031	7,394
	b($\hat{\theta}$)	0,036	-1,173	0,387	0,021	-1,157	0,357	0,010	-1,085	0,351	0,002	-1,031	0,394
	EQM($\hat{\theta}$)	0,005	0,005	4,755	0,004	0,004	2,385	0,003	0,003	2,034	0,002	0,002	1,757
100	$\hat{\theta}$	1,535	-2,774	7,091	1,520	-2,703	7,138	1,509	-2,618	7,150	1,500	-2,601	7,174
	b($\hat{\theta}$)	0,035	-0,774	0,091	0,020	-0,703	0,138	0,009	-0,618	0,150	0,000	-0,601	0,174
	EQM($\hat{\theta}$)	0,003	1,715	1,094	0,002	1,556	0,958	0,001	1,346	0,830	0,001	1,555	0,759
200	$\hat{\theta}$	1,533	-2,344	7,002	1,519	-2,286	7,060	1,508	-2,242	7,093	1,499	-2,172	7,093
	b($\hat{\theta}$)	0,033	-0,344	0,002	0,019	-0,286	0,060	0,008	-0,242	0,093	-0,001	-0,172	0,072
	EQM($\hat{\theta}$)	0,002	0,779	0,505	0,001	0,624	0,454	0,001	0,506	0,415	0,001	0,353	0,339
500	$\hat{\theta}$	1,534	-2,026	6,889	1,519	-2,017	6,961	1,508	-2,019	7,006	1,500	-2,019	7,033
	b($\hat{\theta}$)	0,034	-0,026	-0,111	0,019	-0,017	-0,039	0,008	-0,019	0,006	0,000	-0,019	0,033
	EQM($\hat{\theta}$)	0,001	0,100	0,193	0,001	0,043	0,165	**0,000**	**0,027**	**0,144**	0,000	0,017	0,128
		$\theta_2(\beta_0$=1,5; β_1=-2,00; Y =1)											
N		**30%**			**20%**			**10%**			**0%**		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}$
50	$\hat{\theta}$	1,748	-1,942	1,068	1,646	-1,980	1,059	1,570	-2,009	1,051	1,509	-2,041	1,044
	b($\hat{\theta}$)	0,248	0,058	0,068	0,146	0,020	0,059	0,070	-0,009	0,051	0,009	-0,041	0,044
	EQM($\hat{\theta}$)	0,220	0,220	0,050	0,159	0,159	0,042	0,125	0,125	0,035	0,109	0,109	0,028
100	$\hat{\theta}$	1,740	-1,903	1,031	1,641	-1,948	1,028	1,566	-1,984	1,022	1,498	-2,007	1,021
	b($\hat{\theta}$)	0,240	0,097	0,031	0,141	0,052	0,028	0,066	0,016	0,022	-0,002	-0,007	0,021
	EQM($\hat{\theta}$)	0,133	0,255	0,020	0,084	0,207	0,017	0,060	0,181	0,015	0,051	0,162	0,014
200	$\hat{\theta}$	1,732	-1,898	1,023	1,636	-1,943	1,022	1,559	-1,977	1,019	1,494	-1,988	1,019
	b($\hat{\theta}$)	0,232	0,102	0,023	0,136	0,057	0,022	0,059	0,023	0,019	-0,006	0,012	0,009
	EQM($\hat{\theta}$)	0,092	0,127	0,010	0,052	0,108	0,009	0,032	0,091	0,008	0,026	0,079	0,006
500	$\hat{\theta}$	1,734	-1,897	1,007	1,635	-1,942	1,007	1,559	-1,978	1,005	1,498	-2,001	1,004
	b($\hat{\theta}$)	0,234	0,103	0,007	0,135	0,058	0,007	0,059	0,022	0,005	-0,002	-0,001	0,004
	EQM($\hat{\theta}$)	0,070	0,058	0,003	0,031	0,044	0,003	0,015	0,037	0,003	0,010	0,032	0,002

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

De modo geral, os resultados das simulações indicam que os EMV para os parâmetros do MRLLD são acurados, com desempenho melhorando conforme o tamanho da amostra aumenta e a taxa de censura diminui. Além disso, a forma da função de risco – unimodal ou decrescente – não compromete a qualidade das estimativas, desde que o tamanho amostral seja suficientemente grande.

4 APLICAÇÃO

Para ilustrar a aplicabilidade do MRLLD definido em (2.2) foram utilizados dados sobre o tempo de permanência de alunos em um curso superior. Inicia-se com a descrição dos dados e em seguida é feita uma análise descritiva desses dados. Foram comparadas as modelagens não paramétricas com a distribuição LLD e Weibull discreta de Nakagawa e Osaki (1975). Como os resultados foram melhores para a distribuição LLD, os dados foram analisados via o MRLLD. Todas as análises foram realizadas por meio do *software* R (TEAM, 2025).

