UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

LETÍCIA HELEN KLEMS

ANÁLISE DE EVENTOS RECORRENTES APLICADO A DADOS DE E-COMMERCE

PRESIDENTE PRUDENTE

2024

LETÍCIA HELEN KLEMS

ANÁLISE DE EVENTOS RECORRENTES APLICADO A DADOS DE E-COMMERCE

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina TCC II. Orientador: Prof. Dr. Mário Hissamitsu Tarumoto

PRESIDENTE PRUDENTE

2024

K64a

Klems, Letícia Helen

Análise de eventos recorrentes aplicado a dados de e-commerce / Letícia Helen Klems. -- Presidente Prudente, 2024 50 p. : il., tabs.

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. Eventos recorrentes. 3. Modelos semi-paramétricos. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

LETÍCIA HELEN KLEMS

ANÁLISE DE EVENTOS RECORRENTES APLICADO A DADOS DE E-COMMERCE

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:

Orientador: Documento assinado digitalmente gov.br MARIO HISSAMITSU TARUMOTO Data: 17/12/2024 11:04:33-0300 Verifique em https://validar.iti.gov.br

Prof. Dr. Mario Hissamitsu Tarumoto
Departamento de Estatística

Documento assinado digitalmente gov.br MANOEL IVANILDO SILVESTRE BEZERRA Data: 18/12/2024 16:10:46-0300 Verifique em https://validar.iti.gov.br

Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

Presidente Prudente, 06 de dezembro de 2024.

RESUMO

O comércio de produtos de forma online vinha crescendo nos últimos anos a nível mundial, tendo seu processo acelerado pela pandemia de 2020. O desafio do empreendedor de se manter relevante para os clientes também cresceu, considerando que mesmo os que compram online, fazem ampla pesquisa na rede, com as possibilidades de escolha e comparação de produtos, preços e prazos de entrega. Nesse caso, o marketing se tornou chave e conhecer os gostos e comportamentos dos consumidores é essencial para o sucesso. Entre os vários indicadores que podem ser utilizados, o intervalo de tempo entre as compras dos clientes pode indicar que o caminho percorrido está na direção correta, e os lojistas têm a chance de criar oportunidades para vendas adicionais e aumentar a receita. A estimação destes intervalos entre compras pode ser modelada utilizando métodos de análise de sobrevivência multivariada, mais especificamente as técnicas de eventos recorrentes com o intuito de estimar o tempo entre as compras e a probabilidade de o cliente voltar a consumir na loja. Existem diferentes técnicas aplicadas em eventos recorrentes, para esse estudo, foram exploradas as adaptações do modelo semi- paramétrico de Cox em casos multivariados. Neste relatório final, são apresentados os modelos e suas aplicações na base de dados já tratada. O modelo PWP obteve melhor ajuste para o problema, assim como melhor interpretação dos coeficientes.

Palavras-chave: Análise de sobrevivência; Marketing; Comércio; Eventos recorrentes; Modelo de Cox; Múltiplos eventos.

ABSTRACT

Online shopping has been growing worldwide in recent years, and the 2020 pandemic has accelerated the process. The challenge for retailers to remain relevant to customers has also grown, considering that even those who buy online do a lot of research on the web, with the possibility of choosing and comparing products, prices and delivery times. In this case, marketing has become key and knowing consumers' tastes and behaviors is essential for success. Among the various indicators that can be used, the time interval between customer purchases can indicate that the path taken is in the right direction, and shopkeepers have the chance to create opportunities for additional sales and increase revenue. The estimation of these intervals between purchases can be modeled using multivariate survival analysis methods, more specifically recurring event techniques in order to estimate the time between purchases and the probability of the customer returning to the store. There are different techniques applied to recurring events. For this study, adaptations of the Cox semi- parametric model were explored in multivariate cases. In this final report, the models and their applications in the database already treated are presented. The PWP model provided better help for the problem, as well as better interpretation of the coefficients.

Keywords: Survival analysis; Marketing; Retail; Recurrent events; Cox model; Multiple events.

Lista de ilustrações

Lista de tabelas

Sumário

1 INTRODUÇÃO

A pandemia de Covid-19, que se agravou em março de 2020, impactou todos os aspectos da sociedade. A interação com outras pessoas, seja em ambientes de trabalho ou lazer, tornou-se inviável devido à necessidade de distanciamento social. Essa transformação foi profundamente impactante em aspectos como saúde, sociabilidade, meios de subsistência e até mesmo hábitos de consumo.

Com a adoção de medidas de isolamento social, muitos lojistas foram obrigados a fechar temporariamente seus comércios, uma vez que o contato pessoal tornou-se uma perigosa forma de contágio da doença. Ao mesmo tempo, os cidadãos em isolamento encontraram uma nova maneira de consumir, dado que a rotina costumeira de ir até a loja e escolher o produto desejado já não podia acontecer.

A solução encontrada pelos consumidores foi recorrer à internet para fazer suas compras, seja em sites próprios das lojas, grandes plataformas de venda e até redes sociais. Apesar de já existirem empresas vendendo seus produtos online, os números de vendas não eram tão expressivos, principalmente se tratando de itens do dia-a-dia. O comércio virtual, antes do início da pandemia, era dominado por grandes empresas já estabelecidas no mercado. Entretanto, esse número mudou quando o pequeno empreendedor se viu na necessidade de levar a loja até o cliente através da internet. A inserção desse público e a diversificação dos produtos impactou de forma significante a economia digital nos últimos anos.

Segundo o índice MCC-ENET[1], desenvolvido pelo Comitê de Métricas da Câmara Brasileira da Economia Digital (camara-e.net) em parceria com o Neotrust | Movimento Compre & Confie, o E-commerce cresceu 73,88% no ano de 2020 quando comparado ao mesmo período de 2019 no Brasil. Em termos de faturamento, esse crescimento foi de 83,68% de acordo com a mesma fonte.

E esse crescimento não se limita ao primeiro ano da pandemia, conforme o índice MCC-ENET[2] o setor cresceu 35,36% em 2021 em comparação a 2020. Esses

---

[1] FATURAMENTO e vendas do e-commerce mais que dobram, revela índice Câmara-e.net. E-commercebrasil, 2 dez. 2020. Disponível em: https://www.ecommercebrasil.com.br/noticias/e-commerce-brasileiro-cresce-dezembro. Acesso em: 18 nov. 2023.

[2] FATURAMENTO do e-commerce brasileiro tem alta de 48,4% em 2021: Outro dado positivo foi o crescimento das vendas do setor, de 35,36%. [S. l.]: Mercado e Consumo, 24 jan. 2022. Disponível em: https://mercadoeconsumo.com.br/24/01/2022/noticias/faturamento-do-e-commerce-brasileiro-tem- alta-de-484-em-2021/. Acesso em: 16 out. 2023.

dados indicam que a tendência de consumo virtual é crescente, visto que os consumidores se habituaram com as facilidades de poder fazer compras sem sair de casa.

Para os clientes, ter acesso aos mais variados leques de produtos na palma da mão são benéficos. Isso permite comparar preços, ler as avaliações de outras pessoas sobre o produto desejado, além de poder receber a compra em casa. Por esse motivo, os varejistas foram motivados a buscar a inserção nesse mercado cada dia mais competitivo.

Com o aumento de lojas digitais, a concorrência nesse meio também aumenta. Assim, é necessário que a empresa se destaque em meio ao mar de opções fornecidas ao consumidor. Um planejamento de marketing bem direcionado é essencial para que qualquer empresa obtenha sucesso em suas vendas online.

O diretor de Operações da agência Google Premier Digmax[3] Brasil, Danilo Jacomel, comenta a importância do marketing digital para empresas:

> "O marketing digital tem potencial para ajudar qualquer tipo de negócio, de qualquer ramo de atividade e qualquer região. Todos estão conectados com o dispositivo na palma da mão, e o grande espectro de possibilidades de conectar a sua oferta com a demanda existente", (Terra, 2020).

O marketing desempenha um papel fundamental na retenção de clientes em um e-commerce. Conhecer o comportamento do seu público a ponto de prever a probabilidade de um cliente retornar e fazer compras no próximo mês é essencial para criação de estratégias de marketing eficazes.

Como por exemplo a personalização das campanhas com base nas preferências do cliente, a segmentação de esforços para grupos específicos e a automação de mensagens oportunas. Isso resulta em maior retenção de clientes, crescimento consistente para o negócio e otimização de custos de propaganda.

Uma possibilidade de estratégia é estimar o tempo que um perfil de cliente demora para voltar a comprar e desenvolver campanhas de propaganda de acordo com esse tempo.

Ao estimar o intervalo entre as compras dos clientes, as empresas podem criar oportunidades para vendas adicionais e aumentar a receita, contribuindo para um

---

[3] ESPECIALISTA afirma que Marketing Digital é saída para crise econômica. [S. l.]: Terra, 9 abr. 2020. Disponível em: https://www.terra.com.br/noticias/dino/especialista-afirma-que-marketing-digital-e-saida-para-crise-economica,22474ed873913815bb38bbd89aa74b1c5g8ghatn.html. Acesso em: 25 out. 2023.

crescimento econômico mais sustentável e previsível. Conforme as empresas aumentam suas vendas e receitas por meio da fidelização de clientes, também podem criar mais oportunidades de emprego produtivo, especialmente em funções de marketing, análise de dados e atendimento ao cliente.

O objetivo geral deste trabalho é estimar o intervalo de tempo entre as compras online efetuadas por um cliente e a probabilidade de que ele faça uma nova compra, utilizando um modelo de regressão de eventos recorrentes.

Para alcançar o objetivo principal, deverão ser atingidos os seguintes objetivos específicos: Organizar a base de dados, identificar a distribuição que melhor se adequa aos dados e construir modelos semi-paramétricos usando diferentes técnicas de análise de eventos recorrentes.

Esse propósito se alinha diretamente com o Objetivo de Desenvolvimento Sustentável (ODS) 8 da ONU (Organização das Nações Unidas), intitulado "Trabalho Decente e Crescimento Econômico". Esse objetivo estabelece metas essenciais para promover um crescimento econômico sustentável, o emprego produtivo e o trabalho digno. Metas que, em geral, incentivam o crescimento econômico e o desenvolvimento de forma inclusiva, assim como o presente trabalho.

2 REVISÃO BIBLIOGRÁFICA

A análise de sobrevivência é um campo da estatística voltado para o estudo do tempo decorrido até determinado evento acontecer. A partir desse objetivo inicial, foram desenvolvidos métodos para análise de sobrevivência para eventos que podem ocorrer mais de uma vez com o mesmo indivíduo, ou seja, eventos recorrentes.

Kleinbaum e Klein (2012, p 398), fazem um estudo completo das técnicas de Análise de Sobrevivência. Desde as definições iniciais de censura, até modelos robustos para dados no tempo. Os autores dedicam um capítulo inteiramente para dados de eventos recorrentes, ou seja, quando o evento pode ocorrer mais de uma vez em determinado intervalo de tempo.

Uma abordagem explanada pelos autores para esse caso, é a modelagem paramétrica de fragilidade compartilhada. Esse tipo de modelo é muito utilizado no caso univariado, quando os dados são relacionados a indivíduos que mesmo expostos ao mesmo ambiente, são inerentemente diferentes.

Os modelos de fragilidade compartilhada no caso multivariado são usados para entender a diferente fragilidade em grupos distintos. Esses modelos paramétricos que tem por base distribuições de probabilidade de variáveis não-negativas e possuem um fator de risco influenciando no risco do indivíduo. Esse fator de risco, ou fragilidade, varia para os diferentes grupos dos dados. Quando um indivíduo possui maior fragilidade, este é mais suscetível a falha.

Apesar de ser uma técnica robusta, os modelos de fragilidade compartilhada possuem maior grau de complexidade e pressupostos a serem atendidos, quando comparado a métodos semi-paramétricos. Além de ser necessário um bom ajuste dos dados na distribuição escolhida.

