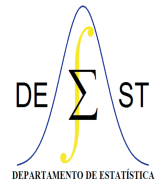




UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA



Regressão Logística Ordinal: Um Estudo Sobre os Níveis de Cuidado Contra a Covid-19

Vinicius Starlino Rolim Ferreira

Ouro Preto-MG
Março de 2023

Vinicius Starlino Rolim Ferreira

Regressão Logística Ordinal: Um Estudo Sobre os Níveis de Cuidado Contra a Covid-19

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientador

Dr. Eduardo Bearzoti

UNIVERSIDADE FEDERAL DE OURO PRETO – UFOP
DEPARTAMENTO DE ESTATÍSTICA – DEEST

Ouro Preto-MG

Março de 2023



FOLHA DE APROVAÇÃO

Vinicius Starlino Rolim Ferreira

Regressão logística ordinal: um estudo sobre os níveis de cuidado contra a covid-19

Monografia apresentada ao Curso de Estatística da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 20 de março de 2023

Membros da banca

Dr. Eduardo Bearzoti - Orientador (Universidade Federal de Ouro Preto)
Dr. Fernando Luiz Pereira de Oliveira - (Universidade Federal de Ouro Preto)
Dra. Graziela Dutra Rocha Gouvea - (Universidade Federal de Ouro Preto)

Professor Dr. Eduardo Bearzoti, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 20/03/2023



Documento assinado eletronicamente por **Eduardo Bearzoti**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/03/2023, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Fernando Luiz Pereira de Oliveira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/03/2023, às 15:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Graziela Dutra Rocha Gouvea**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 28/03/2023, às 13:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0492752** e o código CRC **C1C736ED**.

Ao meu querido avô (in memoriam), que sempre demonstrou orgulho em expressar sua
confiança em mim, especialmente neste momento.

Agradecimentos

À minha mãe, Cláudia Starlino, por ser meu imenso exemplo de força e superação, além de ter feito tudo ao seu alcance para que o caminho até aqui pudesse ser percorrido. Ao meu pai, Vander Rolim, e às minhas irmãs, Vânia e Vanessa, por serem constantes encorajadores e apoiadores dessa trajetória, além de sempre acreditarem em meu potencial.

É com grande satisfação que também expresso minha gratidão aos meus grandes amigos Ronan Abreu, Luíz Diniz, Rogéria Gerçóssimo, Jussara Pereira e Cristiane Martins, por compartilharem bons momentos, conhecimentos e por serem uma parte essencial desta jornada, tornando-a mais agradável e proveitosa.

Por fim, meus agradecimentos aos ilustres professores dos Departamentos de Estatística, Matemática e Computação, pelos valiosos ensinamentos, orientações e contribuições tornando-se fundamentais para minha formação acadêmica.

"Se vi mais longe foi por estar de pé sobre ombros de gigantes."

Sir Isaac Newton.

Regressão Logística Ordinal: Um Estudo Sobre os Níveis de Cuidado Contra a Covid-19

Autor: Vinicius Starlino Rolim Ferreira

Orientador: Dr. Eduardo Bearzoti

Resumo

Diante da crise sanitária mais grave enfrentada pela humanidade desde a criação das Nações Unidas, a pandemia da Covid-19 trouxe repercussões significativas nos aspectos sociais, econômicos e educacionais, culminando em um aumento acentuado do número de óbitos e um risco iminente de colapso do sistema de saúde brasileiro. Para conter a disseminação da doença no meio acadêmico, medidas como a suspensão das atividades presenciais e sua subsequente adaptação remota nas instituições de ensino foram implementadas. Em resposta a esse cenário, a Universidade Federal de Ouro Preto desenvolveu o projeto de pesquisa "Percepções e perspectivas da comunidade universitária na pandemia da Covid-19: o que é importante para o retorno presencial?", que investigou como a pandemia afetou os estudantes e funcionários por meio da aplicação de um questionário eletrônico oferecido a toda a comunidade acadêmica no período de 23 de novembro de 2020 a 31 de janeiro de 2021. Este projeto possibilitou o desenvolvimento deste estudo, que teve como objetivo avaliar os níveis de cuidado contra a Covid-19 adotados pelos respondentes, utilizando estatística descritiva bivariadas e regressão logística ordinal. Com base nos resultados obtidos, observaram-se mudanças significativas nesses níveis ao considerar o gênero feminino, o estado civil casado/união estável e a categoria sem histórico de infecção até o momento da resposta, que foram os principais fatores que influenciaram as respostas em níveis mais elevados de cuidado.

Palavras-chave: Estatística Descritiva Bidimensional, Covid-19.

Ordinal Logistic Regression: A Study on Levels of Care against Covid-19.)

Author: Vinicius Starlino Rolim Ferreira

Advisor: Dr. Eduardo Bearzoti

Abstract

In light of the most severe health crisis faced by humanity since the establishment of the United Nations, the Covid-19 pandemic has brought about significant repercussions in social, economic, and educational domains, culminating in a marked escalation in mortality rates and an impending threat of a breakdown of the Brazilian healthcare system. To curb the dissemination of the disease within academic circles, measures such as the cessation of in-person activities and their subsequent remote adaptation in educational institutions have been implemented. In response to this scenario, the Federal University of Ouro Preto developed the research project "Percepções e perspectivas da comunidade universitária na pandemia de COVID-19: o que é importante para o retorno presencial?", which investigated how the pandemic affected students and employees through the application of an electronic questionnaire offered to the entire academic community during the period from November 23, 2020 to January 31, 2021. This project enabled the development of this study, which aimed to assess the levels of care against Covid-19 adopted by the respondents, using bivariate descriptive statistics and ordinal logistic regression. Based on the results obtained, significant changes were observed in these levels when considering the female gender, married/common-law marital status, and the category without a history of infection up to the time of response, which were the main factors that influenced the responses at higher levels of care.