4.1 Banco de dados

Para a aplicação foi utilizado o banco de dados dos alunos do curso de Ciência da Computação da Universidade Estadual da Paraíba - Campus I, com o total de 709 alunos. O período de acompanhamento é compreendido entre o primeiro semestre de 2010 e o segundo semestre de 2016.

O evento de interesse é a evasão do aluno, logo a variável resposta T é o tempo, em semestres, desde a matrícula até a ocorrência do evento de interesse ou uma censura. Nesta pesquisa, a censura ocorre quando o aluno ainda está cursando ou se formou e o evento de interesse (falha) é a evasão do curso. Nota-se também que $T = 0$ indica que o aluno está cursando o primeiro semestre do curso, sendo esse, compreendido ao último período de observação dos nossos dados. Além da variável T, tem-se a presença de outras informações dos alunos, denotadas como covariáveis. Neste banco de dados, a proporção de censuras é de 53,4%. Na Tabela 8, tem-se a descrição das covariáveis que foram analisadas.

Tabela 8 – Descrição das covariáveis dos alunos do curso de Computação

Covariáveis	Categorias	N (%)
Sexo	0 – Masculino	599 (84,5%)
	1 - Feminino	110 (15,5%)
Idade	0 - <20 anos	301 (42,5%)
	1 - 20 anos	408 (57,5%)
Origem	0 – Outras Cidades	242 (34,1%)
	1 – Campina Grande	467 (65,9%)
Tipo de Escola	0 – Privado	274 (38,6%)
	1 – Público	435 (61,4%)
Forma de Ingresso	0 – Vestibular	300 (42,3%)
	1 - ENEM	409 (57,7%)
Turno	0 - Diurno	344 (48,5%)
	1 - Noturno	365 (51,5%)

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

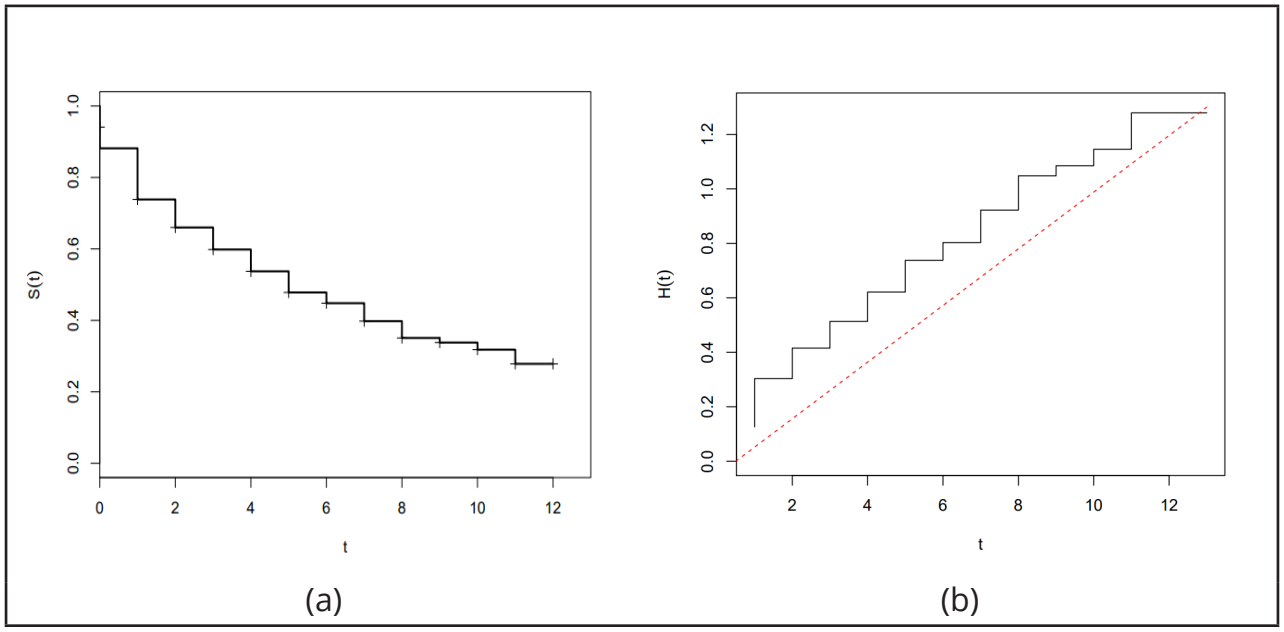
4.2 Análise descritiva

Inicia-se a análise descritiva dos dados observando o comportamento da curva de sobrevivência obtida pelo estimador de Kaplan e Meier (1958). A partir do gráfico da função de sobrevivência e da função de risco acumulada, pode-se obter informações sobre modelos que podem se ajustar aos dados.

