



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Presidente Prudente

LETICIA TIAGO MAIA SANTOS

**CONSTRUÇÃO DE UM MODELO DE REGRESSÃO EM ANÁLISE DE
SOBREVIVÊNCIA PARA O ESTUDO DO PERFIL DE COMPRADORES DA OLIST
QUE REALIZARAM COMPRAS ONLINE NO PERÍODO DE 2016 A 2018 EM SÃO
PAULO**

PRESIDENTE PRUDENTE

2024

LETICIA TIAGO MAIA SANTOS

**CONSTRUÇÃO DE UM MODELO DE REGRESSÃO EM ANÁLISE DE
SOBREVIVÊNCIA PARA O ESTUDO DO PERFIL DE COMPRADORES DA OLIST
QUE REALIZARAM COMPRAS ONLINE NO PERÍODO DE 2016 A 2018 EM SÃO
PAULO**

Relatório Final de Trabalho de Conclusão de
Curso apresentado ao Curso de Graduação
em Estatística da FCT/Unesp para
aproveitamento na disciplina de Trabalho de
Conclusão de Curso.

Orientador: Prof. Dr. Mário Hissamitsu
Tarumoto.

PRESIDENTE PRUDENTE

2024

S237c	Santos, Leticia Tiago Maia Construção de um modelo de regressão em análise de sobrevivência para o estudo do perfil de compradores da Olist que realizaram compras online no período de 2016 a 2018 em São Paulo / Leticia Tiago Maia Santos. -- Presidente Prudente, 2024 58 p. Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente Orientador: Mário Hissamitsu Tarumoto 1. Análise de sobrevivência. 2. Modelo Paramétrico. 3. Modelo de Regressão Weibull. 4. Olist. 5. Marketplace. I. Título.
-------	--

TERMO DE APROVAÇÃO

LETICIA TIAGO MAIA SANTOS

CONSTRUÇÃO DE UM MODELO DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA PARA O ESTUDO DO PERFIL DE COMPRADORES DA OLIST QUE REALIZARAM COMPRAS ONLINE NO PERÍODO DE 2016 A 2018 EM SÃO PAULO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Documento assinado digitalmente
 **MARIO HISSAMITSU TARUMOTO**
Data: 16/12/2024 10:08:49-0300
Verifique em <https://validar.iti.gov.br>

Orientador: _____

Prof. Dr. Mário Hissamitsu Tarumoto.

Departamento de Estatística

Documento assinado digitalmente
 **GUILHERME APARECIDO SANTOS AGUILAR**
Data: 16/12/2024 09:30:00-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Guilherme Aparecido Santos Aguilár.

Departamento de Estatística

Presidente Prudente, 06 de dezembro de 2024

Resumo

O aumento crescente de usuários na internet impulsionou o surgimento do segmento *E-commerce*¹, com empresas adotando o modelo de *marketplace*² possibilitando conectar clientes a diferentes lojas e serviços. Empreendimento como a Olist, uma startup brasileira fundada em 2015 por Tiago Dalvi, emergiram com a proposta de oferecer espaço para pequenos empresários nos principais *marketplaces* do país. Este estudo tem como propósito analisar os dados fornecidos pela Olist³ no período de 2016 a 2018 no estado de São Paulo para investigar como o tipo de compra influencia o intervalo de tempo entre a penúltima e a última compra realizada pelos clientes. Será empregado um modelo de regressão em análise de sobrevivência para determinar qual modelo se adapta melhor à distribuição dos dados, seja ele semiparamétrico ou paramétrico. Como a estratégia de lucro da Olist está diretamente ligada à quantidade de produtos vendidos por vendedores ativos na plataforma, a compreensão das covariáveis que afetam o tempo entre as compras, como tipo de produto e valor do frete, pode fornecer insights importantes para a empresa, possibilitando a implementação de medidas como a abertura de novos centros de distribuição em pontos estratégicos, visando reduzir os custos de frete e, consequentemente, alcançar maior número de consumidores. Uma das informações notadas nesse trabalho foi que os clientes que compram mais frequentemente tendem a fazer compras de maior valor além disso seu tempo mediano de recompra é de 83 dias, enquanto pessoas que compram menos de 2 compras possui tempo de recompra de 140 dias. O modelo de regressão que melhor ajustou foi o de Weibull.

Palavras-chave: Análise de sobrevivência; Olist; *Marketplace*; Modelo Paramétrico; Modelo Semiparamétrico; Modelo de Regressão Weibull.

¹ *E-commerce* é o comércio virtual, ou seja, o e-commerce é composto por todas as empresas que vendem seus produtos virtualmente.

² *Marketplaces* são as plataformas onde os produtos são vendidos. Alguns exemplos de marketplaces são: Amazon, Americanas, Mercado Livre, etc.

³ Olist é uma empresa de departamentos e tem como principal objetivo facilitar pequenos vendedores a se inserir em grandes lojas virtuais.

Abstract

The increasing number of internet users has driven the emergence of the *E-commerce* segment, with companies adopting the *marketplace* model, enabling them to connect customers to different stores and services. Enterprises like Olist, a Brazilian startup founded in 2015 by Tiago Dalvi, have emerged with the proposition of providing space for small entrepreneurs on the country's major *marketplaces*. This study aims to analyze the data provided by Olist from 2016 to 2018 to investigate how the type of purchase influences the time interval between the penultimate and the last purchase made by customers. A survival analysis regression model will be employed to determine which model better fits the data distribution, whether it be semiparametric or parametric. As Olist's profit strategy is directly linked to the quantity of products sold by active sellers on the platform, understanding the covariates that affect the time between purchases, such as product type and shipping cost, can provide important insights for the company, enabling the implementation of measures such as the opening of new distribution centers in strategic locations, aiming to reduce shipping costs and consequently reach a greater number of consumers. One of the information noted in this work was that customers who buy more frequently tend to make higher value purchases and their median repurchase time is 83 days, while people who buy less than 2 purchases have a repurchase time of 140 days. The best fitting regression model was Weibull.

Keywords: Survival Analysis; Olist; *Marketplace*; Parametric Model; Semiparametric Model. Weibull Regression Model.

Lista de Ilustrações

Figura 1 - Função de sobrevivência teórica.....	13
Figura 2 - Exemplos de funções de risco.....	14
Figura 3 - Gráfico de Kaplan-Meier proveniente dos dados deste trabalho.....	15
Figura 4 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Exponencial.	17
Figura 5 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull.....	18
Figura 6 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Normal.....	19
Figura 7 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Gama.	21
Figura 8 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Logística.	22
Figura 9 - Esquema de organização das bases.	30
Figura 10 - Histograma da variável Tempo.....	35
Figura 11 - Curva de Sobrevivência categorizada por quartis do total da compra. ..	37
Figura 12 - Curva de Sobrevivência categorizada por quartis do preço do produto.	38
Figura 13 - Curva de Sobrevivência categorizada por quartis do valor do frete.	39
Figura 14 - Gráficos de dispersão das covariáveis numéricas.	40
Figura 15 - Volumetria de pedidos ao longo das safras.	41
Figura 16 - Curva de sobrevivência para a covariável nc.	42
Figura 17 - Análise dos resíduos de Cox-Snell do modelo de regressão Exponencial ajustado para os dados de compradores da Olist.	48
Figura 18 - Análise dos resíduos de Cox-Snell do modelo de regressão Weibull ajustado para os dados de compradores da Olist.	49
Figura 19 - Análise dos resíduos de Cox-Snell do modelo de regressão Log-Normal ajustado para os dados de compradores da Olist.	50

Lista de tabelas

Tabela 1 - Relação categoria x tipo produto da base.....	33
Tabela 2 - Estatísticas das covariáveis numéricas.	36
Tabela 3 - Análise da variável categórica Tipo do Produto.....	41
Tabela 4 - Estimativas da sobrevivência para os tempos de compra usando-se o estimador de Kaplan-Meier e os modelos Exponencial, Weibull, Log-Normal e Log-Logística.....	45
Tabela 5 - Estimativas para os dados de compras feitas por consumidores da Olist.	46
Tabela 6 - AIC dos modelos de regressão.....	47
Tabela 7 - Resultados dos testes da razão de verossimilhanças (TRV).....	51

SUMÁRIO

1. INTRODUÇÃO	9
2 REVISÃO BIBLIOGRÁFICA.....	11
2.1 ANÁLISE DE SOBREVIVÊNCIA	11
2.1.1 <i>Função de Sobrevivência</i>	12
2.1.2 <i>Função de Risco</i>	13
2.1.3 <i>Estimador de Kaplan-Meier</i>	14
2.2 DISTRIBUIÇÕES DE PROBABILIDADE	16
2.2.1 <i>Distribuição Exponencial</i>	16
2.2.2 <i>Distribuição Weibull</i>	17
2.2.3 <i>Distribuição Log-Normal</i>	18
2.2.4 <i>Distribuição Gama</i>	20
2.3 MODELOS DE REGRESSÃO PARAMÉTRICOS.....	22
2.3.1 <i>Modelo de Regressão Exponencial</i>	22
2.3.2 <i>Modelo de Regressão Weibull</i>	23
2.3.3 <i>Modelo de Regressão Log-Normal</i>	24
2.3.4 <i>Modelo de Regressão Logístico</i>	24
2.4 ESTIMAÇÃO DE PARÂMETROS	25
2.5 ANÁLISE DE RESÍDUOS.....	26
2.5.1 <i>Resíduos de Cox-Snell</i>	26
2.6 ESCOLHA DO MODELO	27
2.6.1 <i>Métrica AIC</i>	27
2.6.2 <i>Método dos gráficos</i>	28
2.6.3 <i>Teste de Hipóteses</i>	28
3 APLICAÇÃO DOS DADOS	30
3.1 TRATAMENTO DA BASE	30
4 RESULTADOS.....	35
4.1 <i>Análise Descritiva</i>	35
4.2 MODELAGEM ESTATÍSTICA	44
4.2.1 <i>Modelos de regressão paramétricos</i>	44
4.2.2 <i>Adequação dos modelos</i>	46
CONCLUSÕES	52
REFERÊNCIAS.....	53
APÊNDICES.....	55
APÊNDICE 1 – DICIONÁRIOS DOS DADOS OBTIDOS NA PLATAFORMA KAGGLE.....	55

1. INTRODUÇÃO

Atualmente vivemos em um mundo cada vez mais conectado, em 2022, segundo informações da pesquisa TIC Domicílios 2022, havia cerca de 149 milhões de usuários de internet. Por conta desse número elevado de pessoas conectadas, impulsionou o crescimento das compras online. Uma pesquisa realizada pela StartSe em colaboração com o Instituto Pesquisa na Hora revelou que aproximadamente 57% dos consumidores têm o hábito de realizar compras pela internet com regularidade.

Sabendo do aumento crescente de usuários que utilizam a internet, algumas empresas começaram a atuar no segmento de *e-commerce* por meio de *marketplace*. Esse formato possibilita conectar clientes a diferentes lojas e serviços, isto é, pode-se comprar um produto em um determinado *marketplace* que está sendo vendido por outra pessoa.

A partir desse modelo de negócio, surgiram empresas com intuito de oferecer aos pequenos empreendedores, normalmente sem presença digital, espaço nos grandes *marketplaces* do país. Esse é o caso da Olist uma startup brasileira fundada em 2015 por Tiago Dalvi com intuito de empoderar o comércio que transforma negócios em gigantes do *e-commerce*.

Com esse sistema de venda, os pequenos empreendedores têm a oportunidade de anunciar seus produtos utilizando o nome da Olist, que por sua vez, anunciam em diversos *marketplaces* como Mercado Livre, Magalu, Amazon, Inter entre outras. Esse fato provoca um aumento nas vendas, já que em lojas virtuais têm o privilégio de poder vender diariamente, 24 horas por dia sem precisar arcar com custos trabalhistas, tem-se um alto alcance geográfico podendo ter produtos vendidos no mundo todo, além disso, lojas virtuais tendem a apresentar custos fixos e variáveis mais baixo, viabilizando o empreendedorismo para número maior de pessoas que não precisam investir muito dinheiro como fariam se fossem abrir uma loja física. Para os *marketplaces* a vantagem é cobrar pelo uso da plataforma e aplicar taxas sobre o produto vendido de cada lojista.

Neste trabalho o objetivo é realizar um estudo com os dados fornecidos pela Olist disponível no site *Kaggle*. Neste estudo o intuito é entender se o tipo de compra afeta o tempo de retorno após a penúltima compra. Para isso será utilizado um modelo de regressão em análise de sobrevivência para poder entender quais as covariáveis que explicam o tempo entre a penúltima e última compra realizadas pelos clientes.

Este trabalho está dividido da seguinte forma: no capítulo 2 foi realizado a revisão da literatura das principais distribuições de probabilidade; no capítulo 3 foi apresentado a aplicação dos dados onde temos a explicação de como foi o tratamento da base; no capítulo 4 é mostrado os resultados; no capítulo 5 é apresentado a conclusão e por fim as referências e apêndice.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção são apresentados alguns conceitos de Análise de Sobrevivência importantes para a realização do estudo, de forma que seja possível atingir o objetivo deste trabalho.

2.1 Análise de Sobrevivência

A Análise de Sobrevivência é um ramo da Estatística que pode ser aplicado também nas áreas de Medicina, Finanças e Engenharia chamada nesta última de Análise de Confiabilidade.