Keywords: Descriptive Bivariate Statistics, Covid-19.

Lista de figuras

1	Probabilidades acumuladas - Modelos com chances proporcionais	p. 22
2	Distribuição de frequência relativa (%) das concordâncias contidas na Questão 31 por Campus.	p. 34

Lista de tabelas

1	Distribuição de frequências absoluta e relativa (%) por Níveis de cuidado, Campus e Gênero.	p. 31
2	Distribuição de frequência absoluta dos eventos Contraiu Covid e Convive com Pessoas em Risco.	p. 32
3	Distribuição de frequência absoluta por Estado civil e Gênero.	p. 33
4	Estimativas, erros padrões e valores- p referentes aos parâmetros de modelo logístico cumulativo ajustado.	p. 35
5	Qualidade do ajuste do modelo de logitos cumulativos, considerando o Critério de Informação de Akaike (AIC).	p. 36
6	Qualidade do ajuste do modelo de logitos cumulativos, considerando o teste de Pulkstenis-Robinson.	p. 36
7	Avaliação da pressuposição da ausência de multicolinearidade entre as covariáveis consideradas.	p. 37
8	Avaliação da pressuposição de chances proporcionais no modelo ajustado, considerando o Teste de Brant.	p. 38
9	Análise de deviance do Tipo II, para julgar a significância de cada covariável no modelo ajustado.	p. 38
10	Razão de chances e intervalos de 95% de confiança, para o modelo logístico ajustado.	p. 39

Lista de abreviaturas e siglas

SARS-Cov-2 – Severe Acute Respiratory Syndrome Coronavirus 2

Covid-19 – Severe Acute Respiratory Syndrome Coronavirus 2

OMS – Organização Mundial da Saúde

GLM – Modelo Linear Generalizado

Sumário

1	Introdução	p. 13
2	Referencial Teórico	p. 15
2.1	Modelos Lineares Generalizados	p. 16
2.1.1	Componente Aleatório	p. 17
2.1.2	Componente Sistemático	p. 17
2.1.3	Função de Ligação	p. 17
2.2	Regressão Logística	p. 18
2.3	Regressão Logística Ordinal	p. 19
2.3.1	Modelagem Via Logitos Cumulativos	p. 20
2.3.2	Modelagem Via Logitos Cumulativos com Chances Não Propor- cionais	p. 21
2.3.3	Modelagem Via Logitos Cumulativos com Chances Proporcionais	p. 21
2.3.4	Modelo Via Logitos Cumulativos com Chances Proporcionais Par- ciais	p. 23
3	Metodologia	p. 25
3.1	Dados Utilizados	p. 25
3.2	Recursos	p. 26
3.3	Modelagem	p. 26
3.4	Pressupostos	p. 28
3.4.1	Multicolinearidade	p. 28
3.4.2	Chances Proporcionais	p. 28

3.5	Diagnóstico do Ajuste	p. 29
3.5.1	Teste de Pulkstenis-Robinson	p. 30
4	Resultados	p. 31
4.1	Estatística Descritiva	p. 31
4.2	Ajuste do Modelo	p. 34
4.3	Validação de Pressupostos	p. 36
4.4	Análise dos Resultados Finais	p. 39
5	Considerações finais	p. 41
	Apêndice A – Códigos Utilizados	p. 44

1 Introdução

O SARS-Cov-2 é um vírus pertencente à família dos coronavírus e conhecido por provocar a patologia respiratória conhecida como Covid-19. Suas primeiras ocorrências foram constatadas em amostras de lavado broncoalveolar a partir de casos de pneumonia sem precedentes em dezembro de 2019 na cidade de Wuhan, China. Dada sua alta transmissibilidade, no Brasil, o número de casos elevou de, até então 132 suspeitos e 1 confirmação em 27 de fevereiro de 2020, para 52 confirmações e 907 casos suspeitos em 11 de março¹, dia em que a Organização Mundial da Saúde decretou estado de pandemia².

Como consequência, diversas instituições de ensino ao redor do Brasil traçaram novas estratégias de contingência que visaram a prevenção e a redução da curva de contágio deste coronavírus. Dessa forma, uma semana depois, a Universidade Federal de Ouro Preto aprovou a adoção de medidas em seu âmbito a partir de uma reunião extraordinária realizada pelo Conselho Universitário, vindo a suspender todas as suas atividades acadêmicas por tempo indeterminado de acordo com a resolução CEPE N.7.981³.

Ao longo do ano, o cenário no Brasil tornou-se cada vez pior, resultado de uma gestão governamental marcada pela pobreza em ações relacionadas ao gerenciamento, apoio, e principalmente, pela minimização do número de vítimas e disseminação de desinformações, fato que impactou negativamente o tempo até a retomada das atividades presenciais da Universidade. Por outro lado, o Comitê de Enfrentamento ao Coronavírus da UFOP elaborou um Protocolo de Biossegurança baseando-se em ações coletivas e individuais, ao avaliar as condições dos cenários geral brasileiro e específico universitário, possibilitando por fim o planejamento do retorno presencial das atividades acadêmicas nas dependências da Universidade.