A Figura 2 (a) mostra o comportamento empírico da função de sobrevivência e pode-se obter informação sobre o decaimento pouco acentuado da curva de sobrevivência. No tempo zero a probabilidade de sobrevivência não é 1, ou seja, neste tempo ocorreram falhas, sendo uma característica de dados discretos. Observa-se que em todos os tempos ocorreram censuras e que, aparentemente, o tempo mediano está em torno de t = 5 semestres.

A Figura 2 (b) mostra uma função de risco decrescente, que uma forma que contempla a distribuição LLD. Com base nessas análises descritivas, é possível supor como possíveis modelos paramétricos o Log-Logística discreta apresentado na Seção 2, deste trabalho, e a distribuição Weibull discreta (WD).

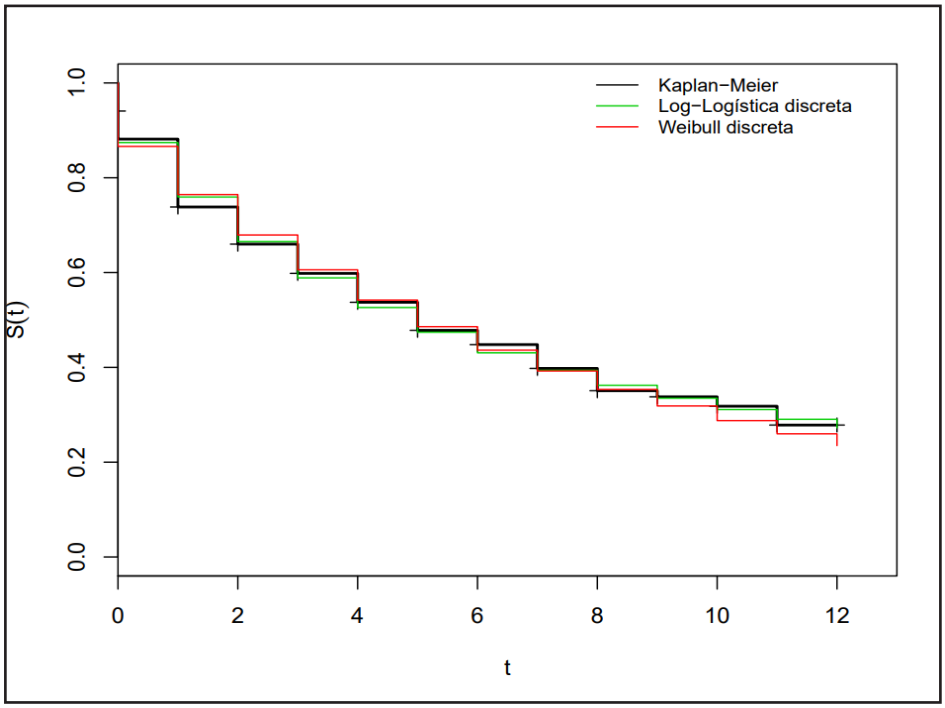
Figura 2 – (a) Curva de sobrevivência estimada pelo método de Kaplan e Meier e (b) Curva do risco acumulado para os dados dos alunos do curso de Computação



Fonte: Elaborado pelos autores com base nos dados da pesquisa (2023)

Ao ajustar os dois modelos, conforme mostra a Figura 3, percebe-se que a distribuição LLD apresenta maior proximidade com a curva empírica do que a distribuição WD.

Figura 3 – Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) e pelos modelos Log-Logística discreta e Weibull discreta



Fonte: Elaborado pelos autores com base nos dados da pesquisa (2025)

Na Tabela 9, apresenta-se as estimativas para os modelos LLD e WD, com seus respectivos erros padrão e o intervalo de confiança de 95%, calculado a partir da matriz de informação de Fisher e utilizando-se da transformação logarítmica. Além disso, para a construção desses intervalos se faz necessário obter uma estimativa para o erro-padrão utilizando-se o método delta, conforme descrito na subseção 2.3.

Tabela 9 – Estimativas dos modelos Log-Logística discreta e Weibull discreta

Modelo	Parâmetros	Estimativas	Erro-Padrão	Intervalo de confiança (95%)
LLD	α	5,479414	0,37952262	[4,82663; 6,22048]
	Y	1,139994	0,05985065	[1,03549; 1,25505]
WD	q	0,866043	0,01168612	[0,85268; 0,87962]
	Y	0,900524	0,04686015	[0,84811; 0,95618]

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Segundo Nakano e Carrasco (2006), o erro máximo cometido na estimação pode ser definido como um teste estatístico de ajuste de modelos, utilizando as diferenças entre as estimativas do modelo estudado e as estimativas de Kaplan e Meier (1958), expresso por:

$$\epsilon = max \left| \hat{S}(t) - \hat{S}_{km}(t) \right| \quad (24)$$

De acordo com a Tabela 10, nota-se que as estimativas dos dois modelos estão próximas das estimativas não paramétricas. No entanto, percebe-se que nos últimos tempos o modelo LLD se aproxima mais do que o WD. Dessa forma, ao utilizar o valor de ε = 0,020962, há indícios que o modelo mais adequado é o LLD.