O estudo de Colosimo e Giolo (2006, p 301) traz diferentes técnicas para eventos recorrentes. Apesar do estudo também fazer menção ao modelo de fragilidade compartilhada, o foco está em modelagem semi-paramétrica dos dados através de modelos marginais e condicionais.

A primeira metodologia explanada é o modelo desenvolvido por Andersen e Gill (1982, p 1101). Essa técnica considera o risco inicial igual para todos os intervalos de tempo e também considera que o indivíduo retorna ao grupo de risco após cada evento com mesmo risco inicial.

O segundo modelo, proposto por Prentice, Williams e Peterson (1981, p 375), parte do pressuposto de que, para que um indivíduo esteja em risco no m-ésimo evento, ele deve ter vivenciado o evento *m-1*. Esse modelo considera a existência de interdependência entre os tempos de ocorrência de eventos para um mesmo indivíduo.

A comparação entre os modelos é fortemente abordada, principalmente levando em consideração o uso de variáveis importantes. O modelo de Andersen e Gill fornece estimativas não viciadas mais próximas mesmo sem a participação de todas as variáveis relevantes no modelo. Em contrapartida, a performance do modelo Prentice, Williams e Peterson pode ser viciada na falta dessas covariáveis. Apesar de imperfeitos, todas as metodologias citadas trazem resultados relevantes quando aplicados corretamente.

Dado que as técnicas de análise de sobrevivência são muito aplicadas na área médica, grande parte de suas aplicações práticas também são nesse campo. A dissertação de mestrado de Mota (2013) aplica os modelos marginais e condicionais abordados acima em dados de pacientes com doença renal crônica em tratamento dialítico, com o objetivo de estudar a incidência de eventos cardiovasculares, considerando também o efeito das covariáveis no resultado final dos modelos.

3 ANÁLISE DE SOBREVIVÊNCIA

.1 Definições

A análise de sobrevivência é o segmento da estatística que tem como variável de interesse o tempo decorrido até determinado evento acontecer. Esse conjunto de técnicas é muito aplicado na área da saúde, como por exemplo, o tempo até a recuperação de pacientes submetidos a diferentes medicações. Mas também é muito aplicada em cenários industriais, podendo avaliar o tempo necessário até a manutenção do maquinário sem prejuízo à rotina.

Apesar da análise de sobrevivência ser mais aplicada nos ramos acima, sua aplicação tem ganhado mais força em cenários distintos como o estudo do tempo até a extinção de uma espécie ou quanto tempo demora para que um cliente se tornar inadimplente. Em suma, nos casos que é necessário estimar o tempo até que algo aconteça, as técnicas de sobrevivência podem ser aplicadas.

Outro benefício dessa área é a possibilidade de trabalhar com dados censurados, ou seja, que foram observados somente até determinado período antes que o evento de interesse aconteça. Isso permite que esses casos agreguem informação no estudo, visto que caso fossem removidas poderiam gerar vícios na análise. Dessa forma, é possível estimar mais corretamente o tempo até o evento, ou até mesmo a probabilidade do evento ocorrer dentro de determinado intervalo de tempo.

A partir desse conceito, é possível estudar o tempo decorrido desde o início da observação até que o evento de interesse, ou falha, ocorra. Então, todos os indivíduos de uma população estudada são considerados em risco até que a falha aconteça ou que a observação seja interrompida, no caso censurada.

.2 Dados censurados

Existem diferentes razões pelas quais uma observação venha a ser censurada. É possível que o estudo seja encerrado após um período pré-estabelecido, e nesse caso, todas os indivíduos que não falharam serão censurados no tempo final. Também é possível que o contato com um dos indivíduos do estudo seja perdido ao longo do

tempo, ou que em caso de estudos clínicos, o paciente deixe de tomar a medicação estudada por motivos além da cura.

Existem três principais tipos de censura, que representam a falta de informação do indivíduo em algum período, porém em diferentes estágios do estudo. São eles:

- Censura à esquerda: é usada quando não é possível definir o risco do evento observado de fato começou. Por exemplo, um paciente diagnosticado com câncer é inserido no estudo, porém não se sabe quando a doença teve início.
- Censura à direita: esse tipo de censura acontece quando o indivíduo estava sendo acompanhado, mas em algum momento do estudo perderam o acompanhamento, o que caracteriza a censura aleatória.
- Censura intervalar: essa censura acontece quando a observação do indivíduo não é contínua, logo é possível afirmar que o evento de interesse aconteceu num intervalo de tempo, ao invés de um tempo t. Por exemplo, o acompanhamento periódico de um paciente que no espaço de tempo entre as consultas desenvolveu a doença observada.

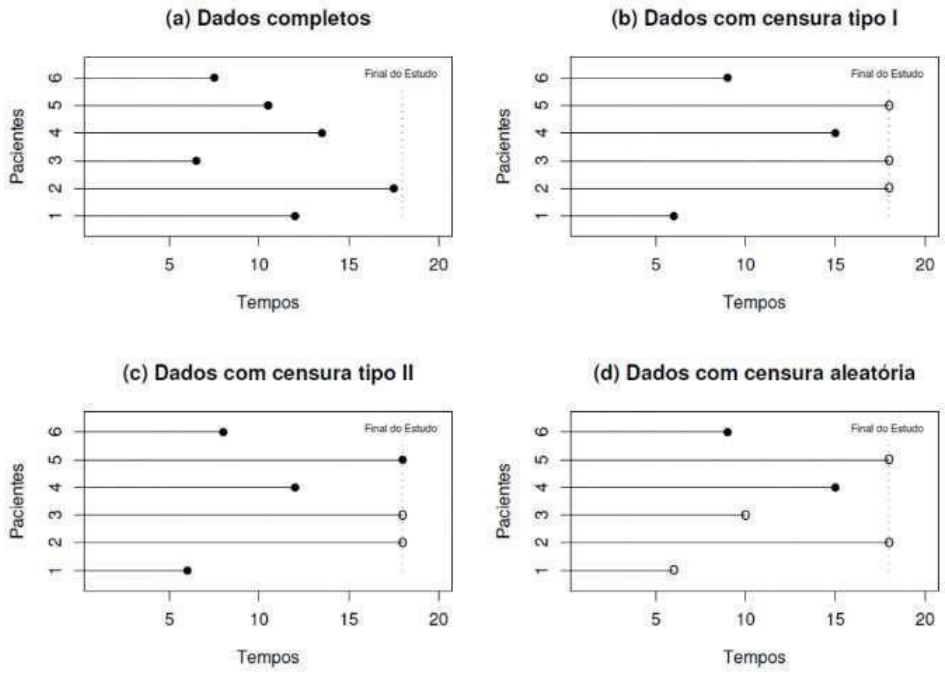
Dentro da censura à direita, existem ainda dois tipos: a censura tipo I que ocorre quando o experimento termina em um tempo pré-determinado, nesse caso os indivíduos que não apresentaram a falha são censurados. Existe também a censura tipo II, que acontece quando o experimento é finalizado após um número pré-determinado de falhas.

A censura pode ser definida usando duas variáveis aleatórias: uma representando o tempo de falha do indivíduo (T), e a outra variável representando a o tempo censurado associado ao mesmo indivíduo (C). Sendo a variável C independente de T. Logo, a marcação de censura ocorre seguindo a representação a seguir:

$$\delta = \begin{cases} 1, se\ T \leq C \\ 0, se\ T > C \end{cases} \quad (1)$$

A variável de censura é representada por δ_i, , com $i = 1, \dots, n$ indivíduos. Assim, para aqueles que sofreram a falha, é atribuído o valor 1. Caso contrário, a censura recebe valor 0.

Figura 1 – Gráficos de diferentes tipos de censura.



Fonte: Colosimo e Giolo (2006).

Nesses casos, ainda que o indivíduo não seja acompanhado até a falha, o seu tempo de permanência no estudo também agrega informação sobre a população estudada. Assim, por meio da análise de sobrevivência, esse tempo observado também é utilizado nas estimações ao invés de apenas descartar a observação.

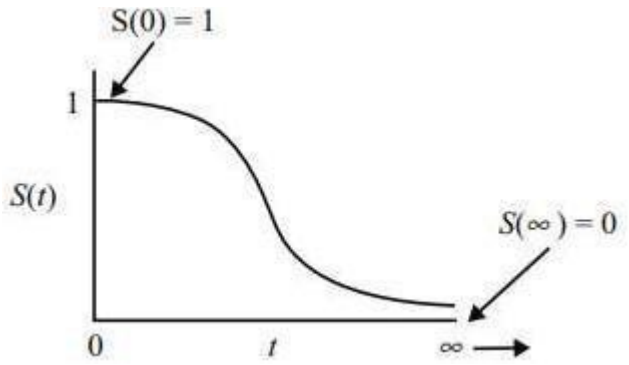
.3 Função de sobrevivência e funções relacionadas

Obtendo o T como variável não-negativa de tempo, é possível estabelecer funções de probabilidade para estudar o comportamento do tempo até o evento. A primeira função descrita é a chamada função de sobrevivência, que tem por objetivo calcular a probabilidade de um indivíduo não falhar até determinado tempo t, ou seja, sobreviver até o tempo t. Essa função pode ser descrita como:

$$S(t) = P(T \geq t) = 1 - F(t) \quad (2)$$

Assim, a função $S(t)$ é contínua, monótona decrescente e pode tender ao infinito em casos que alguma observação nunca apresente o evento.

Figura 2 – Gráfico base da função de sobrevivência.



Fonte: Kleinbaum e Klein (2012).

Uma outra função importante é a função de taxa de falha, ou risco, que estuda o risco específico do indivíduo falhar no tempo t visto que ele sobreviveu até esse tempo. Esta pode ser crescente, decrescente ou constante dependendo da distribuição dos dados. Existem casos ainda, que a função de risco pode obter o formato de “banheira”, que combina as diferentes formas para descrever o comportamento da população. A função de risco é dada por:

$$\lambda(t) = \frac{S(t) - S(t+\Delta t)}{\lim_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t)} \quad \Delta t S(t) \qquad \Delta t \qquad (3)$$

Segundo Colosimo e Giolo (2006, p. 23), a função de risco é mais informativa que a função de sobrevivência visto que suas formas podem diferir significativamente. Enquanto as formas das funções de sobrevivência de diferentes distribuições podem ser muito semelhantes.

A partir da função de taxa de falha, é possível encontrar a função de taxa de falha acumulada. Esta corresponde a taxa de falha acumulada por indivíduo e é dada por:

$$\mathbf{\Lambda}(\mathrm{t}) = \int_0^t \lambda(u) du \qquad (4)$$

Apesar dessa função não possuir interpretação direta, ela é útil para definir a função de taxa de falha. Uma vez que a função de risco é difícil de ser estimada e $\mathbf{\Lambda}(\mathrm{t})$ possui um estimador com propriedades ótimas principalmente em casos de modelagem não paramétrica.

.4 Modelos de regressão

Assim como em outras áreas da estatística, em sobrevivência é possível modelar os dados, entender a relação entre variáveis e até estimar predições. Essa modelagem pode ser paramétrica ou semi-paraméterica. No primeiro caso, o modelo é ajustado com base em uma distribuição de probabilidade já conhecida que melhor se adequa aos dados. Já nos modelos semi-paramétricos o cálculo dos parâmetros é feito empiricamente através da própria base de dados.