Neste contexto, foi elaborado o projeto de pesquisa intitulado “Percepções e perspectivas da comunidade universitária na pandemia de Covid-19 : o que é importante para

¹<https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil>

²<https://www.paho.org/pt/news/11-3-2020-who-characterizes-covid-19-pandemic>

³<https://ufop.br/sites/default/files/cepe7981.pdf>

o retorno presencial?”⁴. Tal projeto foi aprovado pelo Comitê de Ética em Pesquisa da UFOP sob o registro CAAE 39021220.3.0000.5150 e buscou investigar como a pandemia afetou seus estudantes e servidores através da aplicação de um questionário eletrônico disponibilizado a toda a comunidade acadêmica durante o período de 23 de novembro de 2020 a 31 de janeiro do ano seguinte. Dentre as pessoas que participaram da pesquisa, sua grande maioria acreditou no isolamento social, confiou em medidas de proteção indicadas pela OMS, bem como a concordância quanto às ações de contenção da UFOP e aprovação de atividades remotas de estudo/trabalho, considerando a retomada mediante a vacinação em massa de toda a população. A pesquisa trouxe ainda análises sobre questões voltadas aos fatores de risco tanto dos respondentes quanto de pessoas de convivência, além de sentimentos vividos durante a pandemia, estilos de vida, desafios técnicos, contribuindo então para a compreensão das necessidades que a comunidade acadêmica possuía em relação à pandemia e aos reflexos em torno dela através da análise unidimensional dos dados obtidos, sendo sugerida, como proposta futura, a publicação de análises bidimensionais de algumas questões.

Os principais objetivos deste presente estudo visam apresentar, portanto, tais análises propostas, bem como avaliar os níveis de cuidado contra a Covid-19 adotado pelos respondentes. Dessa forma, a utilização de técnicas voltadas à Estatística Descritiva Bivariada e à Análise de Dados Categóricos, por meio da Regressão Logística Ordinal, tornaram-se as mais adequadas ao estudo, uma vez que permitem, respectivamente, avaliar as distribuições conjuntas de pares de questões e modelar o tipo de comportamento levando em consideração a natureza ordinal dos vários níveis da escala de cuidado.

⁴https://saci2.ufop.br/data/pauta/22155_pesquisa_covid19_ufop_resultados_v6.pdf

2 Referencial Teórico

No ano de 1835, o estatístico Adolphe Quetelet sugeriu que populações não poderiam crescer geometricamente ao longo de um grande período de tempo devido aos obstáculos apresentados pela Teoria Malthusiana. Tal publicação influenciou o matemático belga Pierre François Verhulst (1938) a introduzir, pela primeira vez, a função logística como forma de modelar o crescimento populacional humano, animal e vegetal. Paralelamente, esta função também foi descoberta por Pearl e Reed (1920), ao ser desenvolvido um estudo a respeito do crescimento populacional dos Estados Unidos.

A partir desse período, até 1960, esta função tornou-se amplamente conveniente para descrever a relação entre um fenômeno de interesse (variável dependente ou resposta) em função de um ou mais fatores explicativos (conhecidos como variáveis preditoras ou explicativas e frequentemente chamadas de covariáveis), destacando, principalmente, experimentos científicos de Bioensaio.

Bliss (1934ab) considerou a estimação de máxima verossimilhança desenvolvida por Fisher (1912), e introduziu a terminologia *probit*, ou probito (uma abreviação para *Unidade de Probabilidade*, apresentando-a como a inversa da Distribuição Normal Padrão Acumulada), ao analisar a porcentagem de pulgões mortos e a quantidade de nicotina em 100ml de spray. Seus esforços a fim de torná-la bem aceita se estenderam até meados de 1950 através de suas frequentes publicações e pelas análises realizadas por Finney (1947).

Por outro lado, Berkson (1944) levou em consideração a estimação de quadrados mínimos para estabelecer o novo termo *logit*, ou logito, como sendo o inverso da função logística padrão, que guarda uma relação com a razão de chances. O incentivo à sua utilização ocorreu pelos esforços acadêmicos do autor, que também se estenderam ao decorrer das várias décadas seguintes (por exemplo, Berkson, 1951), mas foi pela sua vasta aplicação, vista como vantajosa em relação ao probit, que o logit tornou-se amplamente difundido.

Contudo, foi apenas a partir de 1960, em função da evolução computacional, que grandes avanços ocorreram quanto ao desenvolvimento acerca da modelagem para dados categóricos, propiciando estudos como os de Cox (1966, 1970), e McKelvey e Zanoia (1975), que por sua vez abordam as dificuldades em se aplicar a regressão linear em situações às quais a variável resposta possui uma ordenação natural.

Foi assim sugerido, como alternativa, a extensão do modelo considerando essa natureza ordinal, e assumindo um efeito linear de cada covariável. Uma leitura recomendada é Agresti (2010), que aponta desvantagens significativas, ao realizar esse tipo análise incorretamente, tais como ignorar a variabilidade inconstante das respostas e a de produzir estimativas viesadas para a regressão, resultando em conclusões errôneas. Isso se deve ao fato da pressuposição de variância constante ser violada ao reconhecer que a distribuição dos dados não é normal.

Dessa forma, o modelo logito cumulativo, originalmente proposto por Walker e Duncan (1967), passou a ser conhecido como regressão logística ordinal, sendo sua aplicação o objeto de interesse deste estudo.

2.1 Modelos Lineares Generalizados

Dada a vasta literatura a respeito da regressão linear, seu uso tornou-se comum para a modelagem estatística da grande maioria dos fenômenos aleatórios durante o século XX.

Mesmo sendo esta a principal técnica estatística utilizada, muitos desses fenômenos não atendiam as pressuposições exigidas para sua aplicação. Foi a partir dessas limitações que então surgiram os Modelos Lineares Generalizados, introduzidos por Nelder e Wedderburn (1972), como uma extensão a esta técnica tão utilizada.

Agresti (2013) apresenta de maneira lúcida este conceito; segundo o autor, sua utilização busca generalizar o modelo de regressão linear comum a fim de abordar sua utilização também em situações às quais a resposta não possui a distribuição normal. Quando a variável de interesse corresponde a contagens, por exemplo, torna-se mais adequada a distribuição de Poisson ou binomial negativa; em respostas dicotômicas, a Bernoulli, e em politômicas, tem-se a distribuição multinomial, dentre outras distribuições, como em respostas contínuas assimétricas, onde se utiliza a distribuição gama.