Outros procedimentos para seleção de modelos utilizam critérios de informação. Sua ideia básica é selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Na Tabela 11, apresenta-se os critérios AIC, AICc e BIC para os modelos LLD e WD. Nos três critérios, o modelo LLD apresentou menores valores, e isso significa que a distribuição Log-Logística discreta se ajusta melhor aos dados do que a Weibull discreta.

Tabela 10 – Estimativas da função de sobrevivência pelo método de Kaplan e Meier pelo modelo LLD e WD

Tempo	K-M	LLD	WD
0	0,881523	0,874256	0,866043
1	0,738356	0,759318	0,764543
2	0,659963	0,665236	0,679230
3	0,598284	0,588740	0,605817
4	0,537235	0,526071	0,541876
5	0,478167	0,474156	0,485757
6	0,448073	0,430650	0,436237
7	0,397821	0,393787	0,392363
8	0,350711	0,362231	0,353365
9	0,337958	0,334967	0,318610
10	0,318078	0,311212	0,287569
11	0,278318	0,290357	0,259791
12	0,278318	0,271921	0,234895
	ε	**0,020962**	0,043424

Fonte: Elaborado pelos autores com base nos dados da pesquisa (2025)

Tabela 11 – Critérios de informação AIC, AICc e BIC segundo os modelos LLD e WD

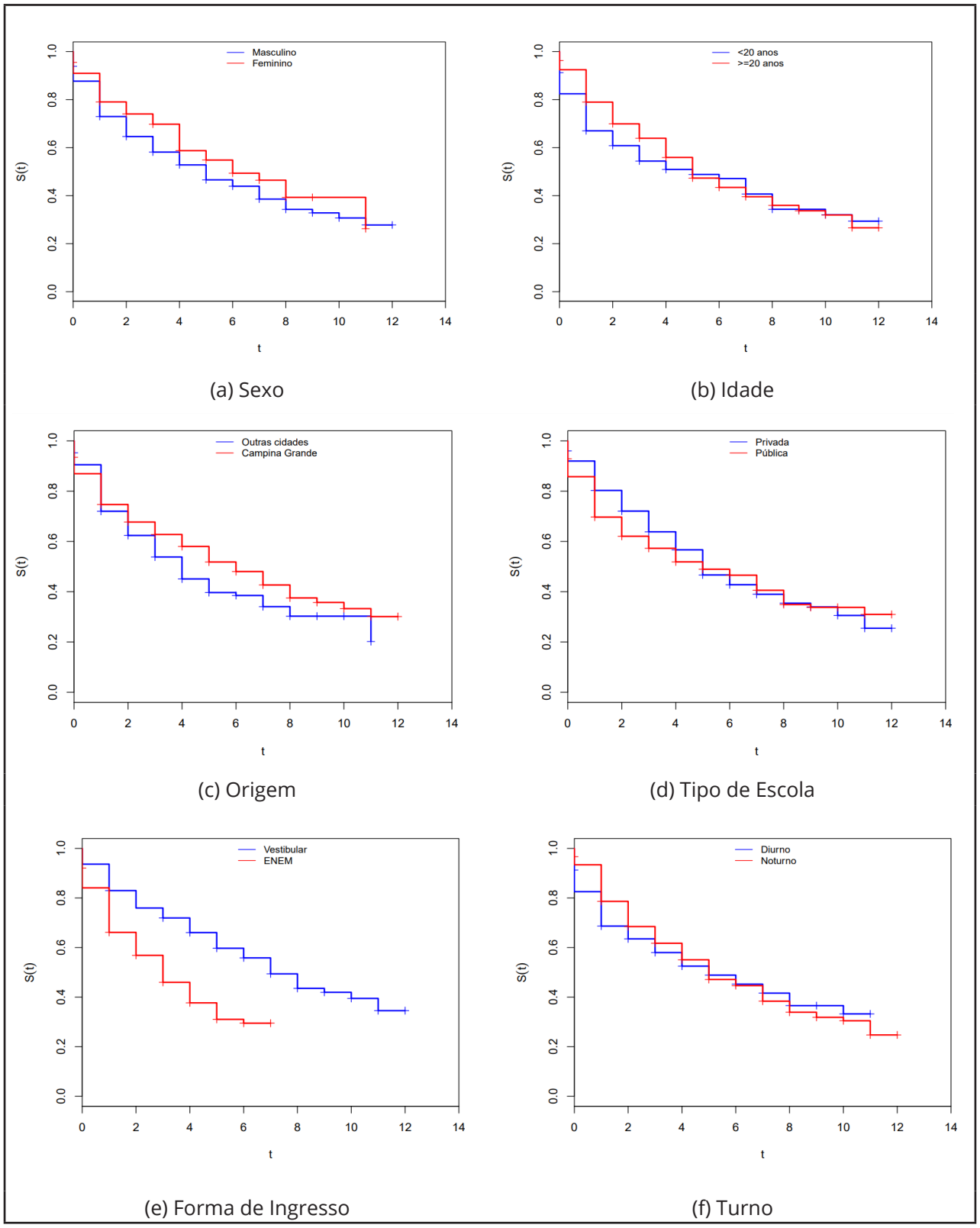
Modelos	AIC	AICc	BIC
LLD	**2036,226**	**2036,260**	**2043,998**
WD	2041,763	2041,797	2049,535