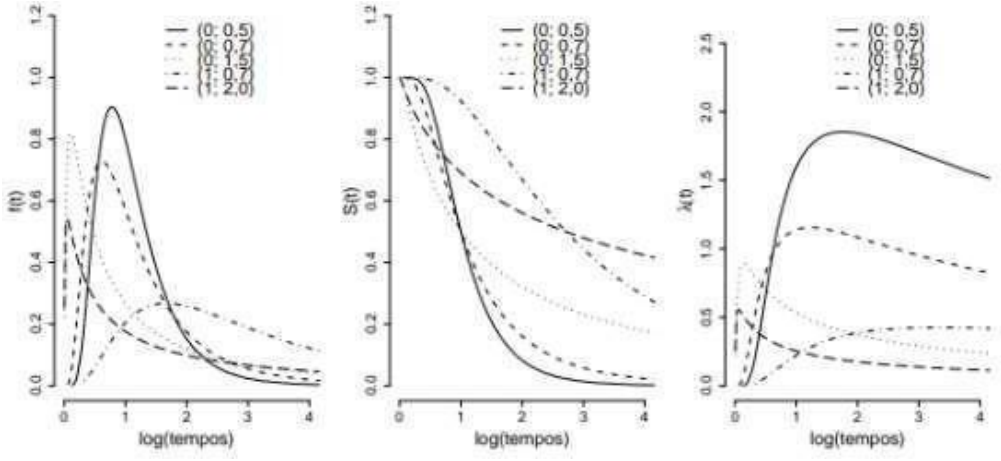
.4.1 Modelos paramétricos

Os dados usados na análise de sobrevivência podem ser considerados tempo de vida, seja de forma literal na medicina, tempo de vida de uma máquina ou duração da rentabilidade de um cliente. Como se trata de medição de tempo, não é possível existir um tempo negativo. Logo, as distribuições usadas para ajustar dados de sobrevivência são também distribuições não-negativas. As principais distribuições de probabilidade são: Exponencial, Weibull, Log-normal e Gamma.

Cada uma das distribuições listadas anteriormente possui diferentes funções de densidade de probabilidade, e consequentemente, diferentes formas de função de sobrevivência e taxa de falha.

A distribuição exponencial, entre as citadas, é a menos utilizada visto que sua função de taxa de falha é constante. Por outro lado, as distribuições Weibull e Log-normal são mais visadas, uma vez que dependendo do valor dos parâmetros de cada função, a forma das funções de sobrevivência e risco assumem curvas muito distintas. Isso permite a maior flexibilidade e mais fácil adequação do modelo aos dados.

Figura 3 – Gráficos da função de densidade, sobrevivência e risco acumulado para diferentes valores de parâmetros para a distribuição Log-normal.



Fonte: Colosimo e Giolo (2006).

A distribuição Gamma e Gamma generalizada também possuem o benefício da flexibilidade a depender dos parâmetros. Uma vez que a distribuição exponencial e Weibull podem ser escritas como um caso particular desta distribuição. Entretanto, a Gamma e Gamma generalizada envolvem maior grau de complexidade para estimação dos parâmetros.

.4.2 Modelagem semi-paramétrica

Um dos tipos de modelos, com muitas aplicações, é o modelo de regressão semi-paramétrico, ou seja, a fórmula do modelo não depende inteiramente de distribuições paramétricas de probabilidade. A metodologia semi-paramétrica de maior significância para a análise de sobrevivência univariada é o modelo de Cox.

O modelo de regressão de Cox (Cox, 1972) é muito utilizado principalmente quando é necessário o estudo de efeito de covariáveis, uma vez que através da estimação dos parâmetros é possível compreender como uma determinada classificação pode aumentar ou diminuir a taxa de falha do indivíduo.

Essa metodologia também é conhecida como modelo de taxas de falha proporcionais, visto que têm como pressuposto que os riscos entre os grupos sejam constantes ao longo do tempo. Isso significa que caso a taxa de falha de um indivíduo

A no início do experimento seja duas vezes maior do que de um indivíduo B, essa proporção deve se manter até o fim do estudo.

O método de Cox é composto por um componente não paramétrico, o que permite maior flexibilidade para o modelo pois esse dado será proveniente da própria distribuição empírica dos dados. A segunda parte da fórmula é composta por uma função não-negativa envolvendo as covariáveis. Como demonstrado abaixo:

$$\lambda(t|x) = \lambda_0(t) g(\boldsymbol{x}'\boldsymbol{\beta}) \tag{5}$$

têm-se que $\lambda_0(t)$ é o componente não-paramétrico, geralmente representado pela função de taxa de falha da base. Já a função $g(x'\beta)$ na maioria das vezes apresenta a forma multiplicativa das covariáveis,

$$g(\boldsymbol{x}'\boldsymbol{\beta}) = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\} = \exp\{\beta_1 x_1 + \cdots + \beta_n x_n\} \tag{6}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros estimados associados as covariáveis e $\boldsymbol{x}$ o vetor de valores para cada variável por indivíduo.

Para esse tipo de modelo, os parâmetros também são estimados maximizando a função de verossimilhança associada. Entretanto, por se tratar de um modelo semi-paramétrico, é usada a verossimilhança parcial para estimação. A grande diferença desse método é que é levado em consideração a probabilidade condicional do *i-ésimo* indivíduo falhar no tempo t_i sabendo quais observações ainda estão em risco no mesmo tempo.

Por esse motivo, a função de verossimilhança parcial que estima β para o modelo de Cox é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp\{\boldsymbol{x}'\boldsymbol{i}\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\boldsymbol{x}_j \boldsymbol{\beta}\}} \right)^{\delta_i} \tag{7}$$

onde δ_i representa a falha daquela observação.

Os parâmetros serão os valores de $\boldsymbol{\beta}$ que maximizam a função descrita acima. Os mesmos podem ser obtidos resolvendo o sistema de equações dado por $U(\boldsymbol{\beta}) = 0$, sendo $U(\boldsymbol{\beta})$ o vetor de escore, ou seja, as derivadas de primeira ordem da função $l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta}))$. Dado pela função a seguir:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{x}_i - \frac{\sum_{j \epsilon R(t)} \boldsymbol{x}_j \exp\{\boldsymbol{x}_j' \boldsymbol{\beta}\}}{\sum_{j \epsilon R(t_i)} \exp\{\boldsymbol{x}_j' \boldsymbol{\beta}\}} \right] = 0 \quad (8)$$

4 EVENTOS RECORRENTES

Os modelos explanados no capítulo anterior são aplicadas nos casos univariados, ou seja, quando o evento de interesse ocorre apenas uma vez para cada participante do estudo. Todavia, existem situações em que o evento pode ocorrer mais de uma vez durante o tempo de observação. Como exemplos podemos citar a reincidência de infecção urinária em pacientes com câncer na bexiga, ou até mesmo a ocorrência de infartos no miocárdio.

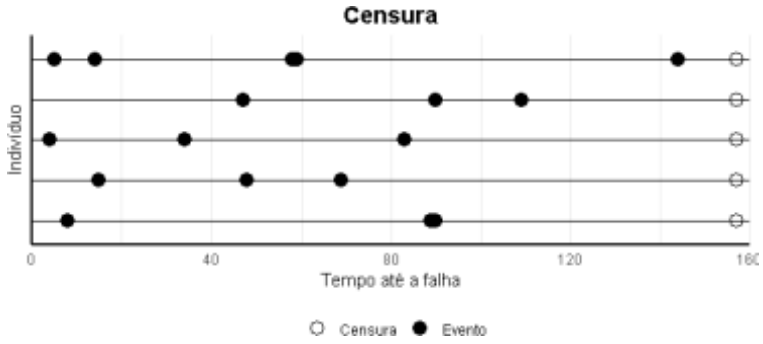
Em casos multivariados, os tempos de falha entre eventos de um mesmo indivíduo pode não ser mutuamente independentes, de forma que não seria possível aplicar o modelo regular de Cox, uma vez que os tempos se sobrepõe e esses intervalos podem estar correlacionados (Carvalho et al., 2011, p. 317)..

Para estas situações, foram criados meios de modelar o tempo entre ocorrências no caso de eventos recorrentes. Assim como para caso univariado, existe pelo menos duas abordagens. Os modelos paramétricos, que são os modelos de fragilidade compartilhada e os semi-paramétricos que se desdobram em diversas outras técnicas. No presente trabalho o foco será nos modelos semi-paramétricos marginais para ajuste dos dados.

A construção da variável de censura também é análoga ao método univariado. Sendo T_{mi} o tempo da m-ésima falha para *o i-ésimo* indivíduo e C_{mi} o tempo de censura para o i-ésimo indivíduo, considerando censura à direita, têm-se que:

$$\delta_{mi} = \begin{cases} 1, & se\ T_{mi} \leq C_{mi} \\ 0, & se\ T_{mi} > C_{mi} \end{cases} \quad (9).$$

Figura 4 – Gráfico representando os casos de censura multivariada para cinco indivíduos da base de estudo.



Fonte: Elaborado pelo autor (2024).

.1 Modelagem marginal

A modelagem marginal tem por objetivo ignorar a correlação entre observações de um mesmo indivíduo e realizar uma correção na variância dos parâmetros estimados. Dessa forma, é encontrado uma estimação mais robusta para o vetor de parâmetros.

Essa correção pode ser feita através da estimativa Jackknife. Esta consiste na retirada de uma ou mais amostras da base e então o estimador é calculado novamente com os valores restantes (Mota, 2013, p. 14). Esse método cria uma estimativa não viciada para dados correlacionados quando a observação deixada de fora for independente dos valores utilizados na conta. Sendo essa técnica agrupada por indivíduo, é retirado um sujeito da base ao invés de observações no tempo.

Pode-se escrever os resíduos de Jackknife agrupados por indivíduo da seguinte forma:

$$J_i = \boldsymbol{\beta} - \boldsymbol{\beta}_{(i)} \quad (10),$$

onde $\boldsymbol{\beta}_{\boldsymbol{i}}$ é o resultado do ajuste sem os valores do indivíduo i.

Outra forma de calcular os mesmos resíduos acima foi proposta por Therneau e Grambsch (2000, p.521), é feita utilizando a metodologia de Newton-Raphson. Dada por:

$$\Delta\beta = 1'(UI^{-1}) = 1'D \tag{11}$$

onde U é a matriz de escore residual, I^{-1} corresponde à variância de β e D representa a mudança em β a cada iteração.

Quando agrupada por indivíduo, essa estimativa pode ser reescrita como:

$$V_j = {}^{n-1}(J - \bar{J})'(J - \bar{J}) \tag{12}$$

n

em que $\bar{J}$ é a matriz das médias das colunas de J. A variância passa, então, a ser escrita como $D'D = I^{-1}(U'U)I^{-1}$, sendo $U'U$ o fator de correção.

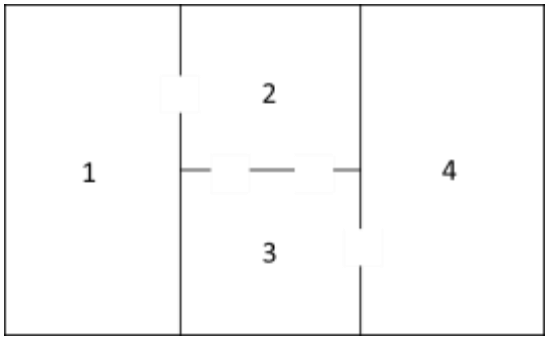
.2 Processo de contagem

A teoria de processos de contagem, ou processos estocásticos, pode ser de extrema utilidade para explicar alguns fenômenos na análise de sobrevivência multivariada. Isso porque esse método consiste em modelar processos que evoluem ao longo do tempo sempre associados a uma probabilidade.

Nesse caso, será dado ênfase para as cadeias de Markov. Estas consistem em uma sequência de variáveis aleatórias x_0, ... , x_n que são condicionalmente independentes entre si. Isso significa que o valor da variável x_m só depende do valor de x_{m-1}, e não diretamente de todas anteriores.

Um exemplo simples seria um indivíduo estando em um labirinto e tendo este que passar por uma das portas do ambiente, é possível associar uma matriz de probabilidades. Considerado a imagem apresentada na Figura 5, ao sair da sala 1 a única escolha é ir para a sala 2, logo a probabilidade de estar na sala 2 no tempo t_2 é 1. Analogamente, estando na sala 2 a probabilidade de voltar para a sala 1 é de aproximadamente 0,33 e de ir para a sala 3 é de 0,67 visto que possui duas portas, e assim sucessivamente.