Nesta área a variável resposta de interesse, geralmente, é o tempo até a ocorrência de um **evento de interesse**, este é denominado **tempo de falha**.

Tempo pode ser anos, meses, semanas ou dias desde o início do acompanhamento de um indivíduo até a ocorrência de um evento; alternativamente, o tempo pode referir-se à idade de um indivíduo quando um evento ocorre.

Por **evento de interesse** entendemos morte, incidência de doença, recaída da remissão, recuperação (por exemplo, retorno ao trabalho) ou qualquer experiência designada de interesse que possa acontecer a um indivíduo. No presente trabalho o evento de interesse é o retorno da compra de um cliente, medido em dias.

A principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Isto é, não se tem, nesses casos, o tempo decorrido até o evento de interesse, mas tem-se o tempo até a ocorrência da censura e, apesar de incompleta, essa informação é muito útil para a análise.

Segundo Colosimo e Giolo (2006) as censuras podem ser classificadas da seguinte maneira:

1. Censura à direita

Ocorre quando o tempo registrado é menor do que o tempo de falha. Isto é, o evento de interesse ainda não havia ocorrido quando o indivíduo foi observado. Esse tipo de censura pode ser subdividido em três casos:

- Censuras do tipo I: Ocorrem em estudos que ao serem finalizados após um período pré-estabelecido de tempo registram, em seu término, alguns indivíduos que ainda não apresentaram o evento de interesse;
- Censuras do tipo II: Ocorrem de estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos;
- Censuras do tipo aleatório: Ocorrem quando um ou mais indivíduos não podem ser acompanhados até o final do estudo ou ainda quando estes falharem por eventos alheios ao de interesse.

2. Censura à esquerda

Ocorre quando tempo registrado é maior do que o tempo de falha. Isto é, o evento de interesse já havia ocorrido quando o indivíduo foi observado;

3. Censura intervalar

A censura intervalar é um tipo mais geral de censura que acontece em estudos que os elementos da amostra são observados em ocasiões periódicas e sabe-se apenas que o evento de interesse ocorreu em um determinado intervalo de tempo. As censuras à esquerda e direita são casos particulares da censura intervalar.

Os dados de sobrevivência para o indivíduo i ($i = 1, \dots, n$) sob estudo são representados pelo par (t_i, δ_i) sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha} \\ 0 & \text{se } t_i \text{ é um tempo censurado} \end{cases}$$

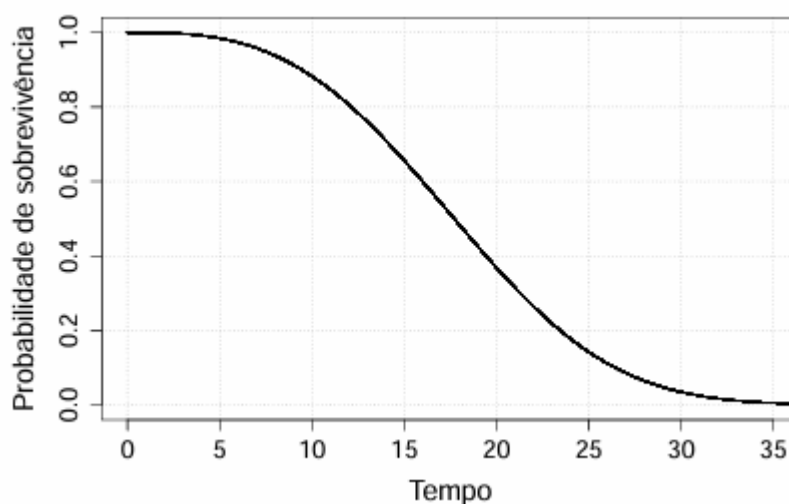
2.1.1 Função de Sobrevivência

Segundo Colosimo e Giolo (2006), a função de sobrevivência é definida como a probabilidade de uma observação não falhar até um certo tempo t , ou seja, a probabilidade uma observação sobreviver ao tempo t . Denotada por $S(t)$

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)$$

Em que $F(t)$ é a função distribuição acumulada da variável T . Essencialmente, a função de sobrevivência calculada no instante t e a probabilidade de sobrevivência por mais do que t .

Figura 1 - Função de sobrevivência teórica.



Fonte: Morettin & Singer (2021)

2.1.2 Função de Risco

Seja T uma variável aleatória positiva referente ao tempo de falha. A função de risco $\lambda(t)$, também chamada de função de **taxa de falha**, representa o risco instantâneo que o indivíduo tem de experimentar o evento de interesse em um determinado tempo t .

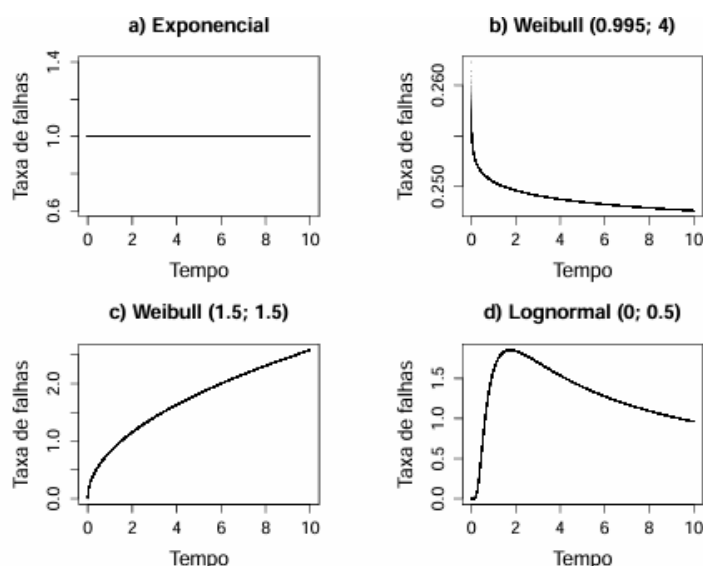
Assim, a função de taxa de falha é definida por:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Na Figura 2, temos exemplos dados por Morettin e Singer (2021) de funções de risco. Em a), o risco de falha é constante, um modelo probabilístico adequado é o modelo exponencial. Em b), o risco de falha decresce com o tempo, um modelo probabilístico adequado é um modelo Weibull. Em c), o risco de falha cresce com o tempo e um modelo probabilístico adequado seria também o de Weibull. Em d),

inicialmente o risco de falha cresce e posteriormente decresce, neste caso um modelo adequado seria o modelo Log Normal.

Figura 2 - Exemplos de funções de risco.



Fonte: Morettin & Singer (2021).

2.1.3 Estimador de Kaplan-Meier

De acordo com Colosimo e Giolo, o estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é também chamado de estimador limite-produto. É uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\hat{S}(t) = \frac{n^{\circ} \text{ de observações } \leq t}{n^{\circ} \text{ total de observações}}$$

Para dados censurados, o estimador de Kaplan-Meier (KM) é o mais utilizado na prática e é baseado na representação da sobrevivência num instante t como um produto de probabilidades condicionais de sobrevivência a intervalos de tempo disjuntos anteriores a t .

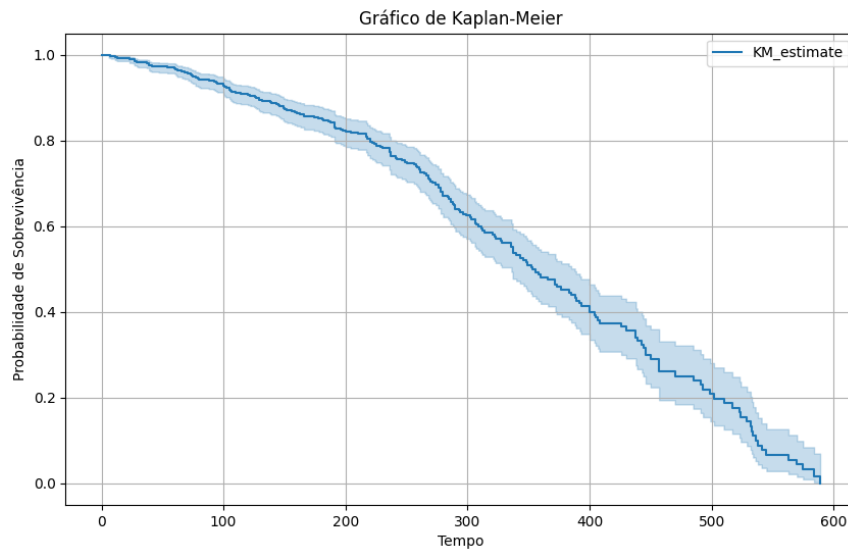
Considerando um estudo envolvendo n indivíduos e os tempos de sobrevivência $t_1, t_2, \dots, t_k$. A expressão geral do estimador KM é definida como

$$\hat{S}(t) = \prod_{j: t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j: t_j < t} \left(1 - \frac{d_j}{n_j} \right)$$

em que n_j é o número de indivíduos que estão sob risco no tempo t_j e d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_j, j = 1, 2, \dots, r$.

Como principais propriedades desse estimador podem-se citar que ele é estimador de máxima verossimilhança de $S(t)$, é considerado não-viciado para amostras grandes, é fracamente consistente e converge assintoticamente para um processo gaussiano. Este estimador também mantém esta forma em estudos envolvendo os mecanismos de censura do tipo I e II, mas não atinge $\hat{S}(t) = 0$, pois as últimas observações são censuradas.

Figura 3 - Gráfico de Kaplan-Meier proveniente dos dados deste trabalho.



Fonte: Autora (2024).

2.2 Distribuições de Probabilidade

Ao realizar modelagem para dados de sobrevivência, é essencial considerar a distribuição de probabilidade que a variável resposta pode ter.

Na Estatística, o modelo normal é frequentemente utilizado. No entanto, devido à característica particular da variável resposta, que só pode assumir valores positivos, utilizam-se distribuições que respeitem essa condição, como as distribuições Exponencial, Weibull, Log-Logística e Log-Normal, que são os casos paramétricos.

2.2.1 Distribuição Exponencial

A distribuição exponencial é um dos modelos probabilísticos mais simples usados para descrever o tempo de sobrevivência. Apresenta um único parâmetro e é a única que apresenta uma taxa de falha constante. A distribuição exponencial apresenta uma propriedade característica chamada de falta de memória da distribuição exponencial, devido a sua taxa de falha ser constante o que significa que tanto uma unidade velha quanto uma nova, que ainda não falharam, apresentam a mesma taxa de falha em um intervalo futuro.

Seja T uma variável aleatória referente ao tempo de falha. Se T segue uma distribuição exponencial, sua função de densidade é dada por:

$$f(t) = \frac{1}{\alpha} \exp\left\{-\frac{t}{\alpha}\right\}, \quad t \geq 0,$$

sendo $\alpha > 0$ o tempo médio de vida.

Ainda, as funções de sobrevivência, $S(t)$, e de taxa de falha, $\lambda(t)$, são dadas, respectivamente, por:

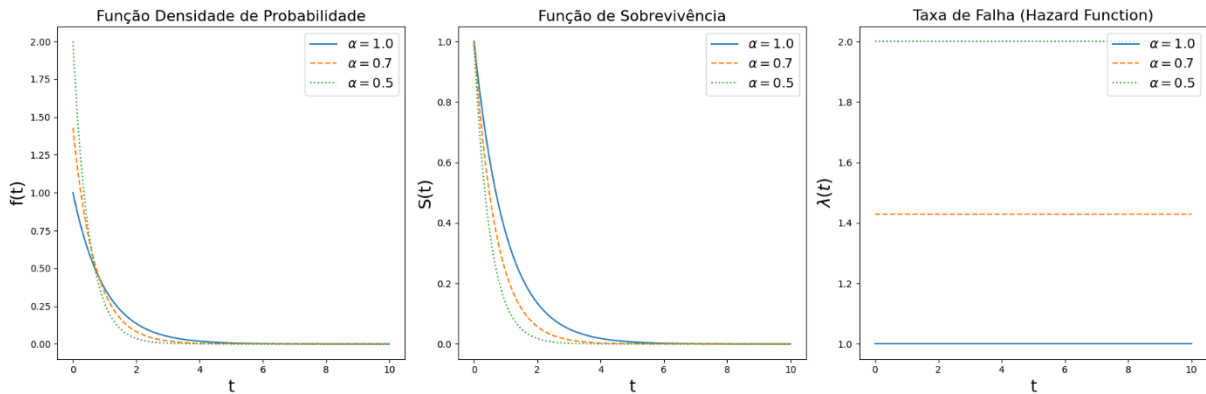
$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

e

$$\lambda(t) = \frac{1}{\alpha} \text{ para } t \geq 0$$

Na figura 4, temos, respectivamente, as funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Exponencial para $\alpha = 1,0$ (—), $0,7$ (--) e $0,5$ (···).

Figura 4 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Exponencial.



Fonte: Autora (2024).

2.2.2 Distribuição Weibull

A distribuição de Weibull é frequentemente utilizada em estudos biomédicos e industriais. Isso se deve ao fato dela apresentar uma grande variedade de formas e uma função de taxa de falha monótona, isto é, ela é crescente, decrescente ou constante.

Seja T uma variável aleatória referente ao tempo de falha. Se T segue uma distribuição de Weibull, sua função de densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad t \geq 0,$$

sendo γ o parâmetro de forma e α o parâmetro de escala, ambos positivos.

As funções de sobrevivência, $S(t)$, e de taxa de falha, $\lambda(t)$, são dadas, respectivamente, por

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}$$

para $t \geq 0$, $\gamma > 0$ e $\alpha > 0$.