Fonte: Elaborado pelos autores com base nos dados da pesquisa (2025)

Pelas análises anteriores, conclui-se que um modelo adequado para modelar o tempo de sobrevivência dos alunos do curso de Ciência da Computação é o modelo LLD. Como neste banco de dados tem-se a presença das covariáveis dos alunos, o próximo passo é a construção de um modelo de regressão, conforme apresentado na Seção 2.2. Para tanto, primeiramente será verificado quais covariáveis estão influenciando no tempo de sobrevivência por meio da análise descritiva.

Uma forma preliminar de verificar a influência da covariável é estimar a função de sobrevivência de Kaplan e Meier em função da covariável de interesse. Caso as curvas de sobrevivência empírica apresentem formas diferentes, há indícios que essa covariável está influenciando a resposta e ela deve ser considerada na modelagem paramétrica.

Figura 4 – Curvas de sobrevivência estimadas pelo método de Kaplan e Meier para as seis covariáveis dos alunos do curso de Computação



Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Pela Figura 4, observa-se o comportamento das covariáveis em estudo. Dentre as seis covariáveis estudadas, a que apresentou uma diferença mais relevante foi a *forma de ingresso*, apresentada na Figura 5(e). As demais covariáveis aparentam ter efeito menos significativos. Outra metodologia que pode ser utilizada para verificar a influência das covariáveis no tempo de sobrevivência são os testes de hipóteses não paramétricos de Wilcoxon ou de Log-Rank. Esses testes consideram as seguintes hipóteses:

{H_0: *As curvas de sobrevivência são iguais* H_1: *As curvas de sobrevivência são diferentes*}

Ao considerar o nível de significância de 5%, como mostra a Tabela 12, as covariáveis *Idade* e *Forma de ingresso* apresentam curvas de sobrevivência diferentes. E ao nível de 10%, as covariáveis *tipo de escola* e *turno* também apresentam curvas de sobrevivência diferentes. Com esse resultado, pode-se construir um modelo de regressão LLD.

Tabela 12 – Teste de hipóteses de comparação entre curvas de sobrevivência utilizando covariáveis dos alunos de Computação

Covariáveis	Teste	Estatística do teste	p-valor
Sexo	Wilcoxon	2,5	0,1110
Idade	Wilcoxon	6,2	0,0131
Origem	Wilcoxon	2,0	0,1560
Tipo de Escola	Wilcoxon	3,3	0,0677
Forma de Ingresso	Log-Rank	41,3	1,32×
Turno	Wilcoxon	3,5	0,0605

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Para as covariáveis *Sexo*, *Idade*, *Origem*, *Tipo de escola* e *Turno* foi utilizado o teste similar de Wilcoxon devido ao comportamento das curvas de sobrevivência. Para a covariável *Forma de ingresso* foi utilizado o teste Log-Rank, por não apresentar nenhum cruzamento entre as curvas de sobrevivência.

4.3 Modelo de regressão Log-Logística Discreta

Por meio da análise descritiva realizada na Subseção 4.2, foi possível identificar as covariáveis que apresentam diferenças mais relevantes nas curvas

de sobrevivência. Essa diferença pode ser indício de que essas covariáveis estão exercendo influência sobre a variável resposta.

Para a construção do modelo de regressão LLD, utilizou-se o processo de seleção de modelo *backward*, que incorpora inicialmente todas as covariáveis e retira-se aquelas com pior desempenho, até obter-se um modelo com todas as covariáveis significativas a um determinado nível de significância.

Dessa forma, a Tabela 13 apresenta as estimativas do modelo de regressão Log-Logística discreta.