Ainda de acordo com Agresti (2010, 2013), de uma forma geral, um modelo linear generalizado tem como objetivo modelar linearmente a média transformada de uma variável resposta, que, por sua vez, possui uma distribuição pertencente à família exponencial natural. Essa generalização é definida a partir de três aspectos: um componente aleatório, um componente sistemático e uma função de ligação.

2.1.1 Componente Aleatório

O componente aleatório consiste essencialmente na distribuição da variável resposta, considerada como pertencente à família exponencial. Considera-se que o vetor aleatório $Y \in \mathbb{R}^k$ é pertencente à família exponencial de dimensão k , se houver a possibilidade de a função de probabilidade ou função densidade $f(y|\theta) : \theta \in \Theta/\Theta \subset \mathbb{R}^k$ ser expressa na forma:

$$f(y|\theta) = \exp \left\{ \sum_{i=1}^k \eta_i(\theta) T_i(y) - B(\theta) \right\} h(y),$$

onde:

- θ é o vetor de parâmetros do modelo;
- $\eta_1, \eta_2, \dots, \eta_k$ e B são funções reais de θ ;
- T_1, T_2, \dots, T_k e h são funções reais de y .

2.1.2 Componente Sistemático

O componente sistemático, também chamado de preditor linear do modelo, relaciona o vetor $(\eta_1, \eta_2, \dots, \eta_k)$ às variáveis explicativas do GLM através de um modelo linear composto pela combinação linear das k variáveis explicativas a partir da fórmula abaixo:

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, k.$$

2.1.3 Função de Ligação

Este último componente relaciona os componentes aleatório e sistemático através de uma função $g(\cdot)$ monotônica e diferenciável, ao compatibilizar o suporte de variação re-

ferente ao $X\beta$, que corresponde ao intervalo dos números reais (\mathbb{R}), com o suporte de variação de $E(Y|X)$, i.e, o valor esperado $\mu(X)$.

Assim, a função de ligação se dá pela associação de valores esperados de resposta a um ou mais preditores lineares e se dá pela igualdade $\mu_i = \eta_i$ onde $\eta_i = g(\mu_i)$, através da seguinte fórmula:

$$g(\mu_i) = \sum_j \beta_j x_{ij}, i = 1, \dots, k.$$

Dessa forma, quando a função de ligação assume a igualdade $g(\mu) = \mu$, tem-se que $\eta_i = \mu_i$. A função $g(\cdot)$ é caracterizada, neste caso, como uma função identidade, recaindo no modelo de regressão linear simples e normalmente distribuído.

2.2 Regressão Logística

Dentre os modelos lineares generalizados, tem-se o modelo de regressão logística, que por sua vez, admite como variável resposta apenas dados binários.

Assim, seu componente aleatório possui uma distribuição binomial para as respostas, uma vez que observa-se sucessos ou fracassos a respeito da ocorrência do evento de interesse no estudo.

A função de ligação, por sua vez, assume a equação:

$$\text{logit}(\pi) = \log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \alpha + \beta x$$

Por fim, como π é restrito ao intervalo $[0; 1]$, o logito assume quaisquer números reais, formando, portanto, o componente sistemático $\eta_i = \alpha + \beta x$.

Mais detalhes a respeito dessa discussão, e sobre as demais funções de ligação, como probit e a log-log, além de outros modelos para respostas politômicas, podem ser encontrados em Agresti (2010).

Como ilustração, seja um experimento cuja variável resposta binária é representada por Y e seja uma variável explicativa dicotômica X , que abrange os gêneros masculino e feminino. É natural considerar que Y_i apresente uma distribuição de Bernoulli. Aqui, as probabilidades de ocorrência de sucesso e fracasso podem ser denotadas, respectivamente, por $P(Y_i = 1) = \pi_i$ e $P(Y_i = 0) = 1 - \pi_i$.

Seja considerado, então, o evento sucesso condicionado ao gênero. O modelo de regressão logística expressará a ocorrência desse sucesso em termos de chances com base nesse preditor. Como sua probabilidade é $P(Y_i = 1|x_i) = \pi_i = \pi(x_i)$, para o caso masculino ($x_i = 1$) essa chance é dada por:

$$\log \left[\frac{\pi(1)}{1 - \pi(1)} \right] = \alpha + 1 \times \beta$$

A fim de possibilitar a interpretabilidade do modelo de uma forma simples, é por fim considerada a exponencialização em ambos os lados. Tem-se então:

$$\frac{\pi(1)}{1 - \pi(1)} = e^{\alpha + \beta}$$

Analogamente, para o caso feminino, a chance de sucesso é dada por:

$$\frac{\pi(0)}{1 - \pi(0)} = e^{\alpha}$$

A razão entre essas chances (em inglês, *Odds Ratio*) expressada abaixo, pode ser compreendida como o tamanho do efeito da probabilidade de observar o evento de interesse em relação à do outro:

$$\frac{\frac{\pi(1)}{1 - \pi(1)}}{\frac{\pi(0)}{1 - \pi(0)}} = \frac{\pi(1)[1 - \pi(0)]}{\pi(0)[1 - \pi(1)]} \Rightarrow \frac{e^{\alpha + \beta}}{e^{\alpha}} = e^{\beta}$$

Os resultados que podem ser obtidos a partir dessa expressão variam entre 0 e infinito; porém, quanto mais próximo de 1 estiver, menor o efeito da mudança de categoria (gênero). Para o modelo acima, tomando o gênero masculino como referência, a chance esperada de ocorrência do sucesso no experimento é e^{β} vezes (maior ou menor) daquela em relação ao gênero feminino.