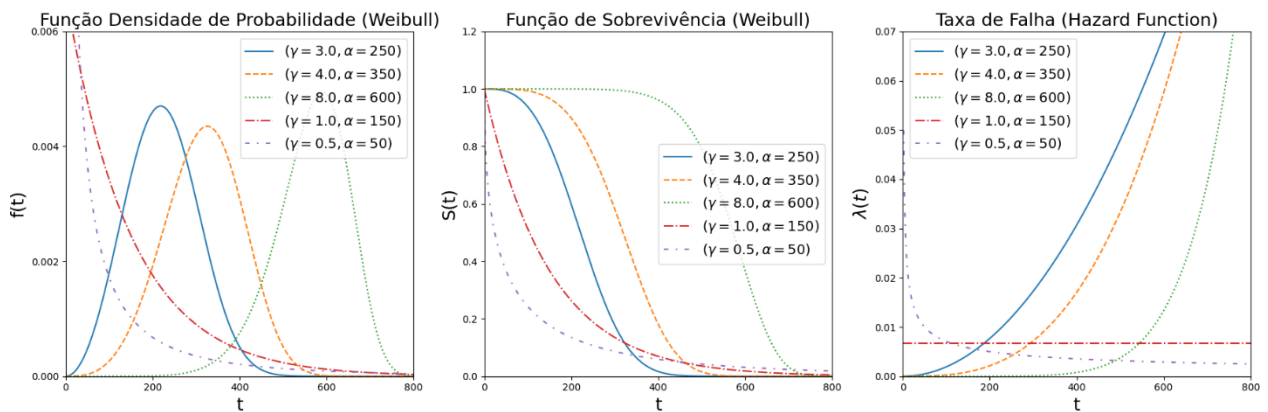
Quando o parâmetro de forma for igual a 1, tem-se como caso particular a distribuição Exponencial, com função de taxa de falha constante. Quando, essa função é crescente e para valores menores do, $\gamma > 1$, ela é decrescente.

A interpretação do parâmetro γ é da seguinte forma:

- Se $\gamma > 1$, aumenta o risco do evento;
- Se $\gamma < 1$, diminui o risco do evento;
- Se $\gamma = 1$, a variável não apresenta efeito no risco.

Na Figura 5, temos respectivamente, a forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull para alguns valores (mostrado na legenda da figura) dos parâmetros (γ, α) .

Figura 5 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição de Weibull.



Fonte: Autora (2024).

2.2.3 Distribuição Log-Normal

A distribuição Log-Normal é muito utilizada para descrever o tempo de vida de produtos e indivíduos, assim como a distribuição de Weibull.

Seja T uma variável aleatória referente ao tempo de falha de um indivíduo. Se T segue uma distribuição Log-Normal, sua função de densidade é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi t\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}, \quad t > 0,$$

sendo $-\infty < \mu < \infty$ e $\sigma > 0$, onde μ é a média do logaritmo do tempo de falha, assim como σ é o desvio-padrão.

As funções de sobrevivência, $S(t)$, e de risco, $\lambda(t)$, não apresentam forma analítica explícita, sendo representadas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right)$$

e

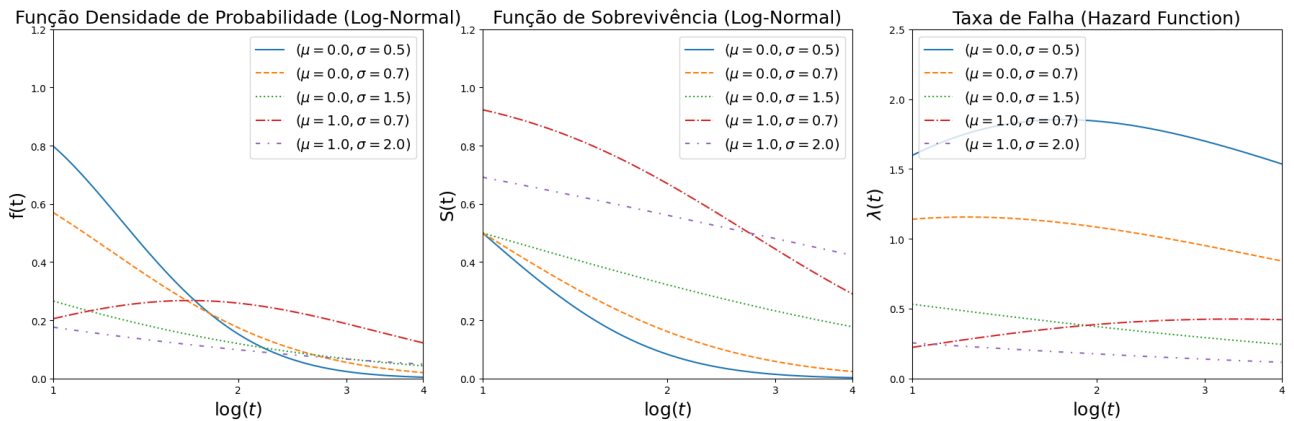
$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi t \sigma}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}}{\Phi\left(\frac{-\log(t) + \mu}{\sigma}\right)}$$

sendo $\Phi(\cdot)$ a função de distribuição acumulada de uma normal padrão.

Diferentemente da distribuição de Weibull, essa distribuição contempla funções de risco unimodais.

Na Figura 6, temos respectivamente, a forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Normal para alguns valores (mostrado na legenda da figura) dos parâmetros (μ, σ) .

Figura 6 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Normal.



Fonte: Autora (2024).

2.2.4 Distribuição Gama

A distribuição Gama é também inclui a exponencial como um caso particular. Sua função de densidade é caracterizada por dois parâmetros, k e α , em que $k > 0$ é chamado parâmetro de forma e $\alpha > 0$ de escala, é expressa por:

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t > 0,$$

com $\Gamma(k)$ a função Gama definida por:

$$\Gamma(k) = \int_0^{\infty} x^{k-1} \exp\{-x\} dx$$

Para $k > 1$, esta função de densidade apresenta um único pico em $t = (k - 1)/\alpha$. A respectiva função de sobrevivência desta distribuição é dada por:

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du$$

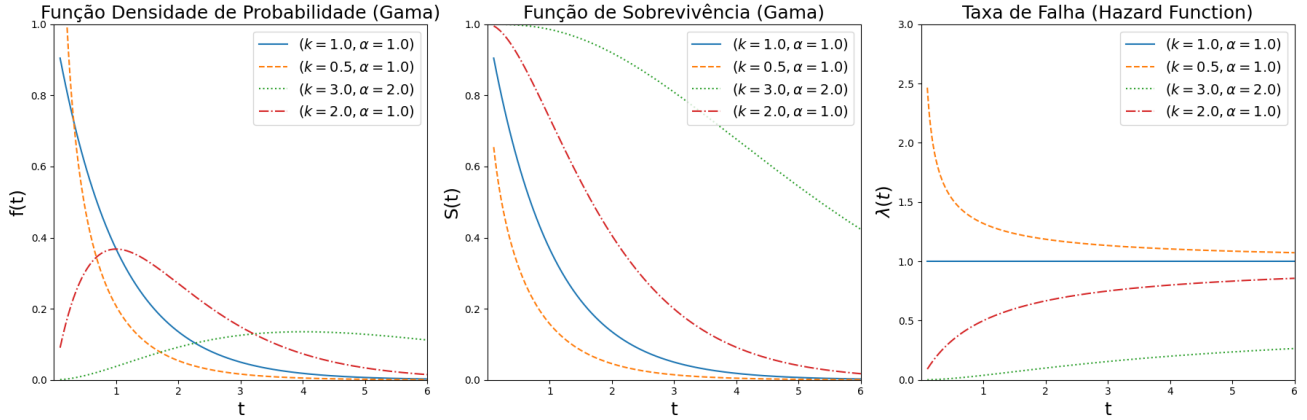
e a função de taxa de falha $\lambda(t)$:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}}{\int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du}$$

A função taxa de falha apresenta um padrão crescente ou decrescente convergindo, no entanto, para um valor constante quando t cresce de 0 a infinito.

Na Figura 7, temos respectivamente, a forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição gama para alguns valores (mostrado na legenda da figura) dos parâmetros (k, α) .

Figura 7 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Gama.



Fonte: Autora (2024).

2.2.5 Distribuição Log-Logística

Seja T , uma variável aleatória com distribuição Log-Logística, a função densidade é expressa por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left(1 + (t/\alpha)^\gamma\right)^{-2}, \quad t > 0,$$

sendo $\alpha > 0$ o parâmetro de escala e $\gamma > 0$ o parâmetro de forma. As funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$ são expressas, respectivamente, por:

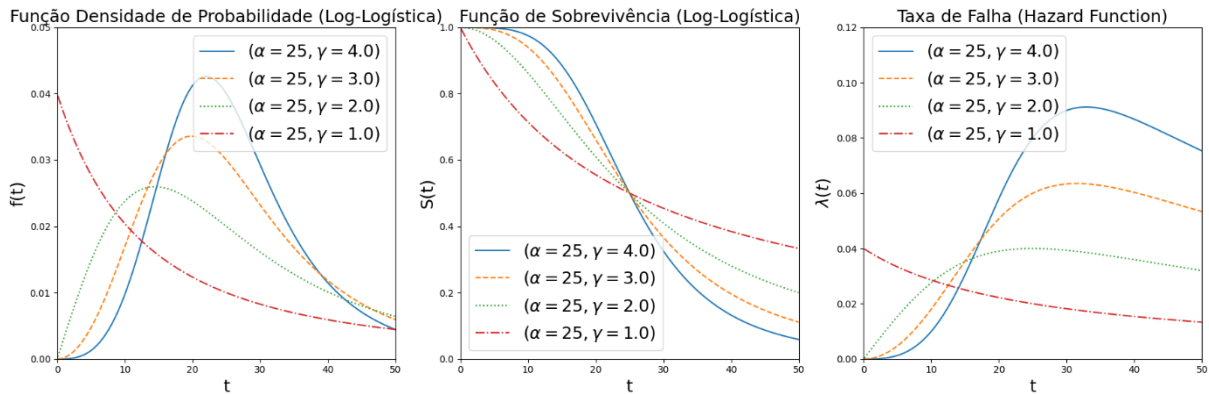
$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}$$

e

$$\lambda(t) = \frac{\gamma (t/\alpha)^{\gamma-1}}{\alpha [1 + (t/\alpha)^\gamma]}$$

Na figura 8, temos respectivamente, a forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Logística para alguns valores (mostrado na legenda da figura) dos parâmetros (α, γ) .

Figura 8 - Funções de densidade de probabilidade, de sobrevivência e de taxa de falha da distribuição Log-Logística.



Fonte: Autora (2024).

2.3 Modelos de Regressão Paramétricos

No estudo de dados de sobrevivência, é comum encontrar situações em que existem variáveis associadas ao tempo de vida, chamadas de covariáveis. Essas covariáveis podem representar tanto a heterogeneidade dos dados quanto as diferenças no tratamento recebido pelos indivíduos.

Uma forma de estudar o efeito das covariáveis no tempo de sobrevivência é reparametrizando um dos parâmetros do modelo para incluir a covariável e assim avaliar sua influência.

Outra abordagem é utilizar duas classes de modelos de regressão: os modelos paramétricos, ou modelos de tempo de falha acelerado, e os semiparamétricos, conhecidos como modelos de riscos proporcionais ou modelos de regressão de Cox.

No presente trabalho será mostrado apenas os modelos paramétricos.

2.3.1 Modelo de Regressão Exponencial

O modelo de regressão Exponencial é o modelo de regressão mais simples, com uma única covariável. Segundo Colosimo e Giolo (2006), a combinação de um componente determinístico e uma distribuição Exponencial com média unitária para o erro ($f(\varepsilon) = \exp\{-\varepsilon\}$) produz o modelo de regressão Exponencial:

$$T = \exp\{\beta_0 + \beta_1 x\} \varepsilon$$

Este modelo é linearizável se for considerado o logaritmo de T , escrito da forma:

$$Y = \log(T) = \beta_0 + \beta_1 X + v$$

onde $v = \log(\varepsilon)$. O erro v segue uma distribuição do valor extremo padrão ($f(v) = \exp\{v - \exp\{v\}\}$). A função de sobrevivência para Y condicional a x é expressa por:

$$S(y|x) = \exp\{-\exp\{y - (\beta_0 + \beta_1 x)\}\}$$

Para T condicional a x , a função de sobrevivência correspondente é:

$$S(t|x) = \exp\left\{-\left(\frac{t}{\exp\{\beta_0 + \beta_1 x\}}\right)\right\}$$

2.3.2 Modelo de Regressão Weibull

O modelo de regressão Weibull, de acordo com David e Kleibaum (2012), é o modelo paramétrico mais utilizado. Assim como no modelo exponencial, será reparametrizado com coeficientes de regressão. O parâmetro adicional γ é chamado de parâmetro de forma e determina a forma da função de risco. A adição deste dá ao modelo de Weibull maior flexibilidade do que o modelo exponencial. Este modelo pode ser expresso como:

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \sigma v = \mathbf{x}'\boldsymbol{\beta} + \sigma v$$

em que $\mathbf{x}' = (1, x_1, \dots, x_p)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$. Neste modelo, T deve ter uma distribuição de Weibull para que $\log(T)$ tenha uma distribuição do valor extremo com parâmetro de escala σ .