Tabela 13 – Estimativas dos parâmetros do modelo de regressão Log-Logística discreta

Modelo	Parâmetros	Estimativas	e^{β}	Erro-padrão	p-valor
MRLLD	β_0	1,0380590	2,823731	0,1151007	$0,0000 \times 10^{-0}$
	β_1 (Idade20 anos)	0,6025797	1,826825	0,1102721	$4,6427 \times 10^{-8}$
	β_2 (Escola=Pública)	-0,3652644	0,694013	0,1056483	$5,4549 \times 10^{-4}$
	β_3 (Ingresso=ENEM)	-0,7235238	0,485040	0,1062902	$9,9614 \times 10^{-12}$
	β_4(Turno=Noturno)	0,3374152	1,401321	0,1059122	$1,4435 \times 10^{-3}$
	γ	2,0220465	-	0,1202715	-

Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

Ao analisar os resultados das estimativas dos parâmetros do modelo de regressão Log-Logística discreta apresentado na Tabela 13, observa-se o valor de $\hat{\gamma} = 2,022$. Sendo assim, o modelo ajustado apresenta taxa de falha unimodal. Em relação à interpretação dos coeficientes de regressão, por β_1 tem-se que para alunos com idade maior ou igual a 20 anos o tempo mediano aproximado até a evasão é 83% maior do que para alunos com idade menor que 20 anos

O coeficiente β_2 indica a procedência no ensino médio dos alunos. Tem-se que o tempo mediano aproximado até a evasão dos alunos é 1,44 maior dos alunos que estudaram em escola privada do que aqueles que estudaram em escola pública.

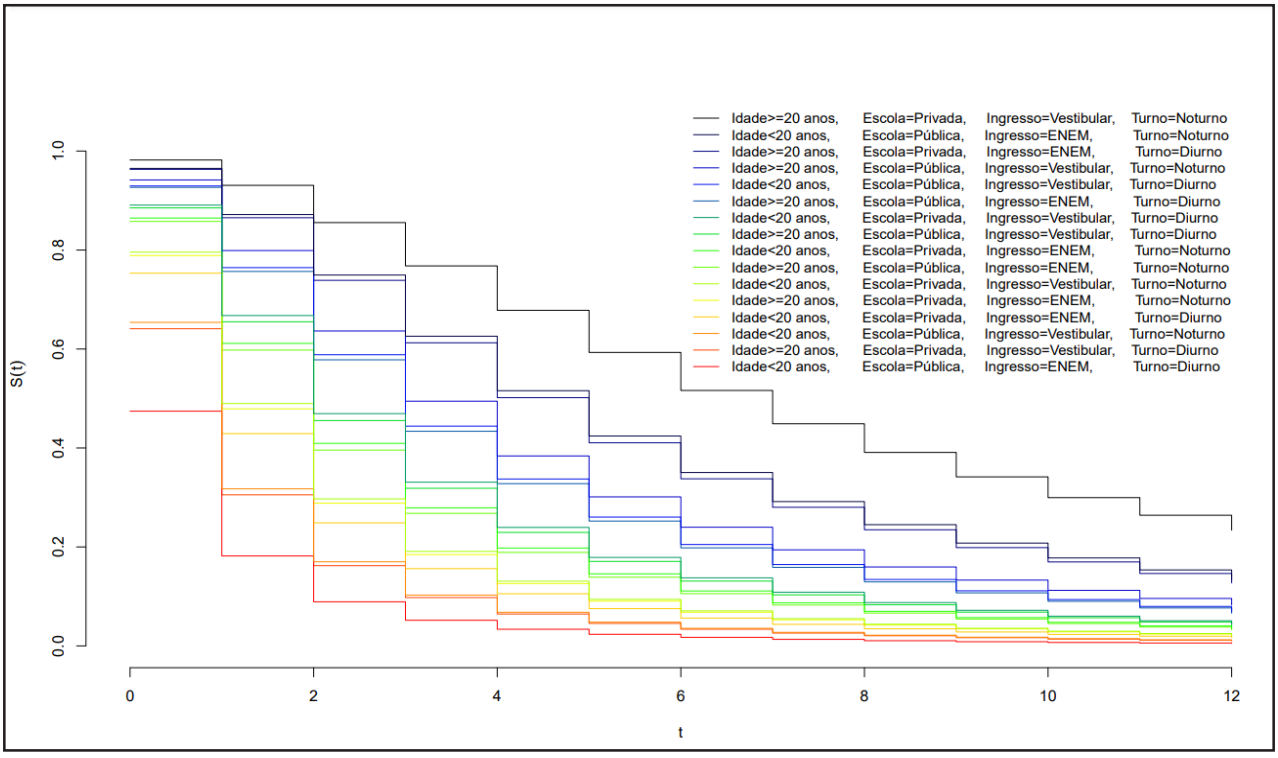
O resultado de β_3 leva a concluir que a forma de ingresso tem influência na evasão. Os alunos que ingressaram no curso via vestibular têm o tempo mediano

aproximado para a evasão duas vezes maior do que os alunos que ingressaram pelo Exame Nacional do Ensino Médio (ENEM).