2.3 Regressão Logística Ordinal

Existem diversas situações às quais a variável resposta é politômica, ou seja, possui três ou mais categorias naturalmente classificadas como nominais ou ordinais. Nessas circunstâncias, a regressão logística multinomial ou ordinal são, respectivamente, úteis para a modelagem (Agresti, 2010).

A primeira é utilizada definindo uma categoria de referência e observando o efeito

das covariáveis, que mudam a probabilidade da categoria de resposta, sendo comparada com essa categoria referencial ou entre si. Assim, o logito para este modelo corresponde ao logaritmo das chances de ocorrência de uma determinada categoria em função da categoria de referência.

Porém, quando há uma ordenação natural dessas categorias, a regressão logística multinomial pode não ser a mais informativa. Neste caso, pode ser considerado o modelo de regressão logística ordinal.

2.3.1 Modelagem Via Logitos Cumulativos

Segundo Agresti (2010, 2013), vários modelos podem ser utilizados para uma modelagem via logitos cumulativos aplicáveis a dados ordinais:

- *Modelo com Chances Proporcionais*: Consiste em uma modelagem seguindo o pressuposto de chances proporcionais;
- *Modelo com Chances não Proporcionais*: Consiste em um ajuste de modelo considerando a violação da suposição de chances proporcionais;
- *Modelo com Chances Proporcionais parciais*: Considera-se que uma parte das covariáveis viola o pressuposto e parte não, cabendo uma modelagem que leve em conta ambos os casos.

Seja um experimento qualquer, cuja variável resposta Y possui c categorias ordinais, sendo suas respectivas probabilidades definidas por $\pi_1, \pi_2, \dots, \pi_c$. Os logitos cumulativos são definidos por:

$$\begin{aligned} \text{logit}[P(Y \leq j)|x] &= \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \\ &= \log \frac{\pi_1(x) + \pi_2(x) + \dots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \dots + \pi_c(x)}, \quad j = 1, 2, \dots, c - 1. \end{aligned}$$

Neste modelo, cada logito cumulativo utiliza todas as c categorias de resposta, considerando a probabilidade de ocorrer a j -ésima categoria como sendo $\pi_j(x)$.

2.3.2 Modelagem Via Logitos Cumulativos com Chances Não Proporcionais

Quando o pressuposto de chances proporcionais é violado, ou seja, quando há uma violação da suposição de que o efeito de cada variável explicativa é o mesmo para logitos de diferentes probabilidades cumulativas, o modelo com chances não proporcionais torna-se apropriado para a modelagem, uma vez que haverá apenas um parâmetro β_{ij} para cada preditor. Este modelo pode ser expresso da seguinte forma reduzida:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j + \beta'_j x_i, \quad j = 1, 2, \dots, c - 1.$$

A diferença deste modelo se encontra no vetor de parâmetros $\beta_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp})'$. Dessa forma, o modelo em sua integralidade é expresso abaixo:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \dots + \beta_{jp}x_p, \quad j = 1, 2, \dots, c - 1.$$

Contudo, uma desvantagem de seu uso se refere ao princípio da parcimônia, a partir do qual pode-se dizer que, mesmo com menor viés ao se estimar as devidas proporções individualmente, há menos parcimônia na modelagem, à medida que os parâmetros aumentam em quantidade.

2.3.3 Modelagem Via Logitos Cumulativos com Chances Proporcionais

O modelo com chances proporcionais foi descrito por McCullagh (1980). Porém, Agresti (2010) mostrou que há várias outras maneiras para sua aplicação quando há uma estrutura ordinal dos dados, tornando o termo *chances proporcionais* vago, devido à existência de outros modelos logitos também aplicáveis à estrutura de chances proporcionais.

Dessa forma, tornou-se conhecido como Modelo Logito Cumulativo com Chances Proporcionais (*proportional odds version of the cumulative logit model*). Este modelo é definido por:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad j = 1, 2, \dots, c - 1.$$

E nesta modelagem, os interceptos representados por a_j , satisfazem a condição $a_1 < a_2 < \dots < a_j$, devido às probabilidades cumulativas $P(Y \leq j)$, enquanto os parâmetros desconhecidos e os coeficientes de regressão podem ser representados, respectivamente, pelos vetores, de tamanho p , $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$ e $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$.

É válido enfatizar que o intercepto a_j varia para cada logito, e há um efeito comum quanto ao coeficiente angular β , fazendo com que as curvas de probabilidades cumulativas se tornem semelhantes, como mostrado na figura 1.

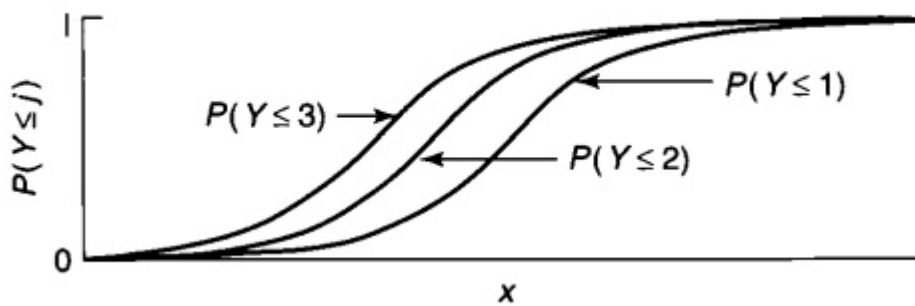


Figura 1: Probabilidades acumuladas - Modelos com chances proporcionais
Fonte: Agresti, 2010, 8.2.2.