Figura 5 –Exemplo de labirinto.



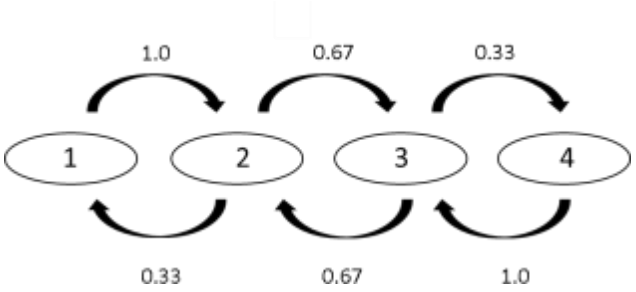
Fonte: Elaborado pelo autor (2024).

Dessa forma, a matriz de probabilidade associada ao movimento desse indivíduo no labirinto é dada por:

$$P = \begin{bmatrix} 0.0 & 1.0 & 0.0 & 0.0 \\ 0.33 & 0.0 & 0.67 & 0.0 \\ 0.0 & 0.67 & 0.33 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \end{bmatrix}$$

Outra forma de representar é através dos estados, explicitando movimentos que podem ser feitos a partir dele e a probabilidade associada, como demonstrado na figura 6:

Figura 6 – Representação da transição dos estados.



Fonte: Elaborado pelo autor (2024).

Essa representação é útil para explicar a relação entre estado de risco e estado em que o evento ocorre para múltiplos eventos, tendo como probabilidade associada o risco basal $\lambda_0(t)$ usado para calcular os modelos.

.3 Modelo de Andersen e Gill (AG)

Proposto por Andersen e Gill em 1982, esse modelo marginal é uma adaptação do modelo semi-paramétrico de Cox. O primeiro grande diferencial desse modelo é que o risco basal é considerado o mesmo para todos os intervalos de tempo analisados. Isso significa que o indivíduo retorna ao grupo de risco após o evento com a mesma taxa de falha base.

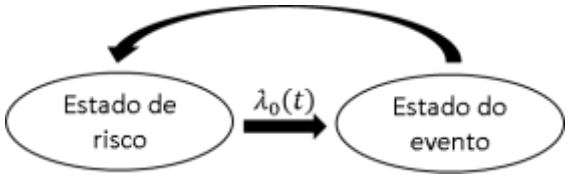
Tal afirmação, implica que os eventos são considerados independentes entre os intervalos de tempo. Se a primeira observação da base começar no tempo de entrada, têm-se que o modelo para o indivíduo i será dado por:

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{x}'(t)\boldsymbol{\beta}\} \tag{13}$$

onde $\lambda_0(t)$ é o risco basal, $\boldsymbol{x}'_i(t)$ são as covariáveis para o indivíduo no tempo t e $\boldsymbol{\beta}$ é o vetor de parâmetros estimados.

O modelo AG pode ser representado através das cadeias de Markov, uma vez que o próximo estado depende somente do anterior. Nesse caso, o indivíduo sai do estado de risco para o estado do evento com a única possibilidade de retornar para o estado inicial visto que o risco basal é o mesmo independentemente da quantidade de eventos. A ilustração é apresentada na Figura 7.

Figura 7 – Representação dos riscos no modelo AG através de uma cadeia de Markov.



Fonte: Elaborado pelo autor (2024).

Como esse modelo têm por pressuposto a independência entre eventos de um mesmo indivíduo, é recomendado que sejam comparadas a variância do modelo de Cox simples e a variância robusta. Onde o esperado é que a variância robusta seja maior em valor.

Como este modelo tem por base os estudos de Cox, a estimação dos parâmetros é feita de forma análoga ao método usado para casos univariados através da função de verossimilhança parcial. A função é adaptada apenas para incorporar as informações de múltiplos eventos. Pode ser representada da seguinte forma:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{m=1}^{k} \left(\frac{\exp\{\boldsymbol{x}'_{mi}\boldsymbol{\beta}\}}{\sum_{j=1}^{n} \sum_{l=1}^{k} \exp\{\boldsymbol{x}'_{lj}\boldsymbol{\beta}\}} \right)^{\delta_i} \quad (14)$$

Nesse caso, o vetor $\boldsymbol{\beta}$ correspondente aos parâmetros corresponde aos valores que maximizam a expressão 14.

.4 Modelo de Prentice, Williams e Peterson (PWP)

O modelo descrito por Prentice, Williams e Perterson (1981) difere no fato de que o indivíduo só estará em risco do *m-ésimo* evento acontecer caso tenha sofrido a falha m-1. Isso significa que as taxas de falha basais serão diferentes de acordo com a quantidade de eventos sofridos por observação.

Por esse motivo, o método PWP também é conhecido como modelagem marginal condicional, uma vez que o indivíduo só estará em risco do evento m caso a condição anterior, falha *m-1*, acontecer. Logo, essa análise é realizada em estratos, sendo estes a quantidade de eventos apresentados considerando a dependência entre os tempos de uma mesma observação.

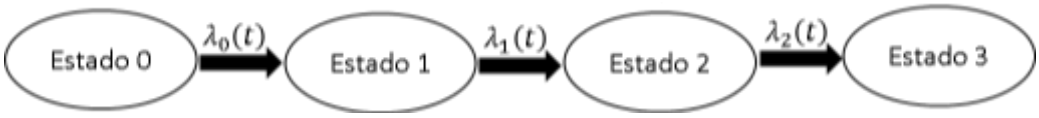
A formulação do modelo condicional incorpora o pressuposto de dependência entre os tempos. Assim, o risco basal para o segundo evento será zero até que a primeira falha aconteça e assim sucessivamente. A função de taxa de falha é dada por:

$$\lambda_{im}(t) = \lambda_{0m}(t) \exp\{\boldsymbol{x}'_{\boldsymbol{i}}(t)\boldsymbol{\beta}_{\boldsymbol{m}}\} \quad (15)$$

onde $\lambda_{0m}(t)$ é o risco basal variando para m eventos, $\boldsymbol{x}'_{\boldsymbol{i}}(t)$ são as covariáveis para o indivíduo no tempo t e $\boldsymbol{\beta}_m$ é o vetor de parâmetros estimados para aquele estrato.

Este modelo também pode ser representado utilizando o processo markoviano. Isso porque, considerando a dependência entre os tempos, o segundo evento só pode acontecer caso o primeiro já tenha acontecido. Assim, dependendo somente do passo anterior como afirma a definição de cadeia de Markov. A Figura 8 ilustra esta situação.

Figura 8 – Representação dos riscos no modelo PWP através de uma cadeia de Markov.



Fonte: Elaborado pelo autor (2024).

Nesse caso, as probabilidades associadas a transição de estados são os riscos basais para cada estrado de eventos.

A estimação dos parâmetros do modelo PWP têm por base a formulação de verossimilhança parcial de Cox. Todavia, como diferença da função correspondente do modelo AG, a iteração correspondente a contagem de eventos se faz presente também no vetor de parâmetros e censura. Isso acontece porque de acordo com o estrato da observação, a função de taxa de falha basal é alterada. Como exemplificado na equação a seguir:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{m=1}^{k} \left(\frac{\exp\{\boldsymbol{x}'_{mi}\boldsymbol{\beta}_m\}}{\sum_{j=1}^{n} \exp\{\boldsymbol{x}'_{mj}\boldsymbol{\beta}_m\}} \right)^{\delta_{mi}} \tag{16}$$

Onde o vetor $\boldsymbol{\beta_m}$ corresponde aos vetores de parâmetros estimados para cada estrato.

5 CRITÉRIOS DE SELEÇÃO DE MODELOS

Para a avaliação do modelo mais adequado aos dados, pode-se utilizar dois critérios de comparação de modelos. O primeiro é a Informação de Akaike (AIC) desenvolvido por Akaike em 1974. Esse critério avalia o ajuste e simplicidade do modelo, pois faz uso do número de parâmetros e da log da função de verossimilhança. É dado por:

$$AIC = -2\log\{L(\boldsymbol{\beta})\} + 2K \tag{17}$$

em que K é o número de parâmetros usados para estimação do modelo. O modelo a ser selecionado é aquele que obtiver o menor valor de AIC.

O segundo critério é o Critério de Informação de Bayes (BIC) e segue uma linha similar ao cálculo do AIC. Este surge do ponto de vista Bayesiano com probabilidade a priori igual para cada modelo. Nesse caso o modelo selecionado também é o que apresenta o menor valor. O critério é calculado da seguinte forma:

$$BIC = -2\log\{L(\boldsymbol{\beta})\} + K\log\{n\} \tag{18}$$

Ainda, uma terceira forma de avaliar o ajuste do modelo é através da medida de concordância. Esse é um método para dados pareados, sendo nesse caso o par formado pelo valor verdadeiro e a predição do modelo, que consiste na razão de pares com valores concordantes em relação ao total (Atkinson, Therneau 2024, p.2). Para essa medida, quanto mais próximo de um for a concordância do modelo, mais adequado para o conjunto de dados ele é.

Outra etapa de verificação do ajuste do modelo é a análise de resíduos para identificar possíveis tendências. Nos casos de ajustes com base no modelo de Cox existe ainda a necessidade de testar o pressuposto de riscos proporcionais entre as variáveis.

Para tal verificação, existem diversos métodos eficientes. Serão destacados dois deles: os resíduos de Schoenfeld e resíduos de Martingale. O primeiro método avalia se o efeito de uma variável é constante ao longo do tempo, independente do momento da ocorrência do evento.

Para cada covariável x_i no tempo do evento t_i, têm-se:

$$r_{ik} = \delta_i(\boldsymbol{x}_{ik} - a_{ik}) \tag{19}$$

$$a_{ik} = \frac{\sum_{j \in R(tj)} \boldsymbol{x}_{ik} \exp(\boldsymbol{x_i \beta})}{\sum_{j \in R(tj)} \exp(\boldsymbol{x}\,\boldsymbol{\beta})} \tag{20}$$

Onde $R(t_j)$ é o conjunto de indivíduos em risco em t_j e x_{ik} representa o valor da covariável k para o indivíduo i presente no grupo de risco.

A análise desse método se dá de forma gráfica, onde o cenário ideal é uma distribuição aleatória dos valores de r_{ik} ao redor do zero. Caso apresente tendência crescente ou decrescente, significa que o modelo falhou no pressuposto de riscos proporcionais e é necessário tratar as variáveis dependentes do tempo.

O segundo método avaliado é chamado de resíduos de Martingale. Essa análise consiste na diferença entre a censura e a expectativa de um evento ao longo do tempo de duração. Para essa metodologia, ambos esperança e soma dos resíduos deve ser igual a zero.

Analogamente aos resíduos de Schoenfeld, também é possível analisar esses resíduos de forma gráfica. Para um modelo bem ajustado, é esperado que os valores estimados estejam aleatoriamente distribuídos ao redor do zero. O cálculo de Martingale se dá por:

$$M_i = \delta_i - \lambda_0(t)\exp\{\boldsymbol{x'_i \beta}\} \tag{21}$$

Esses resultados também são úteis para avaliar outliers e pontos de atenção que podem estar influenciando a qualidade do ajuste.

6 RESULTADOS

Visando alcançar o objetivo geral do estudo, que é estimar o comportamento de compras online, foram utilizados dados provenientes de um e-commerce com foco em vendas de eletrônicos. Dado que são informações públicas usados somente para estudo, as mesmas são anônimas, sendo possível identificar os clientes apenas por um ID disponibilizado. A base contém informações como a data do acesso, tipo de ação, segmento do produto e precificação do item.