O modelo nessa forma também é conhecido como modelo de tempo de vida acelerado. Isto porque a função das covariáveis é acelerar ou desacelerar o tempo de vida. Ele pode ser expresso como:

$$T = \exp\{\mathbf{x}'\boldsymbol{\beta}\}\exp\{\sigma v\}$$

A função de sobrevivência para Y condicional a $\mathbf{x}$ é expressa por:

$$S(y|\mathbf{x}) = \exp \left\{ -\exp \left\{ \frac{y - \mathbf{x}'\boldsymbol{\beta}}{\sigma} \right\} \right\}$$

e para T condicional a $\mathbf{x}$:

$$S(t|\mathbf{x}) = \exp \left\{ -\left(\frac{t}{\exp\{\mathbf{x}'\boldsymbol{\beta}\}} \right)^{1/\sigma} \right\}$$

2.3.3 Modelo de Regressão Log-Normal

De acordo com David e Kleibaum (2012), o modelo Log-Normal possui uma função de risco e uma função de sobrevivência relativamente complicadas, que só podem ser expressas em termos de integrais. A forma da distribuição Log-Normal é muito semelhante à da distribuição Log-Logística e produz resultados de modelo similares. Uma diferença é que, embora o modelo log-Normal acomode um modelo de tempo de falha acelerado, ele não é um modelo de chances proporcionais. As distribuições Weibull e Log-Normal são casos especiais da distribuição Gama Generalizada.

2.3.4 Modelo de Regressão Logístico

Seja $Y = \log(T)$, $\alpha = \exp(\mathbf{x}^T \boldsymbol{\beta})$ e $\gamma = \frac{1}{\sigma}$. Onde T segue uma distribuição Log-Logística, as funções de densidade de probabilidade e de sobrevivência de Y são definidas da seguinte maneira:

$$f(y) = \frac{1}{\sigma} \exp \left\{ \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right\} \left[1 + \exp \left\{ \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right\} \right]^{-2}$$

e

$$S(y) = \left[1 + \exp \left\{ \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right\} \right]^{-1}$$

O modelo de Regressão Logístico, também conhecido como modelo de tempo de vida acelerado, tem forma de locação-escada da mesma maneira que v segue uma distribuição Logística padrão. Da mesma maneira que o modelo Weibull, considera-se $Y = \mathbf{x}'\boldsymbol{\beta} + \sigma v$, dessa forma, tem-se que as funções de densidade de probabilidade e sobrevivência de v são definidas como:

$$f_0(v) = \frac{e^v}{(1 + e^v)^2}$$

e

$$S_0(v) = \frac{1}{(1 + e^v)}$$

2.4 Estimação de parâmetros

Na literatura há diversos métodos de estimação, podemos citar dois deles: o Método de Mínimos Quadrados e o Método de Máxima Verossimilhança. Para estimar pelo primeiro método, deve-se ter como pressuposto a normalidade dos erros. Em casos que há censuras e não temos normalidade dos erros, o Método de Mínimos Quadrados se torna inadequado, visto que não é possível incorporar a censura na função que deve ser maximizada.

Por sua vez, o Método da Máxima Verossimilhança, cujo objetivo é encontrar o valor do parâmetro que maximiza a probabilidade da amostra observada ocorrer, se mostra adequado, pois permite considerar a função de sobrevivência e contribuição das censuras na função de verossimilhança. A função de densidade delas corresponde aos tempos de falha.

Independente do mecanismo de censura à direita adotado, a expressão para a função de verossimilhança é a mesma, dada por (Colosimo e Giolo, 2006):

$$L(\theta) \propto \prod_{i=1}^n [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i}$$

sendo δ_i a variável indicadora de censura.

Ao linearizar a função de verossimilhança aplicando o logaritmo tem-se:

$$l(\theta) = \sum_{i=1}^n [\delta_i \log[f(t_i; \theta)] + (1 - \delta_i) \log[S(t_i; \theta)]] + c$$

em que c é uma constante que não depende de θ .

Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$, ou equivalentemente $l(\theta)$ e são obtidos resolvendo o sistema de equações:

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

2.5 Análise de Resíduos

A avaliação do ajuste do modelo é extremamente importante na análise de dados. O principal objetivo desse passo é examinar a adequação da distribuição escolhida para a variável resposta, verificar as suposições básicas do modelo, identificar pontos extremos, observar a relevância de possíveis fatores omitidos e analisar a forma funcional do modelo.

De acordo com Klein e Moeschberger (2003), essas técnicas devem ser aplicadas para rejeitar modelos claramente inadequados, em vez de “provar” que um modelo específico está correto.

Devido à presença de observações censuradas, uma maneira de se fazer a análise é utilizando os resíduos de Cox-Snell (1968), útil para examinar o ajuste global do modelo.

2.5.1 Resíduos de Cox-Snell

Os resíduos de Cox-Snell (1968) auxiliam no ajuste global do modelo e são quantidades determinadas por:

$$\hat{e}_i = \hat{H}(t_i | \mathbf{x}_i) \text{ ou } \hat{e}_i = \hat{H}(y_i | \mathbf{x}_i)$$

sendo que $\hat{H}(\cdot)$ é a função de taxa de falha acumulada obtida do modelo ajustado, t_i os tempos de falha dos indivíduos e y_i a variável originada a partir da transformação $Y = \log(T)$. Para os modelos de regressão Exponencial, Weibull e Log-Normal, os resíduos de Cox-Snell são dados, respectivamente, por:

$$\text{Exponencial: } \hat{e}_i = [t_i \exp\{-\mathbf{x}_i' \hat{\boldsymbol{\beta}}\}],$$

$$\text{Weibull: } \hat{e}_i = [t_i \exp\{-\mathbf{x}_i' \hat{\boldsymbol{\beta}}\}]^{\hat{\gamma}}$$

e

$$\text{Log-normal: } \hat{e}_i = -\log \left[1 - \phi \left(\frac{\log(t_i) - \mathbf{x}_i' \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \right) \right].$$

Segundo Lawless (2003), os resíduos $\hat{e}_i$ vem de uma população homogênea e devem seguir uma distribuição Exponencial padrão se o modelo for adequado. Em outras palavras, o gráfico $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$ deve ser aproximadamente uma reta com inclinação 1, uma vez que $\hat{H}(\hat{e}_i) = -\log \hat{S}(\hat{e}_i)$. Onde $\hat{S}(\hat{e}_i)$ é a função de sobrevivência dos $\hat{e}_i$'s obtida pelo estimador de Kaplan-Meier (KM). O gráfico das curvas de sobrevivência obtidas a partir do estimador KM também auxilia na verificação da qualidade do modelo ajustado. Quando mais próximas elas se apresentarem, melhor é considerado o ajuste do modelo aos dados.

2.6 Escolha do Modelo

Em problemas que envolvem incerteza, utilizar critérios que auxiliem na comparação e seleção de modelos paramétricos é de fundamental importância em uma análise de dados.

Neste tópico será discutido o critério de comparação de modelos: o AIC, método dos gráficos e o teste de hipóteses.

2.6.1 Métrica AIC

Akaike (1974) propõe um método baseado na medida de Informação de Kullback-Leibler. Seja k o número de parâmetros a serem estimados, n o número de

observações de x e $\hat{\theta}$ uma estimativa de θ , o critério de informação de Akaike (AIC) é obtido por

$$AIC = -2 \log \left(L(\hat{\theta}; x) \right) + 2k$$

Dado um conjunto de modelos para x , ajustados os dados, o melhor será o que apresentar o menor AIC. Além de selecionar um ótimo ajuste, o critério penaliza a adição de parâmetros, desencorajando *overfitting*.

2.6.2 Método dos gráficos

O primeiro método gráfico consiste na comparação da função de sobrevivência do modelo proposto com o estimador de Kaplan-Meier. Neste procedimento ajustam-se os modelos propostos aos conjuntos de dados e, a partir das estimativas dos parâmetros de cada modelo, estimam-se suas respectivas funções de sobrevivência. Em seguida comparam-se as funções de sobrevivência estimadas para cada modelo proposto com $\hat{S}(t)$. O modelo será adequado quando aquele em que sua curva de sobrevivência mais se aproximar daquela do estimador de Kaplan-Meier, isto é, aquele modelo cujos pontos da função de sobrevivência estimada estiverem mais próximo dos valores obtidos pelo estimador KM será o melhor.

O segundo método consiste na linearização da função de sobrevivência tendo como ideia básica a construção de gráficos que sejam aproximadamente lineares em casos que o modelo é ajustado adequadamente. Em seguida, deve-se comparar o estimador de KM com o ajuste do modelo proposto.

2.6.3 Teste de Hipóteses

Outro método é utilizar teste de hipóteses. As hipóteses a serem testadas são

$$\begin{cases} H_0: & \text{O modelo de interesse é adequado} \\ H_a: & \text{O modelo de interesse não é adequado} \end{cases}$$

Para testar é utilizado a estatística da razão de verossimilhanças em modelos encaixados (Cox e Hinkley, 1974). Significa dizer que se deve identificar um modelo generalizado tal que os modelos de interesse sejam casos particulares.

O teste é realizado a partir dos seguintes ajustes:

1. Modelo generalizado e obtenção do valor do logaritmo de sua função de verossimilhança $\log L(\widehat{\theta}_G)$;
2. Modelo de interesse e obtenção do valor do logaritmo de sua função de verossimilhança $\log L(\widehat{\theta}_M)$.

Assim, é possível calcular a estatística da razão de verossimilhança, a partir de:

$$TRV = -2 \log \left[\frac{L(\widehat{\theta}_M)}{L(\widehat{\theta}_G)} \right] = 2 [\log L(\widehat{\theta}_G) - \log L(\widehat{\theta}_M)]$$

que sob H_0 tem aproximadamente uma distribuição qui-quadrado com número de graus de liberdade igual a diferença do número de parâmetro dos modelos $(\widehat{\theta}_G, \widehat{\theta}_M)$ sendo comparados.

A decisão é por meio da comparação do valor de TRV com o valor crítico da distribuição qui-quadrado para o nível de significância desejado, se TRV for maior que o valor crítico, rejeita-se a H_0 indicando que o modelo $L(\widehat{\theta}_G)$ proporciona um ajuste melhor aos dados do que o modelo $L(\widehat{\theta}_M)$.

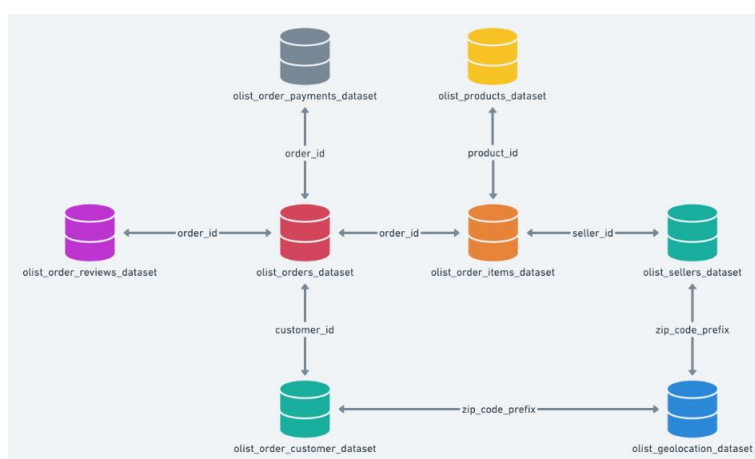
3 Aplicação dos dados

Nesta seção o intuito é mostrar como foi feito o tratamento da base de dados disponibilizada pela Olist no *Kaggle*.

3.1 Tratamento da base

Os dados estão divididos em 8 base de dados que se integram a partir de variáveis chave, isto é, variáveis que estão presentes em duas ou mais bases. O intuito inicial é unir as tabelas utilizando a biblioteca *pandas* do *Python*. O esquema abaixo mostra como estão organizadas essas bases.

Figura 9 - Esquema de organização das bases.



Fonte: *Kaggle* (2022).

A união das bases foi feita com um LEFT JOIN que incluem todos os registros da primeira de duas tabelas (a da esquerda), mesmo se não houver valores correspondentes na segunda tabela (à direita). Essa união resultou em 118.434 registros com 34 colunas repetindo apenas as colunas que foram a chave para a união.

A primeira etapa foi excluir essas repetidas, em seguida fizemos alguns filtros (a explicação da variável está nos apêndices):

- Utilizando a coluna “customer_state” filtramos para apenas os dados do estado de São Paulo;

- A partir da variável “order_status” filtramos para ter o status apenas “delivered”, pois o que nos importa é apenas as compras feitas que foram entregues;
- Filtramos “order_id” para termos apenas registros únicos, pois futuramente poderia dar erro na contagem de pessoas que fizeram compras mais de uma vez.
- Excluimos colunas desnecessárias para o estudo, entre elas são:

'shipping_limit_date','customer_state','order_estimated_delivery_date','order_delivered_customer_date','order_delivered_carrier_date','order_approved_at','order_status','product_name_length','shipping_limit_date','product_description_length','product_photos_qty','product_weight_g','product_length_cm','product_height_cm','order_item_id','payment_sequential','payment_type','payment_installments','product_width_cm','customer_zip_code_prefix','customer_id'.