Como este independe do índice j , a relação entre x e Y é não dependente da categoria em questão. Pode-se, ainda, reescrever esta equação com a seguinte parametrização alternativa:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j - \beta'x_i, \quad j = 1, 2, \dots, c - 1.$$

Diferentemente do modelo anterior, este implica em uma única estimativa da razão de chances para todas as categorias comparadas, o que faz com que a maximização da função de verossimilhança, bem como a interpretação do modelo, se tornem menos complexas, uma vez que há um número menor de parâmetros. Essa função de verossimilhança é expressa por:

$$\begin{aligned} L &= \prod_{i=1}^n \left[\prod_{j=1}^c \pi_j(\mathbf{x}_i)^{y_{ij}} \right] = \prod_{i=1}^n \left\{ \prod_{j=1}^c [P(Y \leq j | \mathbf{x}_i) - P(Y \leq j - 1 | \mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^n \left\{ \prod_{j=1}^c \left[\frac{\exp(\alpha_j + \beta^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \beta^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \beta^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \beta^T \mathbf{x}_i)} \right]^{y_{ij}} \right\}, \end{aligned}$$

onde $y_{ij} = 1$ em casos aos quais a resposta do indivíduo i , $i = 1, \dots, n$, pertence à categoria j , com $j = 1, \dots, c$, e $y_{ij} = 0$ caso contrário, considerando $\sum_{j=1}^c y_{ij} = 1$.

Como este modelo assume chances proporcionais, torna-se necessário testar a hipótese de que os efeitos das variáveis dependentes não diferem entre logitos, ou seja, $\beta_j = \beta \forall j = 1, \dots, c - 1$. Uma maneira de avaliar este pressuposto é considerar a aplicação do teste da razão de verossimilhanças (TRV), avaliando as hipóteses:

$$\begin{cases} H_0 : \beta_j = \beta \\ H_1 : \beta_j \neq \beta \quad j = 1, \dots, c - 1 \end{cases}$$

e considerando a seguinte estatística de teste:

$$\Delta = -2 \log \left[\frac{L_{H_0}}{L_{H_1}} \right]$$

onde L_H é a função de verossimilhança. Essa estatística de teste segue uma distribuição aproximadamente igual à uma qui-quadrado, ou seja, $\Delta \cong \chi_{(gl, 95\%)}^2$ sob a hipótese nula H_0 , onde gl são os graus de liberdade obtidos a partir da diferença entre o número de parâmetros dos modelos considerando as hipóteses H_0 e $H_1 : \beta_j \neq \beta$.

2.3.4 Modelo Via Logitos Cumulativos com Chances Proporcionais Parciais

Proposto por Peterson e Harrell Jr. (1990), este modelo aborda situações em que uma ou mais covariáveis violam a suposição de chances proporcionais. Quando este fato ocorre, o que é comumente verificado em diversas situações, a abordagem parcial pode se tornar útil, uma vez que é levado em consideração a suposição para um subgrupo de covariáveis caracterizado por atender o pressuposto e uma abordagem apropriada às demais. Destacam-se duas abordagens: modelos restritos e os sem restrição.

Para o modelo sem restrição, é acrescentado um coeficiente γ_i , correspondendo ao efeito associado ao i -ésimo logito cumulativo, abordando assim as variáveis explicativas que não atendem o pressuposto. Este modelo pode ser expresso da seguinte forma:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j + (\beta' + \gamma'_j)x_i, \quad j = 1, 2, \dots, c - 1.$$

Assim, considerando a violação da suposição de proporcionalidade em q variáveis, pode-se reescrever a equação acima na seguinte forma:

$$\text{logit}[P(Y \leq j)|x] = \alpha_j + (\beta_1 + \gamma_{1j})x_1 + \dots + (\beta_q + \gamma_{qj})x_q + (\beta_{q+1})x_{q+1} + \dots + \beta_p x_p, \quad j = 1, 2, \dots, c-1.$$

Também são estimados para este modelo $c-1$ interceptos, além de p coeficientes compreendendo os vetores de tamanho $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$ e $\gamma' = (\gamma_1, \gamma_2, \dots, \gamma_p)$, enfatizando que, para o subgrupo em que a suposição é atendida, $\gamma_j = 0$.

Dessa forma, o modelo de chances proporcionais sem restrição pode ser visto como um caso particular para o modelo anterior, uma vez que dada uma situação à qual $\gamma_j = 0 \forall j$, obtêm-se o modelo de chances proporcionais.

Por outro lado, estes mesmos autores também propuseram o caso especial em que parte do modelo satisfaz certas restrições, adequando-o a situações em que tendências decorridas de uma relação não proporcional entre uma covariável e a variável resposta são observadas, sendo incorporada ao modelo acima da seguinte maneira:

$$\begin{aligned} \text{logit}[P(Y \leq j)|x] = & \alpha_j + \tau_j [(\beta_1 + \gamma_{1j})x_1 + \dots + (\beta_q + \gamma_{qj})x_q] \\ & + (\beta_{q+1})x_{q+1} + \dots + \beta_p x_p, \quad j = 1, 2, \dots, c-1. \end{aligned}$$

Neste caso, τ (τ_j) é um escalar específico para cada logito, e é multiplicado por γ_j , que por sua vez é independente dos pontos de corte (níveis críticos nos quais os eventos começam a mudar significativamente). Essas restrições caracterizam o modelo de chances proporcionais parciais restrito, e visam portanto, a interpretação dessa relação linear entre cada chance proporcional e a variável resposta.