Para aplicação dos modelos, foi necessário uma transformação na base de dados com o intuito de deixar no formato desejado, com marcação de censura e tempo entre eventos. Essa transformação foi realizada através do software R, software Python e aplicativo Microsoft Excel.

.1 Apresentação e descrição dos Dados

Os dados utilizados para a aplicação dos modelos são relacionados a venda de produtos online. A base foi disponibilizada de forma virtual contendo as seguintes variáveis:

- Event_time: dia e horário que o evento aconteceu;
- Event_type: qual foi o tipo do evento, podendo ser visualização, carrinho ou compra;
- Product_id: código do produto;
- Category_id: código da categoria do produto;
- Category_code: nome da categoria do produto;
- Price: preço do produto;
- User_id: código identificador único para cada cliente.

Segue exemplo da base original sem tratamentos e suas respectivas variáveis:

Tabela 1 – Amostra da base de dados sem tratamento.

event_time	event_type	product_id	category_id	category_code	price	user_id
24/09/2020	view	1996170	214441	electronics	31.9	1515915625
24/09/2020	view	139905	214442	computers	17.16	1515915625
24/09/2020	view	635807	214442	computers	113.81	1515915625

Fonte: Elaborado pelo autor (2024)

A base tem a primeira observação registrada no dia 24/09/2020 e a última no dia 28/02/2021, sendo essas as datas iniciais e finais do estudo.

Para adequação dos dados, foi necessário retirar todas as observações que possuíam a variável “event_type” marcada como visualização, dessa forma o evento estudado foi considerado como sendo “event_type” = carrinho ou “event_type” = compra. No estudo, todos os dados são censurados no fim do período observado, caracterizando censura do tipo I.

Tabela 2 – Volume de dados.

Volume de dados		
Base	N	IDs Base
inicial	885129	407283
Base sem views	91381	38200

Fonte: Elaborado pelo autor (2024)

Após esse filtro, ficaram na base apenas os 38.200 clientes que compraram pelo menos uma vez no site. Além disso, é interessante destacar que mais de 90% da base apresentam apenas 1 evento, sendo 13 o máximo de compras observadas. Assim, pode-se afirmar que não existe muita recorrência nas compras nesse site em um período de menos de seis meses.

Após esse filtro, a distribuição de eventos e censuras é dada pela tabela 3, o maior volume de censuras se dá pela grande presença de clientes que compraram apenas uma vez.

Tabela 3 – Volume de eventos na base.

Volume	%
Evento	47.26%
Censura	52.74%

Fonte: Elaborado pelo autor (2024)

A base de dados para aplicação desse modelo exige um formato específico. Nesse caso, cada indivíduo pode ser representado por mais de uma linha, em casos de mais de um evento ou uma situação em que após a falha aquele sujeito continuou sendo acompanhado no estudo.

Também é necessário que o tempo de ocorrência da falha seja descrito como início e fim do período. Para essa transformação foi calculada a diferença, em dias, entre os eventos obtendo uma coluna de tempo inicial e uma de tempo de ocorrência do evento.

A Tabela 4 apresenta a base formatada, sem covariáveis, contendo a variável censura onde o valor 1 representa evento e 0 a informação censurada. A variável ordem_id também é essencial para a aplicação do modelo PWP, pois representa a ordem e contagem dos eventos por ID.

Tabela 4 – Modelo da base tratada sem covariáveis

user_id	Start	stop	censura	ordem_id
151591562535328	0	9	1	1
151591562535328	9	12	1	2
151591562535328	12	157	0	3

Fonte: Elaborado pelo autor (2024)

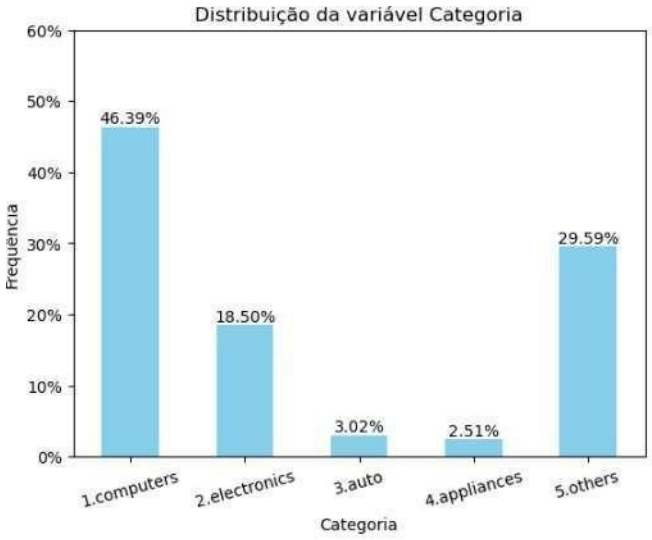
.2 Construção de covariáveis

Foram tratadas três variáveis da base original para o uso no modelo. A primeira variável a ser tratada foi a variável “category_code”, que originalmente, possuí 15 tipos de seções diferentes. Para melhor análise dos dados, foram mantidas as quatro categorias com maior volumetria: “Computers”, “Eletronics”, “Auto” e “Appliances”. As demais foram agrupadas na categoria “others”, contendo as seguintes informações:

“acessories”, “apparel”, “construction”, “contry_yard”, “furniture”, “jewelry”, “kids” e “medicine”.

Uma vez agrupadas, foi calculada a moda das categorias por id para captar qual tipo de produto aquele cliente mais comprou. Na figura 9, pode-se notar que os produtos mais procurados estão na categoria de computadores, com destaque também para eletrônicos que ocupa 18.5% do volume da variável.

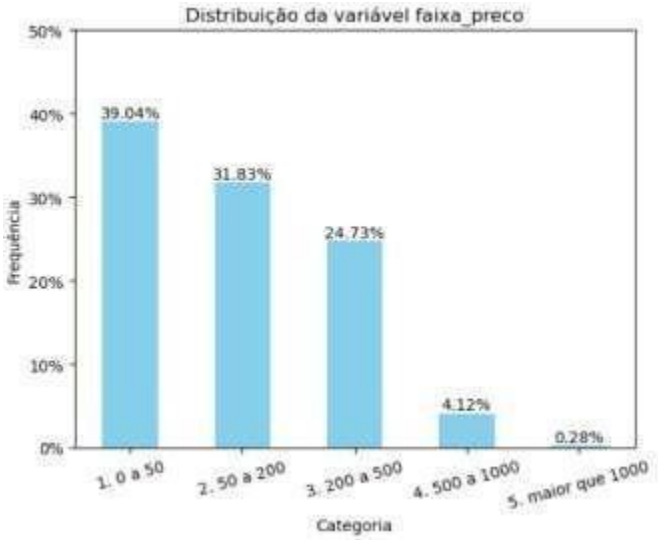
Figura 9 – Distribuição da variável “Categoria”



Fonte: Elaborado pelo autor (2024)

Outra variável a ser tratada foi “Price”, ou seja, o preço. A partir da mesma, foi calculada a média do preço gasto por ID, sendo uma nova covariável contínua. Através dessa variável contínua foi agrupado a média do preço em faixas de valor, criando uma variável categórica.

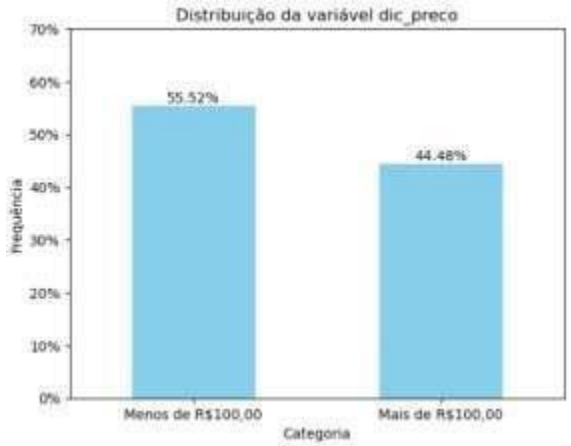
Figura 10 – Distribuição da variável “faixa_preco”.



Fonte: Elaborado pelo autor (2024)

E por fim, o preço do produto foi dicotomizado, recebendo valor 1 para compras acima de R$100,00 e 0 para compras abaixo desse valor, como apresentado na tabela 3. Foram testados os três formatos de variáveis no modelo para averiguar qual apresentava maior ganho no resultado final.

Figura 11 – Distribuição da variável “dic_preco”.

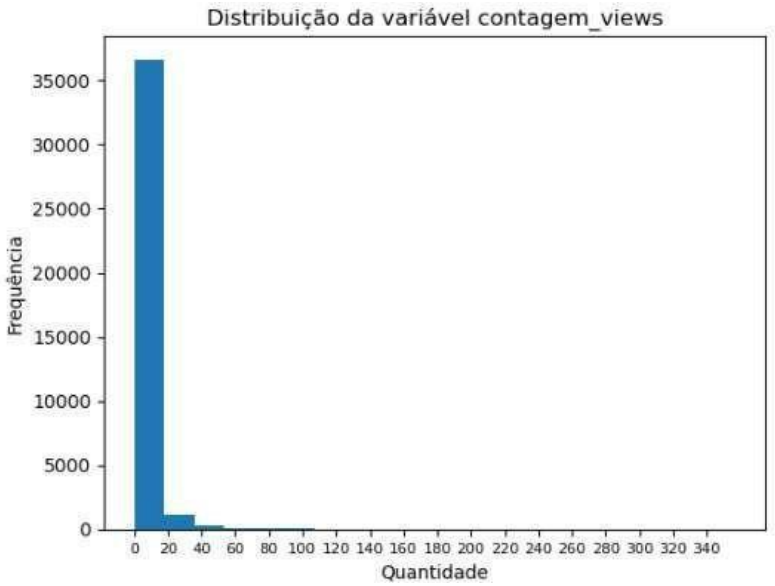


Fonte: Elaborado pelo autor (2024)

A variável “event_type” foi usada para definir o evento mas também para a criação de uma nova covariável que indica a quantidade de vezes que o cliente apenas olhou os produtos, sem realizar a compra ou movimentação para o carrinho. Isso foi

feito através da contagem das observações da variável “event_type” marcadas como visualização por ID.

Figura 12 – Distribuição da variável “contagem_views”.



Fonte: Elaborado pelo autor (2024)

Através da figura 12, é observável que a maior parte do público acessou o site menos de 20 vezes até realizar a compra. Essa variável apresentou um outlier, onde um ID realizou 357 consultas no site antes do evento.

.3 Teste de amostra

Após os filtros abordados acima, a base de dados a ser utilizada para modelagem possuía 90% de clientes com apenas uma compra, ou seja, sem recorrência. Esse alto volume influenciou os modelos testados de forma a não conseguirem aprender o comportamento dos clientes que voltaram a comprar.

Para diminuir o efeito do público censurado, foram realizados teste de amostras com base na proporção do volume de IDs com mais de uma compra. Considerando o total de 3329 usuários recorrentes e fixos, foi retirada uma amostra aleatória dos IDs com somente uma compra considerando 100% do total de usuários recorrentes, ou seja, 3329 IDs. Assim, a amostra final possuiu um volume de 6658 IDs distintos.

Essa mesma lógica foi utilizada para outras proporções, como demonstrado na tabela 5. Para seleção da amostra mais adequada para treinar o modelo, foram treinados modelos de Cox para cada situação e analisados as métricas AIC e BIC.

Visto que o modelo mais adequado aos dados é o que possui o menor valor de AIC e BIC, pode-se afirmar que quanto menor o volume de IDs não recorrentes, melhor o ajuste do modelo.

Tabela 5 – Comparação das amostras.