Com respeito a β_4 que se refere ao turno do curso, como sua estimativa foi positiva, significa que os alunos que pertencem ao curso noturno têm o tempo mediano aproximado de 1,4 vezes maior do que os alunos que pertencem ao curso do turno diurno.

De maneira geral, essas interpretações estão coerentes com a análise descritiva dos dados. Isso demonstra a consistência do modelo ajustado. Além disso, outra maneira de verificar a contribuição de cada covariável na probabilidade de sobrevivência é através da construção das curvas de sobrevivência com base no modelo, em função das combinações das covariáveis. Dessa forma, observa-se que a maior curva de sobrevivência está associada ao grupo de estudantes com idade maior ou igual a 20 anos, que concluíram o ensino médio em escola privada, em que o ingresso no ensino superior foi por meio do vestibular e com o turno do curso noturno, ver Figura 5.

Figura 5 – Curvas de sobrevivência estimadas pelo modelo de regressão Log-Logística Discreta



Fonte: Elaborado pelos autores conforme dados da pesquisa (2025)

5 CONSIDERAÇÕES FINAIS

Neste trabalho foi estudada a distribuição Log-Logística discreta e introduzido o modelo de regressão LLD. As estimativas dos parâmetros do MRLLD foram obtidas pelo método de máxima verossimilhança. A performance das estimativas de máxima verossimilhança foram testadas via simulação Monte Carlo. O método mostrou ser robusto especialmente para valores de parâmetros de forma e escala próximos. Na aplicação em um banco de dados de alunos do curso de Ciência da Computação o MRLLD permitiu dar uma boa interpretação das covariáveis no tempo de evasão (variável resposta). Além disso, a representação gráfica dos resultados dos modelos de sobrevivência para cada combinação de covariáveis permitiu uma visualização mais clara dos grupos que apresentam maiores e menores probabilidade de sobrevivência. Dessa forma, observou-se que alunos com idade maior ou igual a 20 anos, que veio de escola privada, que ingressou via vestibular e estuda em turno noturno, apresentou maior probabilidade de sobrevivência, já a combinação oposta tem a menor probabilidade de sobrevivência (<20 anos, escola pública, ingresso por ENEM, turno diurno).

Esse resultado sugere que a Coordenação do curso de Ciência da Computação da Universidade Estadual da Paraíba poderia desenvolver estratégias de acompanhamento e monitoramento dos estudantes com perfil mais propenso à evasão - caracterizado por idade inferior a 20 anos, origem em escola pública, ingresso via ENEM e matrícula no turno diurno. Tais estratégias também poderiam ser estendidas a todos os alunos do curso.

Uma possível estratégia seria realizar o acompanhamento mais próximo dos discentes nos quatro primeiros semestres, a fim de identificar disciplinas e situações associadas à evasão. A partir desse diagnóstico, seria viável oferecer atividades de monitoria voltadas especificamente para as disciplinas com maior influência nesse processo.

Além disso, recomenda-se disponibilizar tais disciplinas durante o período de Verão Acadêmico e incentivar a participação dos estudantes, possibilitando o avanço na grade curricular, a redução do tempo de permanência e o aumento da motivação para a continuidade no curso.

Também se sugere a implementação de ações que despertem o interesse dos alunos, apresentem as diversas áreas de atuação da Computação e promovam maior aproximação com a prática profissional. Projetos de Extensão, de Iniciação Científica e palestras com profissionais atuantes no mercado de trabalho, por exemplo, podem favorecer o contato mais próximo entre discentes, professores e a comunidade, possibilitando, assim, maior engajamento na vida universitária.

REFERÊNCIAS

ALDAHLAN, M.A. Alpha Power Transformed Log-Logistic Distribution with Application to Breaking Stress Data. **Advances in Mathematical Physics,** v. 1, p. 1-9, 2020.

ASHKAR, F.; MAHDI, S. Fitting the log-logistic distribution by generalized moments. **Journal of Hydrology**, v. 328, n. 3, p. 694–703, 2006.

BENNETT, S. Log-Logistic Regression Models for Survival Data. **Journal of the Royal Statistical Society**. Series C (Applied Statistics), v. 32, n. 2, p. 165-171, 1983.

COLOSIMO, E. A.; GIOLO, S.R. **Análise de Sobrevivência Aplicada**. 2. ed. São Paulo: Blucher, 2024.

COLLET, D. **Modelling Survival Data in Medical Research**. London: Chapman and Hall, 2003.

FILHO, R. L. L. S.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, p. 641-659, 2007.

FRITSCH, R.; ROCHA, C. S.; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. **Educação em Questão,** v. 52, n. 38, p. 81 – 108, 2015.

HASHEMIAN, A.H.; GARSHASBI, M.; POURHOSEINGHOLI, M.A.; ESKANDARI, S. A comparative study of cox regression vs. loglogistic regression (with and without its frailty) in estimating survival time of patients with colorectal cancer. **Journal Medical Biomedical Sciences**, v. 6, n. 1, p. 35–43, 2017.