3 Metodologia

3.1 Dados Utilizados

Os dados utilizados neste presente trabalho foram obtidos de forma primária a partir do projeto de pesquisa intitulado “Percepções e perspectivas da comunidade universitária na pandemia de Covid-19: o que é importante para o retorno presencial?”. Este projeto teve sua aprovação efetivada pelo Comitê de Ética em Pesquisa da Universidade Federal de Ouro Preto e foi desenvolvido pela equipe composta por Mariza A. Costa Pena, Gustavo Meirelles Ribeiro, Víctor Morelli A. Barbosa, Rondon Marques Rosa e Eduardo Bearzoti. O objetivo foi avaliar como a pandemia afetou estudantes e servidores desta comunidade acadêmica através da aplicação de um questionário eletrônico composto por 39 perguntas durante o período de 23 de novembro de 2020 a 31 de janeiro do ano seguinte. Cinco delas foram então selecionadas para compor o objetivo deste estudo:

- *Gênero;*
- *Estado Civil;*
- *Você contraiu a COVID-19 até o momento?;*
- *Neste momento de pandemia você convive com alguém que apresenta algum fator de risco/comorbidade para a COVID-19?;*
- *Em relação aos cuidados de isolamento social e as recomendações da OMS em relação ao uso de máscaras, higienização das mãos e alimentos, distanciamento etc, você se considera.*

As 4 primeiras questões foram apresentadas respectivamente como: *Genero*, *Estado_civil*, *Contraiu_covid* e *Convive_fator_risco*. Já o nível de cuidado adotado foi representado por *lvl_cuidado* e tratado como a variável dependente. Adicionalmente, foi elaborado de modo a seguir uma Escala Likert de 5 pontos, onde a pontuação mínima (1) se referiu

aos respondentes que se consideraram nada cuidadosos e a pontuação máxima (5) indicou aqueles extremamente cuidadosos.

Houve o registro de 1151 respostas, das quais 1082 foram validados a partir de critérios de exclusão mantendo apenas aquelas pertencentes aos indivíduos que concordaram em participar da pesquisa, bem como apenas uma resposta em casos onde verificou-se duplicidade.

3.2 Recursos

Para validar toda a modelagem, análises e pressupostos nas próximas seções, foi utilizado a linguagem **R** para computação estatística em sua versão 4.2.2, juntamente com o ambiente de desenvolvimento integrado Posit (RStudio v: 2022.12.0+353).

O modelo de regressão logística ordinal foi ajustado através da função **polr(.)** pertencente à biblioteca **MASS** (Brian Ripley e outros, 2022). Esta função ajusta um modelo estimando seus parâmetros através da estimativa de máxima verossimilhança.

Para a avaliação da qualidade do ajuste, utilizou-se a função **pulkroben**, pertencente à biblioteca **gofcat** (r Ejike R. Ugba, 2022).

Para avaliar o pressuposto de ausência de multicolinearidade, foi criado um modelo de regressão linear simples, de forma a considerar a variável resposta como numérica, para que as variáveis independentes pudessem ser avaliadas através da função **vif(.)** da biblioteca **car** (Fox J., Weisberg S., 2019).

Por fim, para a avaliação do pressuposto de chances proporcionais, foi utilizada a biblioteca **brant** (Benjamin S. and Marco S., 2020).

Todos estes pressupostos e avaliação serão discutidos em detalhes nas próximas seções e os resultados obtidos a partir destas serão apresentados à frente.

3.3 Modelagem

Seja Y a variável resposta do modelo de regressão logística ordinal deste trabalho. Esta mensura o grau de cuidado contra a covid-19 de um indivíduo respondente do questionário e assume os valores contidos no conjunto $\{1, 2, 3, 4, 5\}$, conforme a seguinte escala:

- $Y = 1$: Pessoa nada cuidadosa. Não há qualquer tipo de adoção das medidas preven-

tivas contra a covid-19, como uso de máscara, distanciamento social ou higienização das mãos e alimentos.

- $Y = 2$: Pessoa pouco cuidadosa. A pessoa respondente adota algumas medidas preventivas, mas de forma inadequada ou insuficiente, como uso incorreto de máscara, distanciamento social parcial ou segue parcialmente as recomendações da OMS.
- $Y = 3$: Pessoa moderadamente cuidadosa. Há a adoção pela maioria das medidas preventivas contra a covid-19, mas pode haver falha em algumas situações ou circunstâncias, como uso de máscara inadequada ao tipo de exposição, distanciamento social relativo ou higienização das mãos incompleta.
- $Y = 4$: Pessoa muito cuidadosa. Todas as medidas adotadas seguem as recomendações sanitárias da OMS, como uso correto e adequado de máscara, distanciamento social absoluto e higienização das mãos de forma frequente e completa.
- $Y = 5$: Pessoa extremamente cuidadosa. Todas as medidas preventivas são adotadas seguindo as recomendações sanitárias e ainda promove a conscientização aos demais.

Sejam, ainda, as seguintes covariáveis:

- Gênero: Masculino ou Feminino;
- Estado civil: Solteiro(a), Casado / União estável ou Outro;
- Contraiu covid: Sim ou Não;
- Convive com pessoas em risco: Sim ou Não.

Levando em consideração o modelo logit descrito na seção 2.3.3, a variável dependente e as covariáveis, obtém-se o seguinte modelo:

$$\begin{aligned} \text{logit}[P(Y \leq j)|x] = & \alpha_j + \beta_1 * \text{Genero} + \beta_2 * \text{Estado_civil} + \beta_3 * \text{Contraiu_covid} + \\ & + \beta_4 * \text{Convive_fator_risco}, \end{aligned}$$

sendo α_j o intercepto pertencente ao j -ésimo logito, e β_i , $i = 1, \dots, 4$ os coeficientes de regressão referentes a cada variável independente.