Percentual	Ids recorrentes	Ids não recorrentes	Número de linhas	AIC cox	BIC cox
100,00%	3229	3329	17684	129912,2	129930,7
75,00%	3244	2497	16024	118374,9	118393,3
50,00%	3265	1665	14367	106905,8	106924,2
25,00%	3286	832	12710	95164,54	95182,99
10,00%	3298	333	11716	88681,01	88702,67
5,00%	3303	166	11381	86612,08	86630,53
2,00%	3305	67	11184	83479,56	83498,01

Fonte: Elaborado pelo autor (2024)

A amostra escolhida, com base nos menores valores de AIC e BIC no modelo de Cox, foi a amostra com apenas 2% de Ids não recorrentes. Isso pelo melhor desempenho do modelo e volume suficiente de usuários diferentes para realizar a análise do mesmo.

Essa medida foi necessária para não incorporar o viés de compras únicas visto que esse não é o intuito do estudo. Outro filtro necessário para melhor visualização dos dados foi limitar em até quatro eventos, o que não impactou no volume de usuários distintos na base de dados.

Para avaliar a amostra escolhida, foi realizado um estudo utilizando o gráfico de Kaplan-Meier. Esse é um método empírico que responde à curva de sobrevivência dos dados, como esse é um estudo multivariado foi feito um gráfico para cada grupo de eventos (primeiro, segundo, etc.), como demonstrado na figura 13:

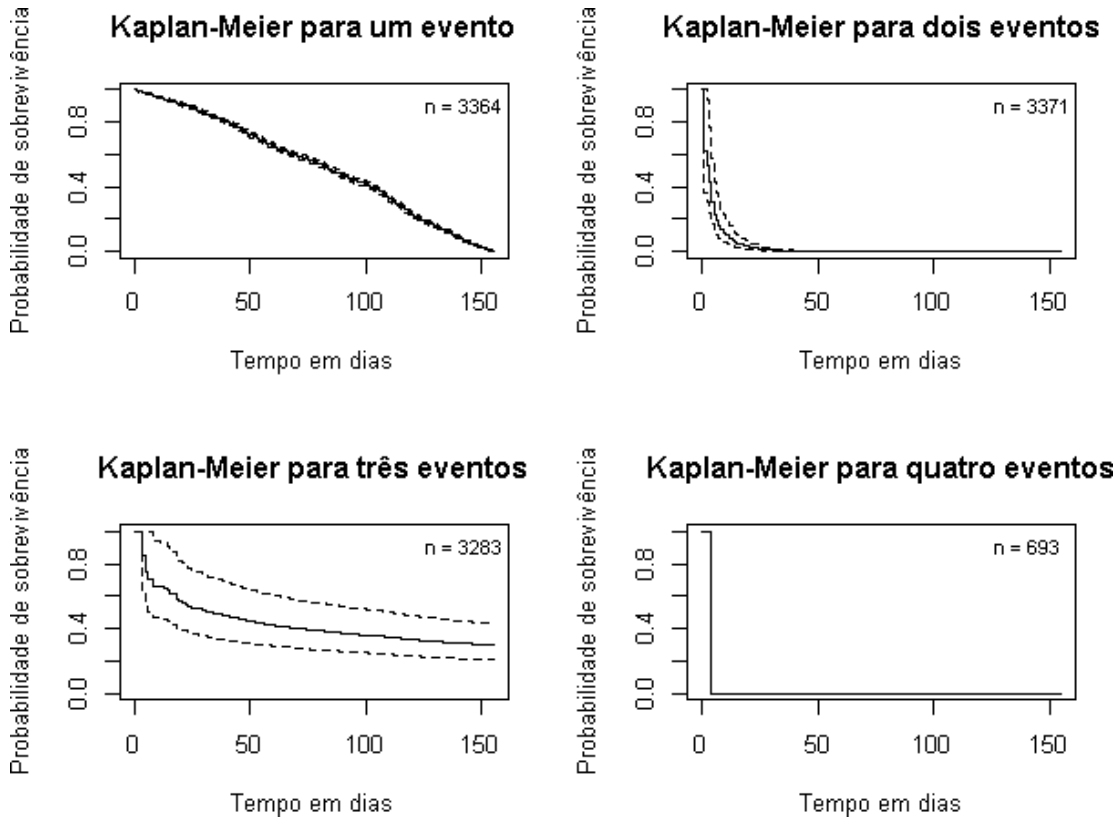
É certo afirmar que o comportamento das curvas de sobrevivência são diferentes para cada ordem de evento:

- Para ordem_id = 1, a probabilidade de sobrevivência (não comprar) decai com comportamento similar a uma reta. No dia 100, a chance de sobrevivência da base é de 40%;

- Para ordem_id = 2, a chance de voltar a realizar uma compra é extremamente maior, decaindo com velocidade;
- Para ordem_id = 3, a curva de sobrevivência se estabiliza por volta do dia 50 com mais de 40% de chance de não voltar a comprar;
- Para ordem_id = 4, essa queda próxima ao dia 0 indica que a chance de sobreviver para quem apresenta 4 marcações é quase nula.

O comportamento para n = 2 e n = 4 não era esperado dado o contexto. Entretanto, visto que esse é o comportamento da base, é esperado o mesmo comportamento nos modelos estimados a seguir.

Figura 13 – Kaplan-Meier por evento.



Fonte: Elaborado pelo autor (2024)

.4 Modelo Andersen e Gill (AG)

Para desenvolvimento do modelo, os dados foram separados em base de treino, contendo 70% dos usuários e base de teste com 30% dos IDs distintos. A taxa de eventos ficou similar em ambas as bases com um percentual geral de 70% de observações não censuradas.

A seleção de variáveis foi realizada através do método stepwise, técnica que adiciona ou remove variáveis do modelo de acordo com o impacto no AIC do modelo treinado. Isso com foco na qualidade de ajuste e na simplicidade do modelo.

Para o modelo AG, através do stepwise, a única variável selecionada foi “contagem_views”. Na tabela 6 estão as métricas do modelo treinado.

Tabela 6 – Métricas de estimação modelo AG

Variável	Estimativa	Erro padrão	Erro padrão robusto	P valor
contagem_views	0.0043899	0.0004501	0.0006464	0.0000000000111

Fonte: Elaborado pelo autor (2024).

O erro padrão robusto é a raiz quadrada da variância robusta, citada na seção 4.1 desse trabalho. Este consiste na correção da variância devido a possível correlação entre os tempos de um mesmo indivíduo. Em geral, esse valor é interpretado da mesma forma que o erro padrão;

Analisando os resultado, têm-se que tanto o erro padrão quando o erro padrão robusto são valores baixos, sugerindo uma baixa variabilidade nas estimativas e, portanto, uma maior precisão das medidas obtidas. Em relação ao p-valor, quando definido um nível de 95% de confiança, é possível afirmar que a variável é significativa para o modelo.

Em modelos não-paramétricos de análise de sobrevivência, a interpretação dos parâmetros é feita de forma exponencial. Isso oferece o aumento na razão de risco a cada aumento de unidade na variável:

- Se exp(coef) > 1, um aumento na variável aumenta o risco do evento;
- Se exp(coef) < 1, um aumento na variável diminui o risco do evento;
- Se exp(coef) = 1, a variável não apresenta efeito no risco.

No caso do modelo AG estimado, para cada visualização que o cliente apresente, o risco dele voltar a comprar aumenta em 0,4%.

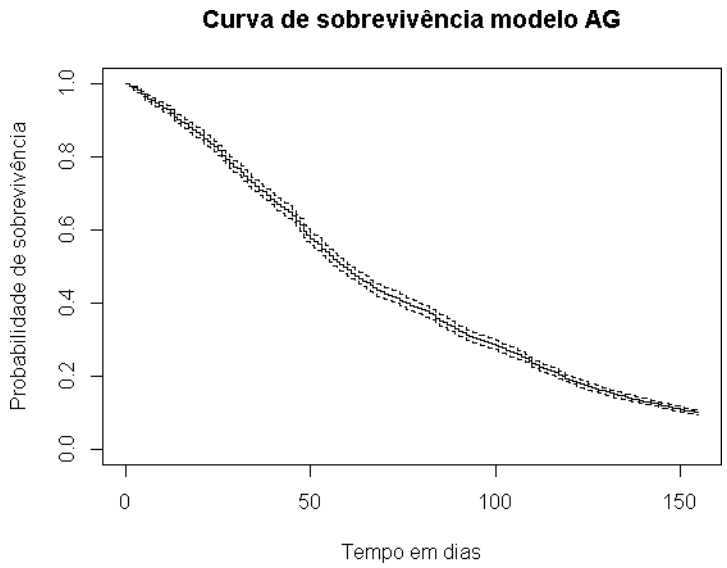
Tabela 7 – Coeficentes estimados do modelo AG.

Variável	exp(coeficiente)	Limite inferior	Limite superior
contagem_views	1.004	1.003	1.006

Fonte: Elaborado pelo autor (2024)

A curva estimada de sobrevivência segue um decaimento sem um ponto de estabilização. O intervalo de confiança das predições é bem ajustado aos dados. Visto que o modelo considera o risco basal sendo o mesmo até o *i-ésimo* evento, é estimada apenas uma curva como apresentado na figura 14:

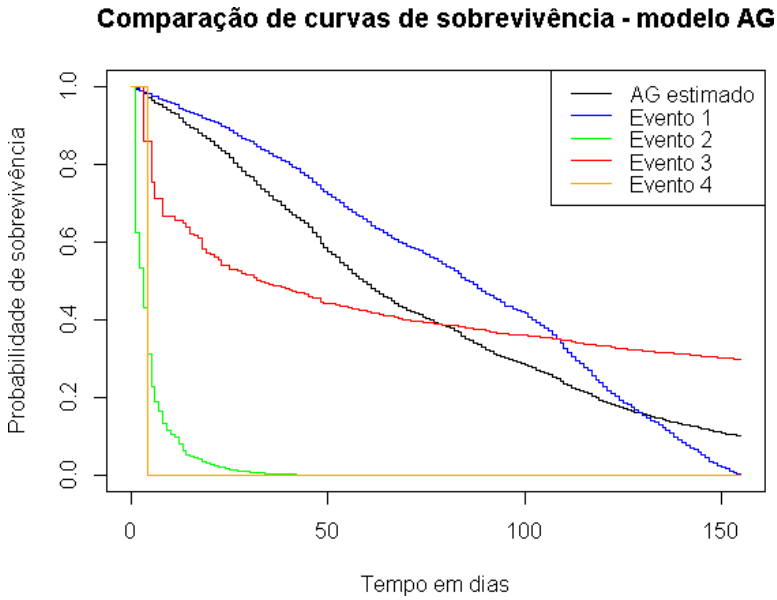
Figura 14 – Curva de sobrevivência modelo AG.



Fonte: Elaborado pelo autor (2024).

Comparando a curva estimada com os gráficos de Kaplan-Meier para cada grupo de eventos, é possível afirmar que o modelo possui um comportamento que generaliza a informação das demais. Logo, apesar do bom ajuste do modelo e da significância da variável, o modelo AG não é capaz de capturar os nuances da base de dados.

Figura 15 – Comparação das curvas de sobrevivência com o modelo AG.



Fonte: Elaborado pelo autor (2024)

.5 Modelo Prentice, Williams & Peterson (PWP)

A seleção de variáveis para o treinamento do modelo PWP também foi realizada através do stepwise. Apresentaram ganho no modelo as variáveis: “media_preco”, “categoria” e “contagem_views”. Na tabela 8 estão as métricas obtidas para o modelo PWP. Visto que a variável categoria não é numérica, é avaliado o ganho em relação a uma categoria estabelecida como base, nesse caso foi “categoria” = computers.

Tabela 8 – Métricas de estimação do modelo PWP.