Aplicados esses filtros ficamos com uma base com 40.501 registros e 10 colunas. A segunda etapa foi a de criação de variáveis:

- “data_compra” que contém a data que o consumidor fez a compra, a partir da variável “order_purchase_timestamp”;
- “data_fechamento” que contém a data da última compra feita na base, essa variável será importante pois será a data considerada de fechamento. Importante considerar que esse cliente que fez a última compra não foi considerado no estudo;
- “numero_compras” que contém a quantidade que cada cliente comprou.

A fim de entender os clientes que compraram **apenas uma vez** criamos uma base com esses registros totalizando 39.156 dados e criamos a variável tempo que é a diferença entre a data da compra do cliente e a data de fechamento da base, acrescentamos a variável “censura”, como essa base contém apenas a censura do estudo, ela é completada com “0”. A partir da variável “customer_unique_id” selecionamos aqueles registros duplicados, isto é, clientes que compraram **mais de uma vez** com 2558 dados. A variável “tempo”, nesse caso, foi calculada como a diferença entre a penúltima e última compra desse consumidor e a variável “censura” foi preenchida com “1” com isso a base ficou com 1213 registros.

Como na base dos clientes que realizaram uma única compra obtivemos um grande volume de dados, teríamos um percentual de censura muito maior o que atrapalharia nosso estudo já que o percentual de censura interfere no vício do estimados. Por conta desse fato, utilizamos uma estratégia estatística para contornar

esse problema: amostramos 20% da base “uma vez” a mais do valor da base “mais vezes” com intuito de ficar com 20% de censura. Feito isso juntamos a base uma vez com a base mais vezes ficando com 1456 registros, em seguida fizemos o último filtro para não atrapalhar nos resultados do modelo eliminando os casos em que o “tempo” era igual a zero, isto é, os casos em que as pessoas fizeram mais de uma compra, porém foi no mesmo dia e as pessoas que compraram uma vez e foi no mesmo dia do fechamento da base, nessa última ficamos com 1132 dados, essa nova base será utilizada para a criação dos modelos estatísticos.

Neste trabalho usamos quatro covariáveis para o modelo: “valor frete”, “preço do produto”, “valor total compra”, o “número de compras” e o “tipo do produto” essas foram padronizadas com os dados da última compra. Nessa última, separamos os tipos do produto em 4 categorias: Produtos para o lar, Produtos de consumo e entretenimento, produtos eletrônicos e tecnologia e outros. Na Tabela 1, temos como ficou disposto cada tipo de produto em cada categoria.

Tabela 1 - Relação categoria x tipo produto da base.

Produtos para o Lar	Produtos de Consumo e Entretenimento	Produtos Eletrônicos e Tecnologia	Outros
eletrodomesticos	beleza_saude	informatica_acessorios	industria_comercio_e_negocios
utilidades_domesticas	instrumentos_musica	telefonica	agro_industria_e_comercio
cama_mesa_banho	relogios_presentes	tablets_impressao_imagem	market_place
moveis_decoracao	esporte_lazer	telefonica_fixa	None
climatizacao	brinquedos	eletroportateis	
construcao_ferramentas_construcao	perfumaria	pcs	
ferramentas_jardim	fashion_bolsas_e_acessorios		
moveis_sala	consoles_games		
moveis_escritorio	papelaria		
casa_construcao	cool_stuff		
moveis_colchao_e_estofado	eletronicos		
construcao_ferramentas_seguranca	livros_interesse_geral		
moveis_quarto	alimentos_bebidas		
sinalizacao_e_seguranca	alimentos		
moveis_cozinha_area_de_servico_jantar_e_jardim	bebidas		
eletrodomesticos_2	audio		
casa_conforto	malas_acessorios		
construcao_ferramentas_iluminacao	artes_e_artesanato		
construcao_ferramentas_ferramentas	livros_importados		
construcao_ferramentas_jardim	fashion_roupa_masculina		

Fonte: Autora (2024).

Na covariável “tipo do produto” por ser categórica, foi preciso transformar em variáveis dummies.

A variável "número de compras" (nc) foi transformada em uma variável binária, indicando se o número de compras é menor ou igual a 2 (valor 1) ou maior que 2 (valor 0).

4 Resultados

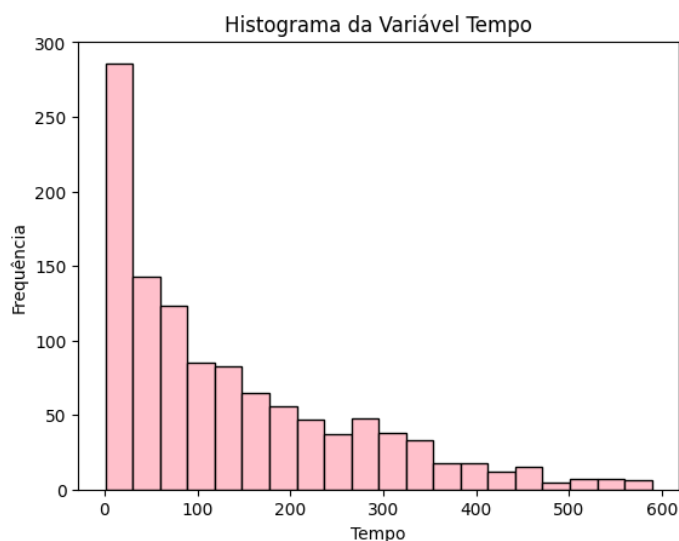
Nesta seção serão apresentados os resultados do estudo desse trabalho.

4.1 Análise Descritiva

Para compreender melhor o comportamento da variável tempo, foi realizada uma análise por meio de um gráfico.

Na Figura 10, temos o histograma da variável tempo. Pode-se notar uma tendência de uma distribuição Exponencial e temos uma assimetria à direita. Vemos que a maioria dos dados está concentrada em valores baixos de tempo, sugerindo que o tempo que as pessoas voltam a comprar é curto.

Figura 10 - Histograma da variável Tempo.



Fonte: Autora (2024).

A fim de compreender as covariáveis numéricas do estudo, fizemos algumas análises estatísticas básicas como mostra na Tabela 2 das covariáveis: preço, valor do frete e o valor total da compra.

Tabela 2 - Estatísticas das covariáveis numéricas.

	Preço	Valor do frete	Valor total da compra
Média	108,02	14,96	134,99
Desvio Padrão	143,38	9,84	152,92
Mínimo	4,99	0,10	0,69
25%	39,90	9,26	54,00
50%	69,00	13,00	93,11
75%	124,90	16,11	161,78
Máximo	1.700	87,74	1.788

Fonte: Autora (2024).

A média do preço é de R\$ 108,02 o que indica que a maioria dos produtos vendidos tem um preço próximo a R\$ 108,02. No entanto, é importante considerar a dispersão dos dados, tendo em vista que o desvio padrão é de R\$ 143,38, isso significa dizer que os preços dos produtos variam consideravelmente, com alguns produtos sendo muito mais caros do que outros. Temos que preço mínimo observado é de R\$ 4,99, indicando que é possível encontrar produtos com preços bem abaixo da média. O preço máximo observado é de R\$ 1.700, indicando assim que existem produtos com preços acima da média.

A média do valor do frete é de R\$ 14,96 que é o valor médio do frete relativamente baixo, o que pode ser um fator importante para os consumidores na hora de decidir onde comprar, o desvio padrão é de R\$ 9,84, o que pode significar que os dados do valor do frete são menores do que os do preço dos produtos. Assim podemos dizer que os valores do frete são mais homogêneos, com menos variação entre um produto e outro.

A média do valor total da compra é de R\$ 134,99 que representa o custo médio total de uma compra, incluindo o preço do produto e o valor do frete.

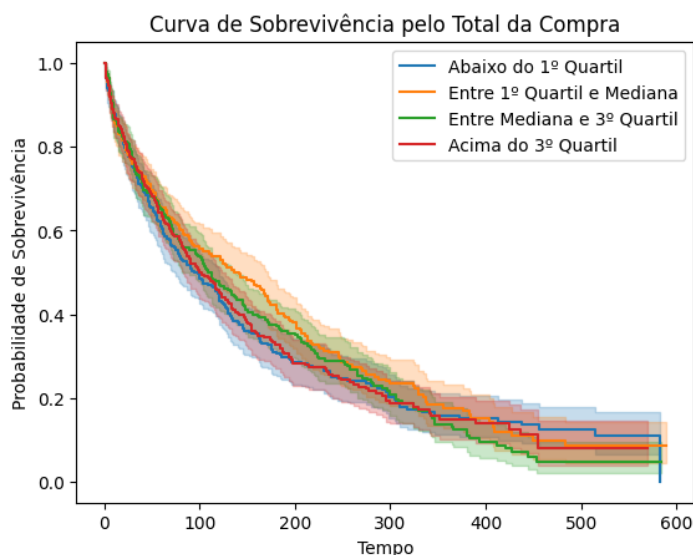
Temos o desvio padrão de R\$ 152,92 que é maior do que a do valor do frete, mas menor do que a do preço dos produtos. Isso significa que o valor total da compra é mais variável do que o valor do frete, mas menos variável do que o preço dos produtos. Isso pode indicar que há promoções que são aplicadas no final da compra, o que acaba fazendo os preços variarem.

Foi feita uma análise descritiva utilizando o estimador não paramétrico de Kaplan-Meier para termos uma noção de como as covariáveis que foram selecionadas para o modelo estão se comportando.

Na Figura 11, temos a curva de sobrevivência categorizada por quartis da covariável total da compra. A curva azul “Abaixo do 1º Quartil”, representa os clientes que gastaram menos do que 25% do total das compras. A curva laranja “Entre o 1º Quartil e a Mediana”, representa os clientes que gastaram entre 25% e 50% do total das compras. A curva verde “Entre a Mediana e o 3º Quartil”, representa os clientes que gastaram entre 50% e 75% do total das compras. A vermelha “Acima do 3º Quartil”, representa os clientes que gastaram mais do que 75% do total das compras.

Podemos observar que há uma tendência onde os clientes que gastam mais (curvas verde e vermelha) têm uma maior probabilidade de sobreviver ao longo do tempo, isto é, voltar a comprar, comparados aos clientes que gastam menos (curvas azul e laranja). Em outras palavras, os clientes que compram mais frequentemente tendem a fazer compras de maior valor.

Figura 11 - Curva de Sobrevivência categorizada por quartis do total da compra.



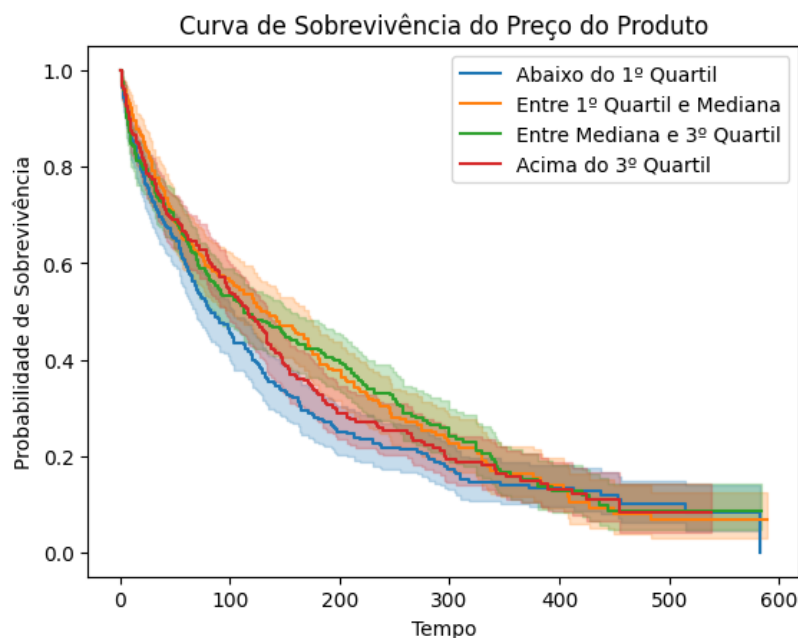
Fonte: Autora (2024).

Na Figura 12 é apresentado a curva de sobrevivência categorizada por quartis do preço do produto. A legenda é a mesma da Figura 11.

Pode-se notar que para a covariável preço, temos a mesma tendência onde produtos de maior preço (curvas verde e vermelha) têm uma maior probabilidade de

sobrevivência ao longo do tempo, comparados aos produtos de menor preço (curvas azul e laranja).

Figura 12 - Curva de Sobrevivência categorizada por quartis do preço do produto.