HOSMER, D.; LEMESHOW, S. **Applied Survival Analysis**. New York: Wiley, 1999.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Censo Demográfico 2022**. 2022. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/22827-censo-demografico-2022.html. Acesso em: 15 jun. 2024.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Sinopse Estatística da Educação Superior 2020**. Brasília: Inep, 2022. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/sinopses-estatisticas/educacao-superior-graduacao. Acesso em: 15 jun. 2024.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete bservations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457–481, 1958.

KHALEEQA, J.; AMANULLAHA, M.; ABDULRAHMANB, A.T.; HAFEZC, E.; ABD EL-RAOUFD, H. Influence diagnostics in Log-Logistic regression model with censored data**. Alexandria Engineering Journal**, v. 61, n. 3, p. 2230-2241, 2022.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. New York: John Wiley Sons, 2011.

LIMA JUNIOR, P.; SILVEIRA, F. L.; OSTERMANN, F. Survival analysis applied to student flow in undergraduate Physics courses: an example from a Brazilian university. **Revista Brasileira de Ensino de Física**, v. 34, n. 1, p. 1-10, 2012.

LIMA, J. G. **Modelo de regressão log-logístico discreto para dados da política de zoneamento e uso do solo na presença de observações censuradas**. 2018. Trabalho de conclusão de curso (Bacharelado em Estatística) - Universidade de Brasília, Brasília, 2018.

LOUZADA-NETO, F.; PEREIRA, B. Modelos em análise de sobrevivência. **Cadernos Saúde Coletiva**, Rio de Janeiro, v. 8, n. 1, p. 8–26, 2000.

NAKAGAWA, T.; OSAKI, S. The discrete weibull distribution. **IEEE Transactions on Reliability**, v.24, n.5, 300–301, 1975.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. **Trends Computational and applied mathematics**, v. 7, n.1, p. 91-100, 2006.

ROSA, C. M. Limites da democratização da educação superior: Entraves na permanência e a evasão na universidade federal de Goiás. **Revista Poíesis Pedagógica**, v. 37, p. 641 – 659, 2007.

R CORE TEAM, R. *et al.* **R**: A language and environment for statistical computing.

Vienna: Statistical Computing, Vienna, Austria, 2025.

SANTOS, D. F. **Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência**. 2017. 81 f. Dissertação (Mestrado em Estatística) - Universidade de Brasília, Brasília, 2017.

SEMESP. **Mapa do Ensino Superior.** 2022. Disponível em: https://www.semesp.org.br/mapa/edicao-12/. Acesso em: 15 jun. 2024.

TAHIR, M.H.; MANSOOR,M.; ZUBAIR, M.; HAMEDANI, G. Mcdonald Log-logistic distribution with an application to breast cancer data. **Mathematics, Statistics and Computer Science Faculty Research and Publications**, v. 13, n. 1, p. 65-82, 2014.

Contribuição de Autoria

1 – Damião Flávio dos Santos

Graduação em Estatística pela UEPB e mestrado em Estatistica pela UnB.
https://orcid.org/0000-0001-9411-1403 • d.flaviostate@gmail.com

Contribuição: Administração de projetos, Análise formal, Conceitualização, Curadoria de dados, Escrita – revisão e edição, Investigação, Metodologia, Obtenção de financiamento, Recursos, Redação – rascunho original, Software, Supervisão, Validação, Visualização

2 – Cira Etheowalda Guevara Otiniano

Graduação em Matemática pela Universidad Nacional de Trujillo, mestrado em Matemática pela UnB e doutorado em Matemática Aplicada pela UnB.
https://orcid.org/0000-0002-5619-0478 • cira@unb.br

Contribuição: Administração de projetos, Análise formal, Conceitualização, Escrita – revisão e edição, Investigação, Metodologia, Obtenção de financiamento, Recursos, Redação – rascunho original, Supervisão, Validação, Visualização

3 – Juliana Betini Fachini Gomes

Graduação em Estatística pela UNESP, mestrado em Agronomia e doutorado em Ciências pela USP.
https://orcid.org/0000-0002-6677-4768 • jfachini@unb.br

Contribuição: Administração de projetos, Análise formal, Conceitualização, Escrita – revisão e edição, Investigação, Metodologia, Obtenção de financiamento, Recursos, Redação – rascunho original, Supervisão, Validação, Visualização,

Como citar este artigo

SANTOS, D. F.; OTINIANO, C. E. G.; GOMES, J. B. F. Avaliação da evasão escolar de alunos de um curso de graduação via um modelo de regressão log-logística discreta. **Ciência e Natura**, Santa Maria, v. 48, e88599, 2026. DOI 10.5902/2179460X88599. Disponível em: https://doi.org/10.5902/2179460X88599.