Variável	Estimativas	Erro padrão	Erro padrão robusto	P valor
media_preco	0.00059667	0.00008608	0.00010743	0.0000000279
categoria2.electronics	0.13589763	0.04225599	0.06539288	0.0377
categoria3.auto	0.03894999	0.07806402	0.10299473	0.7053
categoria4.appliances	0.04785322	0.09603713	0.13237672	0.7177
categoria5.others	0.08931809	0.03777235	0.05745855	0.1201
contagem_views	0.00423875	0.00045385	0.00099264	0.0000195333

Fonte: Elaborado pelo autor (2024)

Analisando os resultado, têm-se que tanto o erro padrão quando o erro padrão robusto são valores baixos, sugerindo uma baixa variabilidade nas estimativas e, portanto, uma maior precisão das medidas obtidas. Em relação ao p-valor, quando definido um nível de 95% de confiança, é possível afirmar que todas as variáveis são significativas para o modelo, visto que o p-valor é menor que 0,05. No caso da variável categoria, ela permanece significativa ainda que nem todos os contrates apresentem ganho de informação.

Na tabela 9 estão apresentados os coeficientes estimados de cada variável e os seus respectivos intervalos de confiança. É possível notar que as mesmas categorias que não possuem um p-valor significativo, possuem intervalos que contém o valor 1. Isso mostra que de fato não são importantes para definir risco de evento, visto que caso assumam o valor 1, nada alteram na estimativa.

A variável “media_preco” indica que para cada aumento de uma unidade, a chance de voltar a comprar diminui em 0.06%. Ou seja, existe maior recorrência para clientes que compram produtos com preços menores. Quando a categoria do produto for eletrônicos, existe 12.71% a menos de chance de compra quando comparado com a computadores. Já em relação a quantidade de visualizações no site, a cada visualização a chance de ocorrer o evento aumenta em 0.42%

No geral, os impactos das variáveis nos riscos estimados é baixa porém significativa.

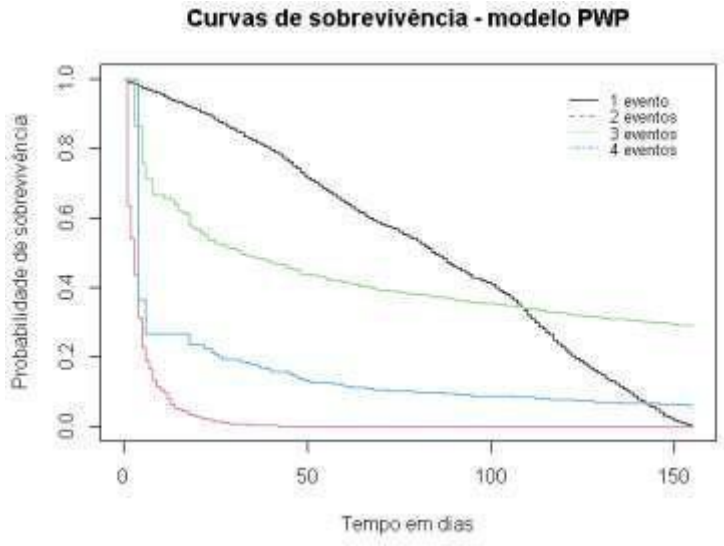
Tabela 9 – Coeficientes estimados do modelo PWP.

Variável	exp(coeficiente)	Limite inferior	Limite superior
media_preco	0.9994	0.9992	0.9996
categoria2.electronics	0.8729	0.7679	0.9923
categoria3.auto	1.0397	0.8497	1.2723
categoria4.appliances	1.049	0.8093	1.3598
categoria5.others	0.9146	0.8171	1.0236
contagem_views	1.0042	1.0023	1.0062

Fonte: Elaborado pelo autor (2024)

Como discutido anteriormente, o modelo PWP estima um modelo para cada nível de evento. Ou seja, têm-se uma curva para quem realizou uma compra, para quem realizou duas e assim por diante. A visualização para as curvas de até quatro eventos estão na figura 16. O modelos treinados apresentam grande divergência de comportamento entre si.

Figura 16 – Curvas de sobrevivência estimadas do modelo PWP.



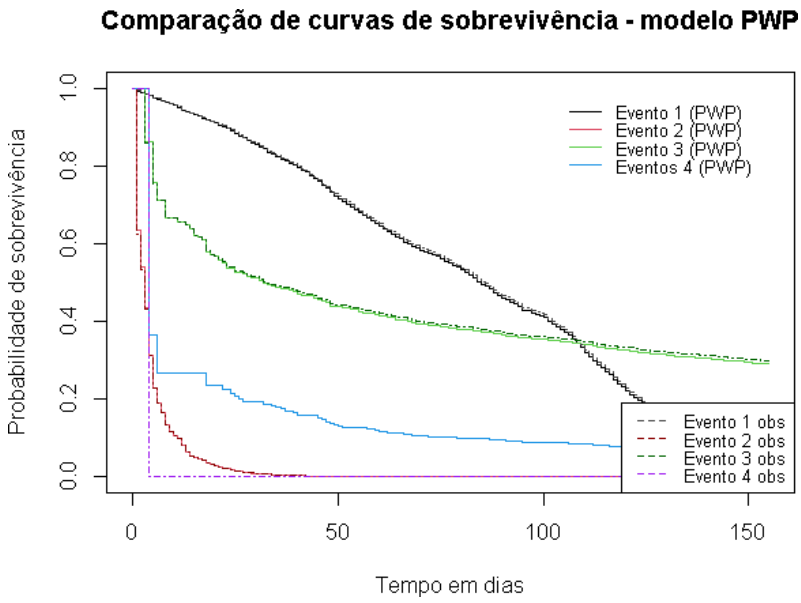
Fonte: Elaborado pelo autor (2024)

A figura 17 ilustra a comparação das curvas de sobrevivência estimadas pelo modelo e a curva empírica para cada ocorrência de eventos. As linhas continuas representam a saída do modelo e as tracejadas os dados reais. É possível afirmar que

para um, dois e três eventos a técnica PWP foi capaz de capturar bem o comportamento dos dados e devolver uma estimativa muito próxima.

Contudo, para quatro eventos o estimado se distanciou da curva verdadeira, comportamento que pode ser justificado pela pouca quantidade de observações nesse cluster.

Figura 17 – Comparação das curvas de sobrevivência com o modelo PWP.



Fonte: Elaborado pelo autor (2024)

Para ambos modelos também foram calculados o AIC, BIC e Concordância. A tabela 10 apresenta os resultados dessas métricas e através deles pode-se concluir que o modelo PWP apresenta o melhor ajuste ao conjunto de dados estudado. Uma vez que apresentou os menores valores de AIC e BIC, e o maior valor de concordância.

Tabela 10 – Medidas de ajuste dos modelos.

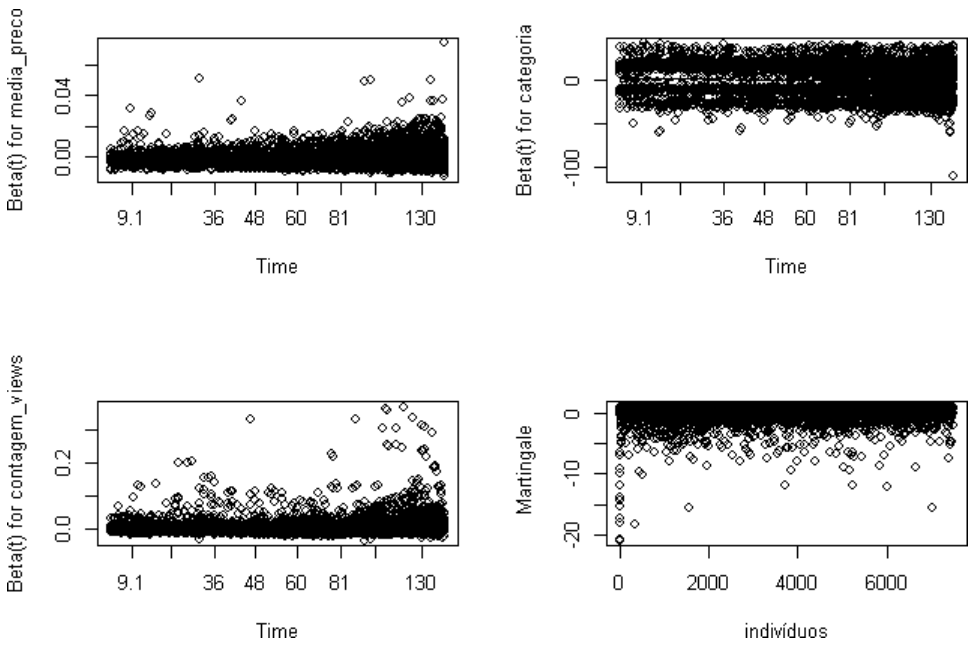
Modelo	AIC	BIC	Concordância
AG	80850.49	80853.57	0.5343
PWP	60732.87	60751.32	0.5588

Fonte: Elaborado pelo autor (2024)

Após a escolha do modelo, é importante analisar os resíduos do mesmo para avaliar a adequabilidade do ajuste utilizando os resíduos de Schoenfeld e os resíduos de Martingale. Na figura 18 estão dispostos os gráficos de Schoenfeld para cada uma das variáveis utilizadas no modelo.

Através da análise gráfica, é possível afirmar que as variáveis “contagem_views” e media_preco não estão aleatoriamente distribuídas ao redor de zero, demonstrando uma tendência crescente. O mesmo comportamento ocorre para os resíduos de Martingale, indicando a falha do modelo no teste de riscos proporcionais.

Figura 18 – Resíduos de Schoenfeld e Martingale para o modelo PWP.



Fonte: Elaborado pelo autor (2024)

Uma solução para amenizar a tendência apresentada nessas variáveis é aplicar a transformação de log. Esse método diminui o impacto de valores extremos, suavizando a informação que vai ser utilizada no modelo sem alterar os resultados de forma significativa.

Por esse motivo, foi desenvolvido um terceiro modelo usando o método PWP com as variáveis log(“contagem_views”), log(“media_preco”) e “categoria”. Na tabela 11 estão dispostas as métricas do novo modelo. Nesse caso, não houveram

mudanças significativas nas estimativas das informações logarítmicas. Entretanto considerando um nível de 95% de confiança, a variável categoria não é mais significativa para o modelo.

Tabela 11 – Métricas de ajuste do modelo PWP com log.

Variável	Estimativas	Erro padrão	Erro padrão robusto	P valor
log(media_preco)	-0.084440	0.01452	0.02172	0.0001010000
categoria2.electronics	-0.123640	0.044570	0.067892	0.0686770000
categoria3.auto	0.109230	0.077600	0.102410	0.2859550000
categoria4.appliances	0.082250	0.095960	0.137880	0.5502310000
categoria5.others	-0.043060	0.039760	0.059400	0.4685220000
log(contagem_views)	0.177140	0.015260	0.026360	0.00000000002

Fonte: Elaborado pelo autor (2024)

As variáveis continuam com similar interpretação, com exceção de categoria que não contribui para o aumento ou diminuição do risco do modelo com uso de log. A interpretação dos coeficientes agora pode ser realizada de forma percentual. Assim, a cada 1% que for acrescentado na variável “media_preco”, ocorre uma redução de aproximadamente 0.08% no risco de compra.

No caso de “contagem_views”, a cada 1% de acréscimo no valor da variável, a chance de acontecer o evento aumenta em 0.2%. As contribuições para movimentação de risco dessas variáveis é baixa porém significativa.

Tabela 12 – Coeficientes estimados do modelo PWP com log.

Variável	exp(coeficiente)	Limite inferior	Limite superior
log(media_preco)	0.9187	0.8803	0.9587
categoria2.electronics	0.8853	0.775	1.0113
categoria3.auto	1.1180	0.9147	1.3665
categoria4.appliances	1.0876	0.83	1.4251
categoria5.others	0.96	0.8545	1.0786
log(contagem_views)	1.1988	1.1379	1.2629

Fonte: Elaborado pelo autor (2024)

A aplicação da transformação logarítmica nas variáveis resultou em uma melhora sutil nas métricas do modelo. Os valores de AIC e BIC diminuíram levemente,

sugerindo um ajuste melhor em comparação com os modelos anteriores. Embora a concordância também tenha apresentado uma ligeira queda, ela ainda permanece superior à obtida pelo modelo AG, indicando que o ajuste do modelo atual continua sendo adequado.