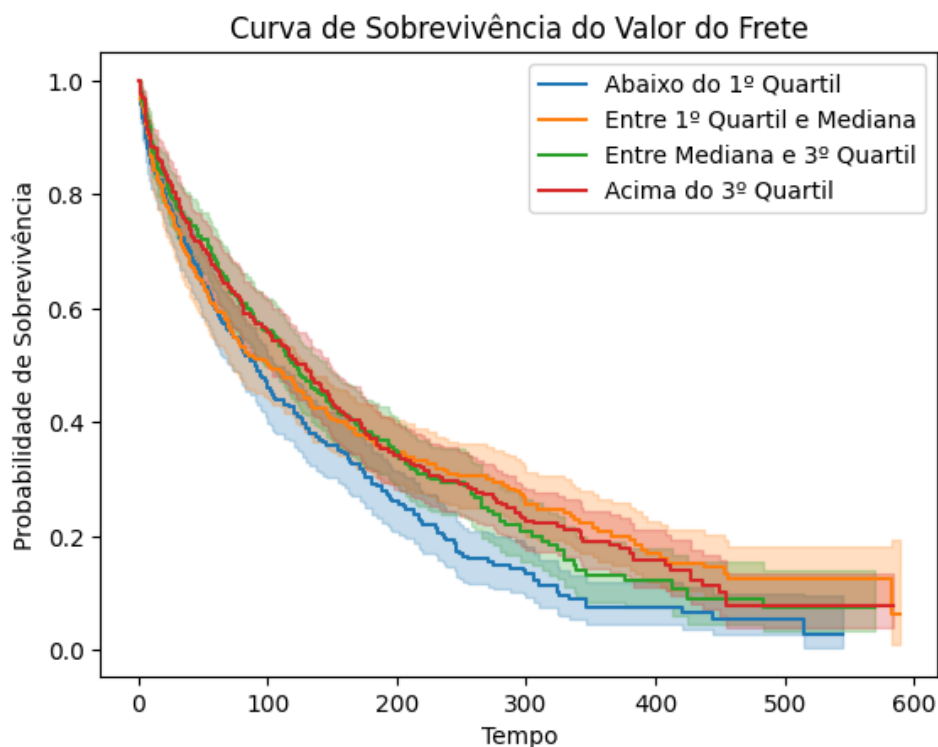


Fonte: Autora (2024).

Na Figura 13, temos a curva de sobrevivência categorizada por quartis do valor do frete. A legenda é a mesma das figuras anteriores.

Neste caso, vemos que a curva azul tem queda mais rápida na probabilidade de sobrevivência, indicando que fretes de menor valor tendem a ter uma menor sobrevivência ao longo do tempo.

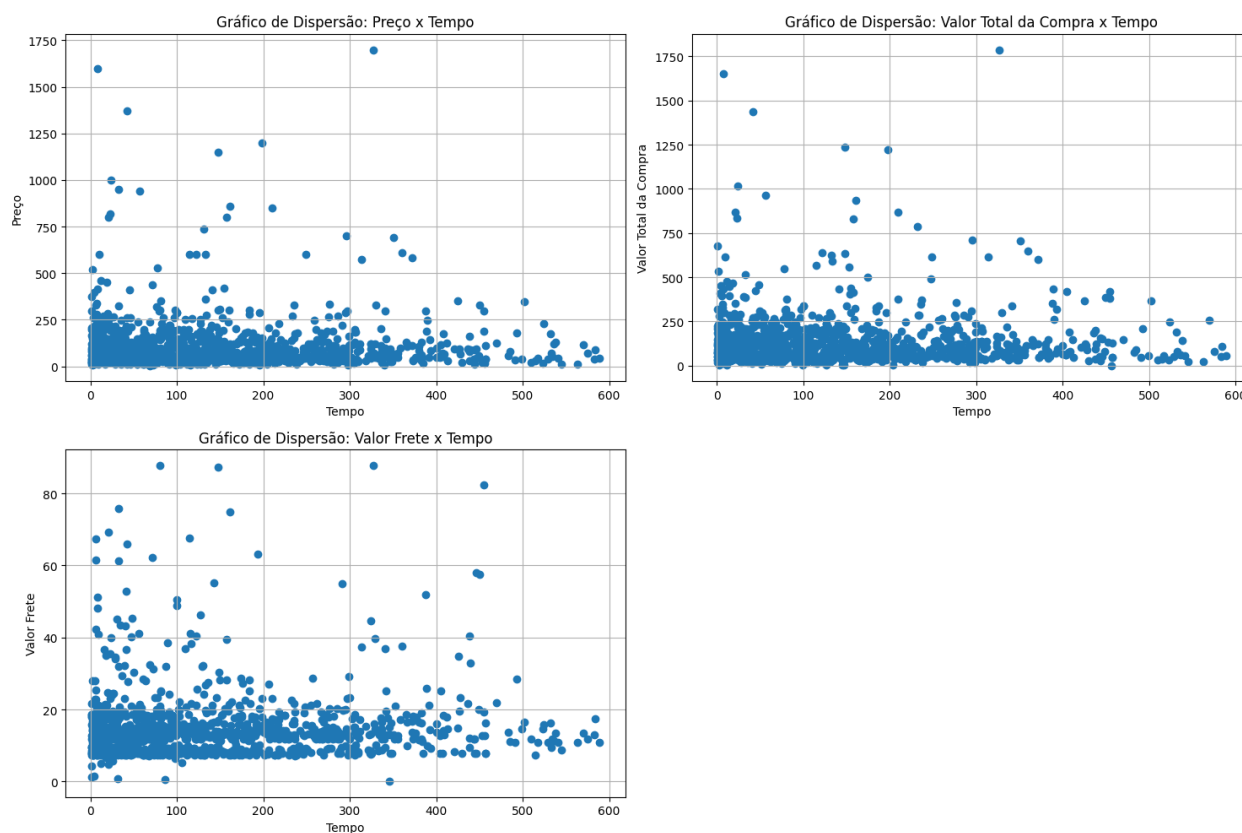
Figura 13 - Curva de Sobrevivência categorizada por quartis do valor do frete.



Fonte: Autora (2024).

Além dessa análise verificamos que há uma correlação moderada entre “preço” e “valor_frete” que é aproximadamente 0,52, há uma correlação muito forte de 0,92 entre “preço” e “valor_total_compra”. E uma correlação moderada entre “valor_frete” e “valor_total_compra” que é de aproximadamente 0,52.

Na Figura 14, temos os gráficos de dispersão das covariáveis preço, total da compra e frete *versus* tempo.

Figura 14 - Gráficos de dispersão das covariáveis numéricas.

Fonte: Autora (2024).

Observa-se que a dispersão dos pontos no gráfico de frete é mais ampla em comparação com os gráficos de preço e valor total da compra, sugerindo uma relação mais fraca e menos linear entre frete e tempo. A nuvem de pontos para o valor total da compra exibe uma forma levemente elíptica com maior concentração no centro, indicando uma relação moderadamente linear com o tempo. Da mesma forma, a nuvem de pontos para o preço também apresenta uma forma elíptica centralizada, evidenciando uma relação moderadamente linear entre preço e tempo.

Foi realizada também uma análise da covariável “tipo do produto”, apresentada na Tabela 3, que é categórica.

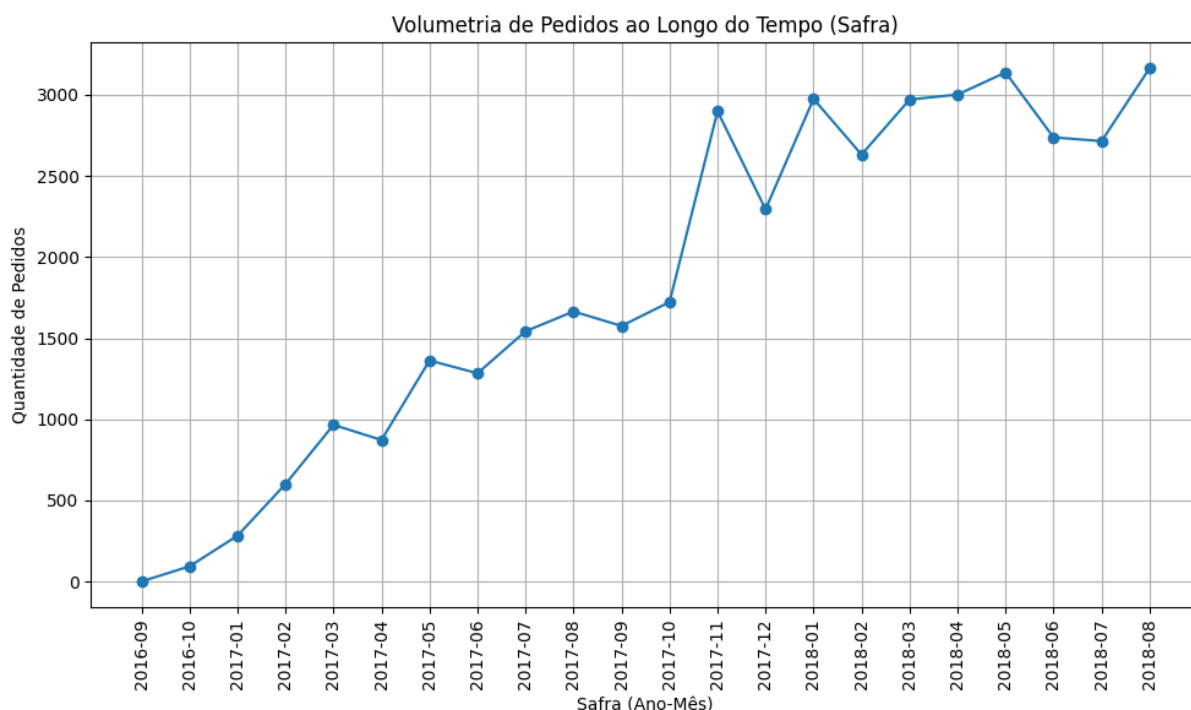
Tabela 3 - Análise da variável categórica Tipo do Produto.

Tipo do produto	Quantidade	(%)
Produtos de Consumo e Entretenimento	555	49,03
Produtos para o Lar	408	36,04
Produtos Eletrônicos e Tecnologia	111	9,80
Outros	58	5,13

Fonte: Autora (2024).

Observou-se que os produtos de consumo e entretenimento foram os mais vendidos na nossa base de estudo, totalizando 555 itens. Em seguida, destacaram-se os produtos para o lar, com 408 itens vendidos.

A Figura 15 mostra a volumetria de pedidos ao longo do tempo (safras mensais), de setembro de 2016 a agosto de 2018.

Figura 15 - Volumetria de pedidos ao longo das safras.

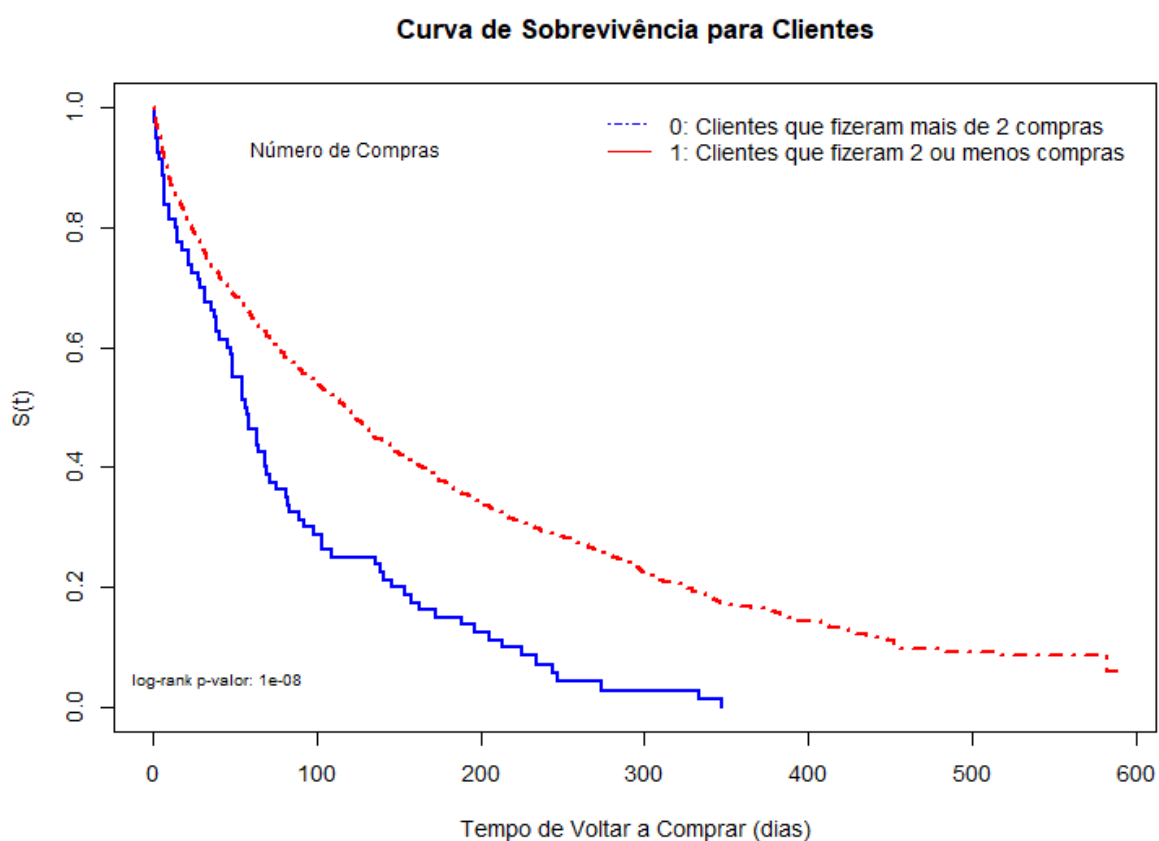
Fonte: Autora (2024).

Observando as tendências ao longo desse período, alguns pontos de sazonalidade e padrão podem ser destacados:

- **Tendência de Crescimento Geral:** De setembro de 2016 até o início de 2018, há uma clara tendência de aumento no volume de pedidos;
- **Sazonalidade:** Algumas flutuações podem refletir sazonalidades anuais, como promoções de final de ano (exemplo: aumento em outubro-novembro de 2017). Esses picos podem estar associados a eventos sazonais, como Black Friday, período em que o consumo geralmente aumenta.

A seguir na Figura 16, é apresentado a curva de sobrevivência para a covariável número de compras categorizada (nc).

Figura 16 - Curva de sobrevivência para a covariável nc.