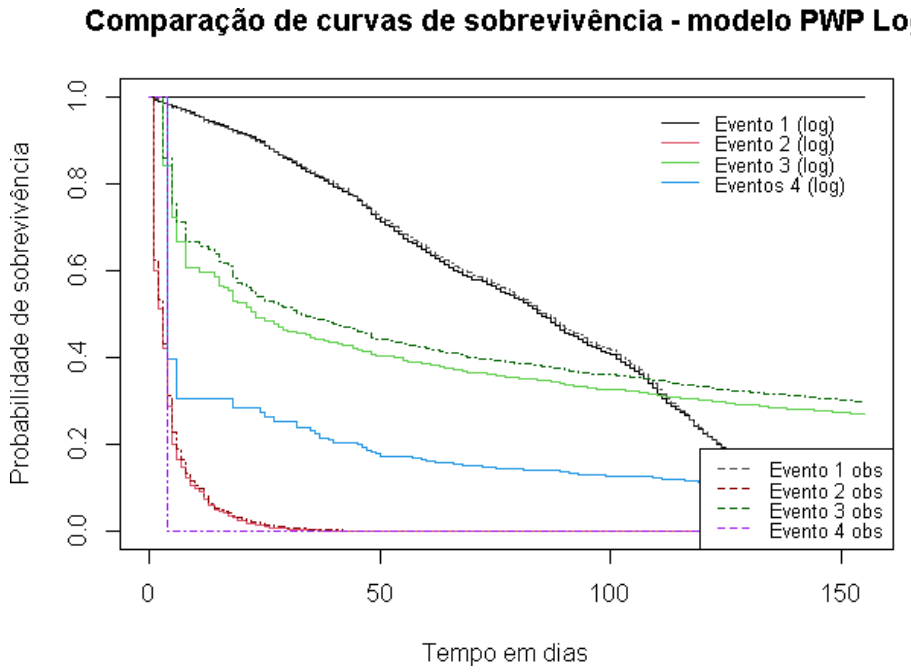
Tabela 13 – Medidas de ajuste dos modelos.

Modelo	AIC	BIC	Concordância
AG	80850.49	80853.57	0.5343
PWP	60732.87	60751.32	0.5588
PWP com log	60700.94	60719.39	0.545

Fonte: Elaborado pelo autor (2024)

Em relação as curvas de sobrevivência estimadas, o novo modelo também apresenta bons resultados. O comportamento das curvas para 1 e 2 eventos se mantém muito próximas ao empírico. Já para 3 eventos, o estimado segue a mesma curva porém com um nível de risco maior. Para 4 eventos o comportamento é similar o modelo PWP original.

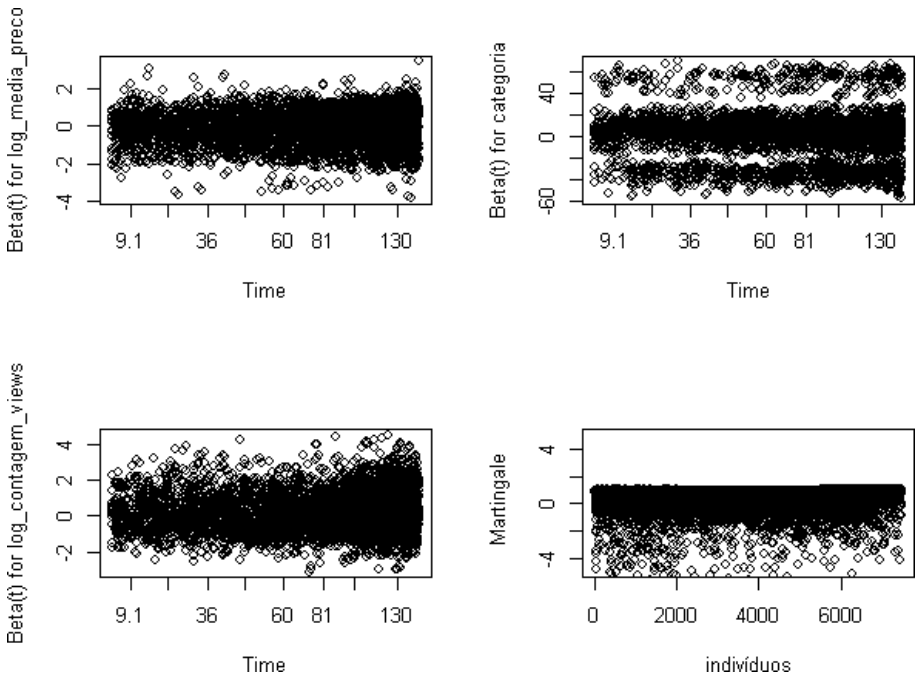
Figura 19 – Comparação de curvas de sobrevivência para o modelo PWP com log.



Fonte: Elaborado pelo autor (2024).

Os resíduos do modelo PWP com log estão melhores distribuídos ao redor de zero quando comparado com o modelo original. Têm-se que as variáveis log(“media_preco”) e log(“contagem_views”) não apresentam mais tendências visíveis. Em relação aos resíduos de Martingale, estes indicam que o ajuste não é totalmente adequado aos dados, com uma maior tendência a resultados negativos.

Figura 20 – Resíduos de Schoenfeld e Martingale para o modelo PWP com log.



Fonte: Elaborado pelo autor (2024).

.6 Caso de uso

O objetivo do modelo é segmentar públicos para melhorar a assertividade de campanhas de marketing, e assim reduzir os custos com propaganda sem perder potencial de vendas. Com esse foco, foi desenvolvido um caso de uso com as informações do modelo PWP treinado.

Para o exemplo, foi dado foco na curva de sobrevivência de um evento. Ou seja, qual o público mais propenso a realizar uma segunda compra no site. Logo, foi calculada a probabilidade de sobrevivência (não compra) para todos os indivíduos passados 30 dias da primeira compra.

Com base na probabilidade, foram separados três grupos: um grupo com alto risco de compra, um grupo com médio risco e outro com baixo risco de voltar a comprar no site em questão. Esses grupos foram quebrados por percentis aproximadamente iguais de 33% de volume em cada segmento.

Tabela 14 – Distribuição do volume dos grupos de risco.

Risco	Volume	
	N	%
Alto	335	33.00%
Médio	335	33.00%
Baixo	345	33.99%
Total	1015	100.00%

Fonte: Elaborado pelo autor (2024).

A probabilidade de não voltar a comprar, ou seja, sobreviver ao evento, foi dicotomizada através de uma distribuição Bernoullii para cada indivíduo. Nesse caso, o 1 representa a compra, devido a menor probabilidade de sobreviver ao evento. E consequentemente o 0 representa a sobrevivência (sem compras recorrentes).

Para a base de um evento a taxa geral de compra é de 12,33%, entretanto essa taxa se movimenta quando observada por grupos de risco, como apresentada na tabela 15. Essas taxas foram usadas como taxas base de conversão para futuras compras por grupos de risco nesse evento. Na realidade, as taxas de conversão assumem valores muito diferentes, com segmentações distintas, porém para fins de estudos esses valores foram utilizados.

Tabela 15 – Taxa de compra por grupos de risco.

Taxa de compra	%
Alto	17.96%
Médio	15.26%
Baixo	10.72%

Fonte: Elaborado pelo autor (2024).

Partindo da variável dicotomizada é possível estimar o potencial de venda da base, usando a variável media_preco como ticket médio de cada cliente. Para níveis de estudo, foi considerado o custo de R$0,10 centavos por e-mail disparado de propaganda para um lote de 1015 leads da base de teste do modelo. Com essas informações foram construídos dois cenários de marketing: com e sem uso do modelo que estão descritos na tabela 16.

No cenário com uso do modelo, foi disparada a propaganda apenas para o público marcado com alto ou médio risco de voltar a comprar. Isso reduziu em aproximadamente 34% o custo com propagandas e alcançou um potencial de vendas que corresponde a quase 75% do potencial da base inteira.

Tabela 16 – Cenários de segmentação de público.

Segmentação com modelo	Valor gasto com propaganda via e-mail	Potencial de venda
Sem uso	R$ 101,50	R$ 24,779.98
Com uso	R$ 67,00	R$ 18,557.65

Fonte: Elaborado pelo autor (2024).

O caso de uso apresentado demonstra que existe ganho financeiro quando usada inteligência estatística para segmentar melhor o público. No exemplo foi utilizado somente um custo estimado de propaganda por e-mail, todavia existem diversas outros meios de marketing que exigem recursos podendo alcançar valores maiores de custo por cliente.

7 CONSIDERAÇÕES FINAIS

A partir dos resultados apresentados conclui-se que o método PWP se mostrou mais adequado para o conjunto de dados. Apesar de falhar no teste de riscos proporcionais, o modelo PWP sem variáveis log apresentou o ajuste mais adequado para todos os casos de quantidade de compras, enquanto a versão com log não estima bem para três eventos.

Além do ajuste, a transformação log também dificulta a interpretação dos parâmetros, quando comparado com o modelo original. Por esses motivos, o modelo escolhido para estimar a probabilidade de novas compras é o PWP com as variáveis: “media_preco”, “categoria” e “contagem_views” sem transformação logarítmica.

As informações de quantidade de visualizações no site e média do preço por produto gasto se mostraram significativas para o modelo, ainda que o impacto no risco seja baixo. Isso indica que as variáveis disponíveis para modelagem não possuem poder preditivo suficiente para ter alto impacto na discriminação de risco.

Tendo em vista que a base de dados não contém um volume grande de eventos recorrentes, foi necessário manipular as informações para aplicar os modelos semi-paramétricos estudados. Melhores ajustes poderiam ser encontrados incluindo novas informações ou utilizando os modelos paramétricos de fragilidade.

REFERÊNCIAS

ANDERSEN, P. K.; GILL, R. D. **Cox's regression model for counting processes: A large sample study**. The Annals of Statistics, 1982, v. 10, p.1100-1120

AKAIKE, H. **A new look at the statistical model identifcation.** IEEE Transactions on Automatic Control, 1974, v. 19, n. 6, p. 716-723.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEC¸ O, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevivência: teoria e aplicações em saúde**. Rio de Janeiro: Fiocruz, 2011. 432p.

COLOSIMO , Enrico Antônio; GIOLO, Suely Ruiz. **Análise de sobrevivência aplicada**. 1. ed. São Paulo, SP: Editora Edgard Blücher Ltda., 2006. 393 p.

COOK, Richard J.; LAWLESS, Jerald F. **The Statistical Analysis of Recurrent Events**. 1. ed. New York, USA: Springer, 2007. v. 403.

COX, D. R. **Partial likelihood**. Oxford, Inglaterra: Biometrika, 1975, v. 62, p.269-276.

GRAMBSCH, Patricia. M.; THERNEAU, Terry. M. **Modeling Survival Data: Extending the Cox Model**. New York, USA: Springer, 2000. p. 350.

KLEINBAUM, David G.; KLEIN, Mitchel. **Survival Analysis**: a Self-Learning Text. 3. ed. Springer, New York, 2012. 700 p.

MOTA, Thiago Santos. **Modelagem de Análise de Sobrevivência com Eventos Recorrentes Aplicada a Dados da Área Médica**. Orientador: Profa. Dra. Luciana Vaz de Arruda Silveira. 2013. 69 f. Dissertação (Mestrado em Biometria) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Botucatu - SP, 2013.

PRENTICE, R. L.; WILLIAMS, B. J.; PETERSON, A. V. **On the regression analysis of multivariate failure time data**. Biometrika, 1981, v. 68, p. 373-379.

THERNEAU, Terry M.; ATKINSON, Elizabeth. Concordance. *In*: **Concordance**. 3.7-0. [*S. l.*], 22 abr. 2024. Software R, p. 20.