Fonte: Autora (2024).

A partir desse gráfico podemos concluir que:

- Os clientes com menos de 2 compras (curva vermelha) têm uma taxa de retorno mais lenta, mas mantêm-se ativos por um período mais longo, mostrando uma possível retenção no longo prazo.
- A curva azul ($nc=0$), que desce mais rapidamente, indica que clientes com mais compras (2 ou mais) tendem a retornar para novas compras mais rapidamente.

O teste log-rank com hipóteses:

$$\begin{cases} H_0: S_1(t) = S_2(t) \\ H_a: S_1(t) \neq S_2(t) \end{cases}$$

Onde t é o tempo do estudo em dias.

Como $pvalor = 8e - 09$ um valor bem pequeno indica que podemos rejeitar a hipótese nula de que não há diferença na sobrevivência entre os grupos. Isso sugere que o grupo $nc = 1$ (clientes com 2 ou menos compras) apresenta um padrão de sobrevivência diferente do grupo $nc = 0$ (clientes com mais de 2 compras).

4.2 Modelagem Estatística

A partir das técnicas de análise de sobrevivência descritas anteriormente e dos dados coletados do *Kaggle*, foram construídos quatro modelos paramétricos (Exponencial, Weibull, Log-Normal e Log-Logística) utilizando o software R para avaliar o comportamento dos compradores da Olist, no período de 2016 a 2018.

4.2.1 Modelos de regressão paramétricos

Conforme apresentado na seção 2.2 distribuições de probabilidade, as estimativas das funções de sobrevivência para os modelos Exponencial, de Weibull, Log-Normal e Log-Logística são, respectivamente,

$$\begin{aligned}\widehat{S}_e(t) &= \exp\{-t/84,41\}, \\ \widehat{S}_w(t) &= \exp\{(-t/77,18)^{0,8044}\}, \\ \widehat{S}_{ln}(t) &= \phi[-\log(t) - 3,77/1,67], \\ \widehat{S}_{lg}(t) &= \frac{1}{1 + (t/48,63)^{1,05}}\end{aligned}$$

Os valores que aparecem nas expressões apresentadas são as estimativas de máxima verossimilhança dos parâmetros de cada um dos modelos.

Os valores estimados para essas funções, por exemplo, no tempo $t = 10$ dias são, respectivamente,

$$\begin{aligned}\widehat{S}_e(t) &= \exp\{-t/84,41\} = 0,888, \\ \widehat{S}_w(t) &= \exp\{(-t/77,18)^{0,8044}\} = 0,824, \\ \widehat{S}_{ln}(t) &= \phi[-\log(t) - 3,77/1,67] = 0,810, \\ \widehat{S}_{lg}(t) &= \frac{1}{1 + (t/48,63)^{1,05}} = 0,840\end{aligned}$$

Notamos que as estimativas obtidas por meio dos modelos de Weibull e Log-Normal são bem próximas. Isso não é observado para o modelo Exponencial e Log-Logística, que apresenta um valor estimado ligeiramente diferente dos obtidos para os outros dois modelos.

A Tabela 4 mostra as estimativas das funções de sobrevivência para os tempos de compra usando-se os modelos Exponencial, de Weibull, Log-Normal, Log-Logística e o Kaplan-Meier.

Tabela 4 - Estimativas da sobrevivência para os tempos de compra usando-se o estimador de Kaplan-Meier e os modelos Exponencial, Weibull, Log-Normal e Log-Logística.

Tempos	Kaplan-Meier	Exponencial	Weibull	Log-Normal	Log-Logística
3	0,950	0,965	0,929	0,945	0,949
5	0,900	0,942	0,895	0,902	0,916
6	0,850	0,931	0,880	0,882	0,900
7	0,800	0,920	0,865	0,863	0,884
8	0,750	0,910	0,851	0,844	0,869
9	0,700	0,899	0,837	0,827	0,855
10	0,650	0,888	0,824	0,810	0,840
12	0,596	0,867	0,800	0,780	0,813
15	0,542	0,837	0,765	0,738	0,775
18	0,481	0,808	0,733	0,701	0,739
19	0,421	0,798	0,723	0,690	0,728
20	0,361	0,789	0,714	0,679	0,718
22	0,301	0,771	0,695	0,658	0,697
25	0,241	0,744	0,668	0,630	0,668
28	0,181	0,718	0,643	0,604	0,641
30	0,120	0,701	0,627	0,588	0,624
40	0,060	0,623	0,555	0,520	0,551
45	0,060	0,587	0,523	0,492	0,520

Fonte: Autora (2024).

Na estimação do modelo de regressão, foram testadas como covariáveis o valor total da compra, o número de compras categorizado e o tipo de produto categorizado. As variáveis de frete e preço não foram incluídas devido à alta correlação com o valor total da compra. No entanto, entre as variáveis testadas, apenas o número de compras categorizado (nc) mostrou-se estatisticamente significativo para explicar o tempo de sobrevivência.

Na Tabela 5 mostra as estimativas dos parâmetros utilizando apenas como covariável “nc” por meio da regressão Exponencial, Weibull, Log-Normal e Log-Logística.

Tabela 5 - Estimativas para os dados de compras feitas por consumidores da Olist.

Modelo	Intercepto	Coefficiente (nc)	$\hat{\gamma}$	Loglik(model)	Chisq	p-valor
Exponencial	44,35	0,7738977	1 (fixo)	-5459,3	35,04	3,23E-09
Weibull	43,46	0,8354433	1,243	-5425,7	26,95	2,09E-07
Log-Normal	37,74	0,7666973	1,673	-5456,5	15,44	8,50E-05
Log-Logística	38,84	0,7328592	0,952	-5453,8	15,84	6,88E-05

Fonte: Autora (2024).

Observa-se a partir da Tabela 5 que o modelo Weibull é recomendado para representar esses dados, dado seu melhor ajuste e flexibilidade. A variável nc é um fator significativo em todos os modelos.

Foi calculado os tempos medianos da covariável nc, eles revelam diferenças de comportamento de recompra entre os grupos:

- Clientes com maior número de compras (nc = 0) têm um ciclo de recompra de aproximadamente 83,36 dias;
- Por outro lado, clientes com poucas compras (nc = 1) demoram mais para retornar, com tempo mediano de retorno para esse grupo é de 140,93 dias.

4.2.2 Adequação dos modelos

Nessa seção vamos verificar a adequação desses modelos por meio dos métodos comentados na seção 2.6 desse trabalho.

4.2.2.1 Métrica AIC

De acordo com AKAIKE (1974) para um conjunto de modelos para x , ajustados os dados, o melhor será o que apresentar o menor AIC. Na Tabela 6 temos o AIC para cada um dos modelos regressão estimados.

Tabela 6 - AIC dos modelos de regressão.

Modelo	AIC
Exponencial	10922,53
Weibull	10857,44
Log-Normal	10918,98
Log-logística	10913,61

Fonte: Autora (2024).

Com base nos valores de AIC, o modelo Weibull seria a escolha preferida, seguido pelo modelo Log-Logística, Log-Normal e, por último, o modelo Exponencial.

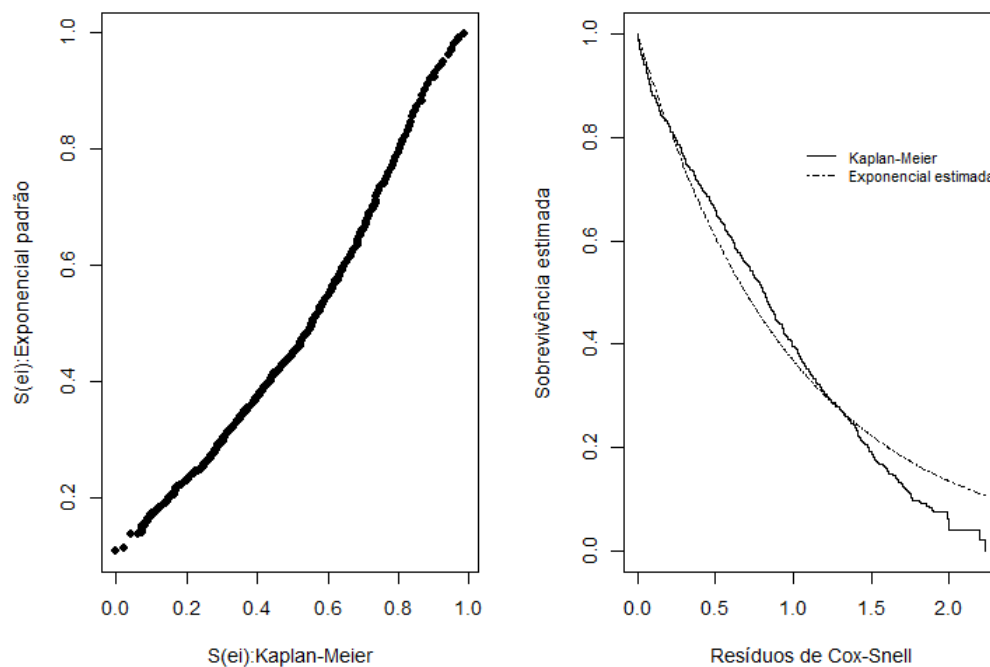
4.2.2.2 Método dos gráficos e resíduos.

Para avaliar o ajuste dos modelos de regressão Exponencial, Weibull, Log-Normal aos dados desse estudo foi utilizado os resíduos de Cox-Snell, definidos na seção 2.5 deste trabalho.

Se o modelo for adequado, esses resíduos, devem ser provenientes de uma amostra aleatória da distribuição Exponencial padrão. A Figura 17 apresenta os gráficos $(\hat{S}(\hat{e}_i)_{KM})$ (curva de sobrevivência desses resíduos obtidas por Kaplan-Meier) e $(\hat{S}(\hat{e}_i)_{exp})$ (curva modelo Exponencial padrão) bem como o gráfico dos pares de pontos $(\hat{S}(\hat{e}_i)_{KM}, \hat{S}(\hat{e}_i)_{exp})$ que devem ser aproximadamente uma reta, para que o modelo ajustado seja considerado satisfatório.

Podemos observar que ambos os gráficos mostram que a Exponencial não indica uma boa adequação.

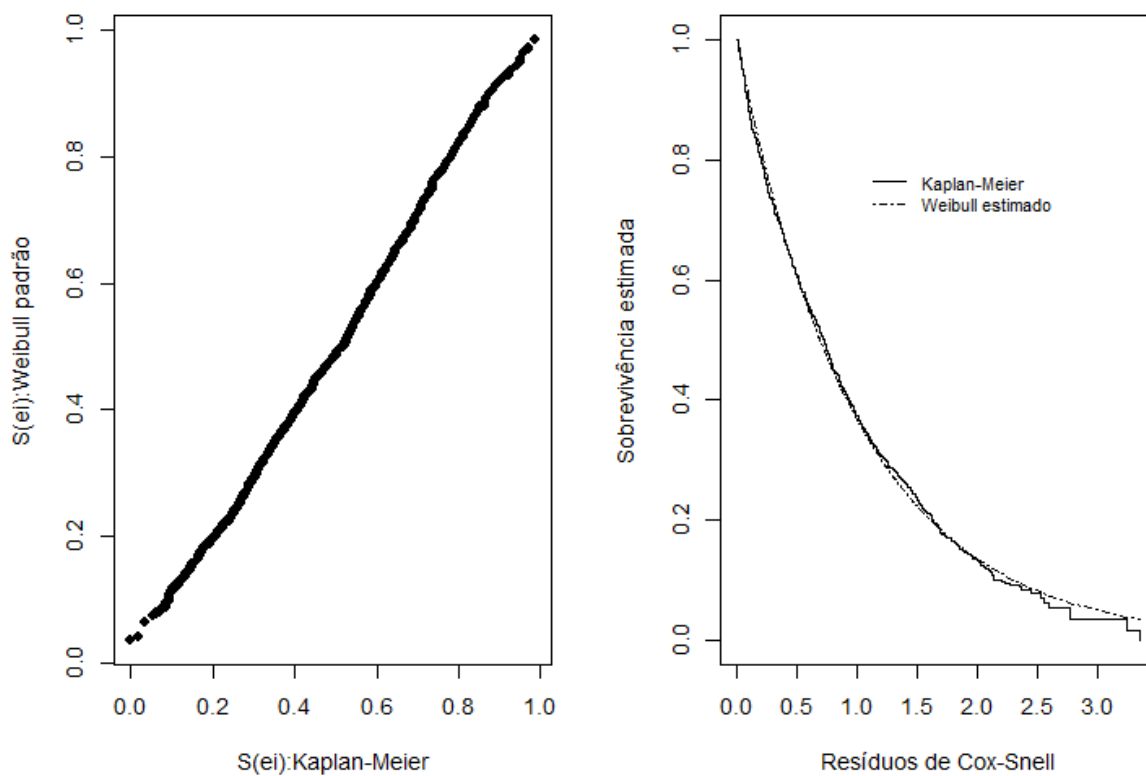
Figura 17 - Análise dos resíduos de Cox-Snell do modelo de regressão Exponencial ajustado para os dados de compradores da Olist.



Fonte: Autora (2024).

Ao contrário do modelo anterior, o modelo Weibull encontra-se bem ajustado, conforme observado na Figura 18.

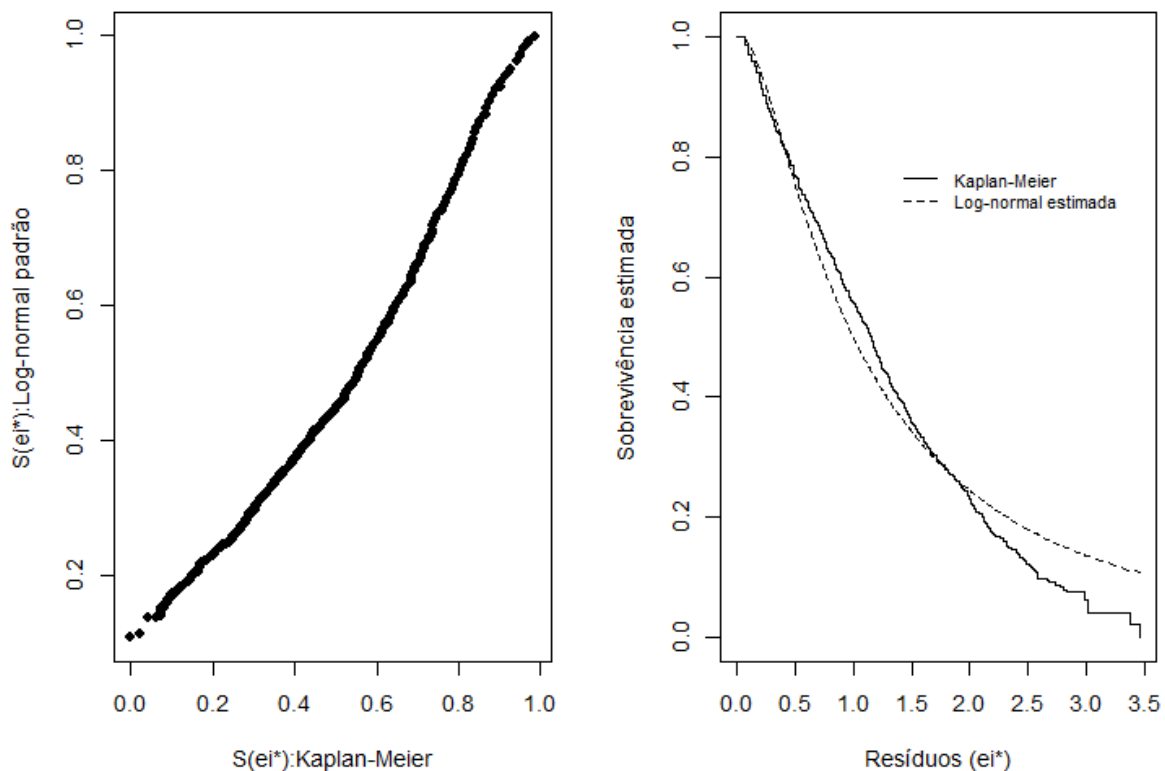
Figura 18 - Análise dos resíduos de Cox-Snell do modelo de regressão Weibull ajustado para os dados de compradores da Olist.



Fonte: Autora (2024).

A partir da Figura 19, observamos que o modelo de regressão Log-Normal não apresentou um bom ajuste quanto o modelo Weibull.

Figura 19 - Análise dos resíduos de Cox-Snell do modelo de regressão Log-Normal ajustado para os dados de compradores da Olist.



Fonte: Autora (2024).

4.2.2.3 Teste de Hipótese

As hipóteses a serem testadas são

$$\begin{cases} H_0: & \text{O modelo de interesse é adequado} \\ H_a: & \text{O modelo de interesse não é adequado} \end{cases}$$

Para testar é utilizado a estatística da razão de verossimilhanças em modelos encaixados (Cox e Hinkley, 1974). Significa dizer que se deve identificar um modelo generalizado tal que os modelos de interesse sejam casos particulares. Na Tabela 7 é apresentado os resultados dos testes do TRV.

Tabela 7 - Resultados dos testes da razão de verossimilhanças (TRV).

Comparação	Estatística da Razão de Verossimilhança	p-valor
Modelo Generalizado vs. Exponencial	6.719.181	2,57E-09
Modelo Generalizado vs. Weibull	0.1017789	0.749
Modelo Generalizado vs. Log-Normal	6.164.876	4,11E-09
Modelo Generalizado vs. Log-Logística	56.278	6,29E-08

Analisando a Tabela 7, temos para os modelos de regressão Exponencial, Weibull, Log-Normal e Log-Logística que o p-valor é muito pequeno, indicando que H_0 deve ser rejeitada, em outras palavras, o modelo Exponencial não é adequado quando comparado ao modelo Generalizado, que oferece um ajuste significativamente melhor para os dados. Já na comparação entre Modelo Generalizado e Weibull, o p-valor é alto (0.7497), indicando que não rejeitamos a hipótese nula. Sugerindo que ambos os modelos se ajustam de maneira semelhante aos dados, e não há uma preferência clara por um em relação ao outro.

Conclusões

Neste projeto, o grande desafio foi a organização da base de dados, estruturada para possibilitar a aplicação de modelos de regressão em análises de sobrevivência, seguido do problema em encontrar variáveis que explicassem a sobrevivência de compras.

Ao final deste estudo, pudemos entender um pouco sobre o padrão de compra no estado de São Paulo dos consumidores da Olist e analisamos que as pessoas que compram mais frequentemente tendem a fazer compras de maior valor, além disso seu tempo mediano de recompra é de 83 dias, enquanto pessoas que compram menos de 2 compras possui tempo de recompra de 140 dias. As compras feitas por todos os clientes da Olist seguem um padrão de crescimento geral ao longo das safras disponibilizadas e percebe-se um aumento pelos meses de outubro e novembro, normalmente fase da Black Friday no Brasil.

Falando sobre as covariáveis, o valor da compra não influencia na sobrevivência, assim como o tipo de produto que antes era uma hipótese que influenciariam. A única variável que foi significativa foi o número de compras. A região de compras estava concentrada na cidade São Paulo, seguido por Campinas e Guarulhos que são grandes metrópoles próximas à capital.

O modelo de regressão que melhor ajustou foi o de Weibull, o que faz sentido já que ele é um modelo mais flexível com o parâmetro adicional de escala permitindo modelar uma taxa de risco variável ao longo do tempo.

Para estudos futuros, seria interessante obter um conjunto de dados mais atualizado, pois o nosso inclui apenas dois meses de 2016, todos os meses de 2017 e dados de 2018 até agosto. Isso permitiria avaliar se o perfil dos compradores mudou ao longo do tempo. Além disso, contar com uma base de dados mais completa e com maior volume de informações ajudaria a enriquecer as análises, oferecendo resultados mais consistentes e detalhados.

Referências

AMAZON. **Principais diferenças entre loja física e virtual** (2021). Disponível em: <<https://venda.amazon.com.br/sellerblog/principais-diferencas-entre-loja-fisica-e-virtual>>. Acesso em: 05 outubro 2023.

COLOSIMO, E. A; GILOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006.

DAVID, G.; KLEINBAUM, M. **Survival analysis: a self-learning Text**. 3ªEd. Springer, 2012.

EUGENIO, D. R. R. **Análise de Sobrevivência Aplicada à Saúde**. Orientador: Prof Dr. Mario Hissamitsu Tarumoto. 2021. Dissertação (Mestrado) – Programa de Pós-Graduação em Matemática Aplicada e Computacional, Faculdade de Ciências e Tecnologia, UNESP, São Paulo, 2021.

FEILER, C. P. **9 em 10 brasileiros têm hábito de fazer compras ou pagamentos online** (2022). StartSe. Disponível em: <<https://www.startse.com/artigos/pagamento-digital-brasileiros-pesquisa/>>. Acesso em: 05 outubro 2023.

KAGGLE. **Brazilian E-Commerce Public Dataset by Olist** (2020). Disponível em: <<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>>. Acesso em: 05 de outubro 2023.

LAWLESS, J. F. **Statistical models and methods for lifetime data**, John Wiley & Sons. 2011.

MORETTIN, P. A; Singer, Júlio da Motta. **Estatística e Ciência de Dados aplicada**. Departamento de Estatística Universidade de São Paulo: São Paulo, 2022.

NAÇÕES Unidas Brasil: **Objetivos de Desenvolvimento Sustentável**. Disponível em: <<https://brasil.un.org/pt-br/sdgs/9>>. Acesso em: 05 outubro 2023.

PEREIRA, A. I. **Análise preditiva de churn em e-commerces brasileiros sobre a ótica da análise de sobrevivência em comparação às técnicas de machine learnings**. 2023. Trabalho de Conclusão de Curso. (Graduação em Estatística) – Universidade Estadual Paulista Júlio de Mesquita Filho. Orientador: Mário Hissamitsu Tarumoto.

PORTO, D. **Mais de 92 milhões de brasileiros acessam a internet apenas pelo celular, diz pesquisa** (2023). CNN Brasil. Disponível em: <[https://www.cnnbrasil.com.br/tecnologia/mais-de-92-milhoes-de-brasileiros acessam-a-internet-apenas-pelo-celular-diz-pesquisa/](https://www.cnnbrasil.com.br/tecnologia/mais-de-92-milhoes-de-brasileiros-acessam-a-internet-apenas-pelo-celular-diz-pesquisa/)>. Acesso em: 05 outubro 2023.

REDAÇÃO OLIST. **O que é Olist? Conheça as soluções e aumente as vendas do seu negócio** (2023). Disponível em: < <https://olist.com/blog/pt/olist/casos-de-uso/o-que-e-olist/>>. Acesso em: 05 outubro 2023.

RUIZ, G. J. **Aplicação de Modelos de Regressão Logística para os Dados Públicos da Empresa Olist**. 2023. Trabalho de Conclusão de Curso. (Graduação em Estatística) Universidade Estadual Paulista Júlio de Mesquita Filho. Orientador: Mário Hissamitsu Tarumoto.

Apêndices

Apêndice 1 – Dicionários dos dados obtidos na plataforma *Kaggle*.

Quadro 1 – Dicionário da base Olist_customers_dataset

Nome da variável	Descrição
customer_id	ID do cliente.
customer_code_prefix	Prefixo do código postal (começo do CEP).
customer_city	Cidade do cliente
customer_state	Estado do cliente

Fonte: Ruiz (2023).

Quadro 2 – Dicionário da base Olist_geolocation_dataset

Nome da variável	Descrição
geolocation_zip_code_prefix	Prefixo do código postal (começo do CEP).
geolocation_lat	Latitude geográfica do cliente.
geolocation_lng	Longitude geográfica
geolocation_city	Cidade do cliente, assim como na base customers
geolocation_state	Estado do cliente, assim como na base customers

Fonte: Ruiz (2023).

Quadro 3 – Dicionário da base Olist_order_items_dataset

Nome da variável	Descrição
order_id	ID do pedido.
order_item_id	Posição do item no pedido, por exemplo, se for comprado um lápis, uma borracha e uma caneta no mesmo pedido, vão aparecer 3 linhas na tabela, a linha onde o produto é o lápis vai aparecer com o order_item_id igual a 1, já o da borracha vai ser igual a 2 e o da caneta igual a 3.
product_id	ID do produto.
seller_id	ID do vendedor
price	Preço do produto
freight	Frete do produto

Fonte: Ruiz (2023).

Quadro 4 – Dicionário da base Olist_order_payments_dataset

Nome da variável	Descrição
order_id	ID do pedido.
payment_sequential	Ordem do pagamento
payment_type	Como foi pago o pedido
payment_installments	Número de parcelas da compra
payment_value	Valor da compra

Fonte: Ruiz (2023).

Quadro 5 – Dicionário da base Olist_order_reviews_dataset

Nome da variável	Descrição
review id	ID da avaliação do pedido
order_id	ID do pedido
review score	Nota dada pelo cliente para aquele pedido
review_comment_title	Título do comentário do cliente
review_comment_message	Mensagem enviada pelo cliente
review_creation_date	Data de criação da avaliação
review_answer_timestamp	Data e hora da resposta da avaliação

Fonte: Ruiz (2023).

Quadro 6 – Dicionário da base Olist_orders_dataset

Nome da variável	Descrição
order_id	ID do pedido
customer_id	ID do cliente
order_status	Status do pedido
order_purchase_timestamp	Data e hora da compra
order_approved_at	Data e hora da aprovação da compra
order_delivered_carrier_date	Data e hora em que o pedido foi entregue para a transportadora
order_delivered_customer_date	Data e hora em que o pedido foi entregue ao cliente
order_estimated_delivery_date	Data de entrega estimada (previsão de entrega)

Fonte: Ruiz (2023).

Quadro 7 – Dicionário da base Olist_products_dataset

Nome da variável	Descrição
product_id	ID do produto
product_category_name	Nome da categoria a qual o produto pertence
product_name_lenght	Número de caracteres no nome do produto
product_description_lenght	Número de caracteres na descrição do produto
product_photos_qty	Quantidade de fotos do produto
product_weight_g	Peso do produto em gramas(g)
product_lenght_cm	Comprimento do produto em centímetros(cm)
product_height_cm	Altura do produto em centímetros (cm)
product_width_cm	Largura do produto em centímetros (cm)

Fonte: Ruiz (2023).**Quadro 8** – Dicionário da base Olist_sellers_dataset

Nome da variável	Descrição
seller_id	ID do vendedor
seller_zip_code_prefix	Prefixo do código postal (começo do CEP) do vendedor
seller_city	Cidade do vendedor
seller_state	Estado do vendedor

Fonte: Ruiz (2023).