3.4 Pressupostos

3.4.1 Multicolinearidade

A multicolinearidade pode se tornar um problema na aplicação da regressão logística ordinal devido ao fato de se tornar difícil separar os efeitos individuais na variável dependente, uma vez que as variáveis preditoras se encontram altamente correlacionadas umas com as outras. Como consequência, pode haver estimativas instáveis e/ou não confiáveis desses coeficientes, ocasionando o aumento da dificuldade em interpretar os resultados do modelo.

No ajuste do modelo, pressupõe-se que as variáveis preditoras não devem ser altamente correlacionadas entre si. Este pressuposto pode ser avaliado utilizando os mesmos métodos de um modelo com covariáveis contínuas ou mistas (categóricas e contínuas), seja através do fator de inflação de variância (VIF) ou índice de condição (CI), utilizadas a fim de medir o grau de multicolinearidade dos dados, promovendo a remoção ou combinação das variáveis altamente correlacionadas. Outras técnicas também podem ser consideradas, como a estimativa de máxima verossimilhança penalizada (Regressão *Ridge* ou Lasso), que possibilita estabilizar as estimativas dos coeficientes na presença de multicolinearidade.

Neste estudo foi utilizada uma forma estendida do VIF, isto é, o Fator de Inflação de Variância Generalizada (GVIF) (Fox e Monette, 1992), onde é feita uma correção conforme os graus de liberdade (g.l) de cada variável preditora, possibilitando comparações entre dimensões independentemente do nível categórico referencial. Nos casos em que há apenas um único coeficiente, ou seja, g.l = 1, o GVIF se reduz ao VIF usual.

3.4.2 Chances Proporcionais

A pressuposição de que as razões de chances são proporcionais admite que as variáveis independentes são afetadas de maneira constante em cada categoria da variável dependente, ou seja, o efeito é o mesmo ao ser considerada uma mudança na variável independente em relação às chances de ocorrência de uma determinada categoria da variável resposta.

Caso o pressuposto for atendido, os efeitos de cada covariável serão constantes nas chances de resposta de cuidado contra a Covid-19, possibilitando que a razão de chances seja interpretada de uma forma direta, uma vez que apenas um modelo é ajustado para abordar todos os seus níveis. Contudo, como essa suposição nem sempre é satisfeita, outros

modelos podem ser levados em consideração, como os apresentados aqui anteriormente.

Esta avaliação pode ser realizada através do Teste de Brant (Brant, 1990), que por sua vez fornece uma comparação entre um modelo de regressão logística ordinal generalizado e o modelo de chances proporcionais utilizado. Este teste foi elaborado com base no teste de razão de verossimilhanças, comparando o ajuste de um modelo com uma determinada suposição (neste caso a de chances proporcionais) com outro sem essa suposição. A estatística de teste é calculada como a diferença nos logaritmos das verossimilhanças de ambos os modelos, sendo então comparada a uma distribuição qui-quadrado como descrito em 2.3.3. Se o logaritmo da verossimilhança de um modelo corresponde ao quão bem o modelo ajusta os dados, este teste torna-se o mais apropriado para apoio à conclusão acerca da validação da suposição.

Assim, sejam as seguintes hipóteses a serem julgadas:

$$\begin{cases} H_0 : \beta_j = \beta \\ H_1 : \beta_j \neq \beta \quad j = 1, \dots, c - 1 \end{cases}$$

Sob H_0 , tem-se que o modelo possui chances proporcionais. Por outro lado, caso rejeitada esta hipótese, haverá pelo menos uma categoria que difere significativamente da outra, em termos de razões de chances.

3.5 Diagnóstico do Ajuste

Uma vez que são atendidos os pressupostos e selecionado o modelo, resta avaliar a qualidade do ajuste realizado, ou seja, o quão bem ele se ajusta aos dados.

Existem estatísticas de qualidade de ajuste tradicionalmente utilizadas, sendo duas delas: a estatística qui-quadrado de Pearson, baseada nos resíduos de Pearson, e a estatística da razão de verossimilhanças, baseada nas *deviances*.

Outra medida de qualidade que também pode ser considerada é o Critério de Informação de Akaike (AIC) (por exemplo, Giolo, 2017), que visa equilibrar o ajuste do modelo e sua complexidade, penalizando modelos com muitos parâmetros, visando modelos mais parcimoniosos.

Neste estudo, foram aplicados os testes Qui-Quadrado e Deviance de Pulkstenis-Robinson para avaliação da qualidade do ajuste proposto, e aplicação do AIC para fins de comparação com o modelo nulo (modelo estatisticamente mais simples possível de ser

explicado para um conjunto de dados).

3.5.1 Teste de Pulkstenis-Robinson

Em um artigo publicado em 2004, Pulkstenis e Robinson propuseram uma adaptação das estatísticas de testes χ^2 de Pearson e G^2 (razão de verossimilhanças). Nesta adaptação, tabelas de frequências observadas e estimadas são geradas a partir do agrupamento de observações com base na combinação de covariáveis categóricas, considerando pontuações ordinais medianas desses padrões.

Dessa forma, as estatísticas de teste podem ser obtidas utilizando essas frequências através das fórmulas abaixo:

$$\chi^2 = \sum_{h=1}^2 \sum_{i=1}^c \sum_{j=1}^r \frac{(O_{hij} - \hat{E}_{hij})^2}{\hat{E}_{hij}}$$

$$G^2 = 2 \sum_{h=1}^2 \sum_{i=1}^c \sum_{j=1}^r \log \left(\frac{O_{hij}}{\hat{E}_{hij}} \right) O_{hij},$$

onde c representa o número de possíveis combinações formados a partir das covariáveis, r o número de categorias pertencentes à variável resposta, e k o número de variáveis categóricas consideradas no modelo. Estas estatísticas de testes, sob H_0 , se aproximam assintoticamente de uma distribuição χ_{gl}^2 , com $g.l. = (2c - 1)(r - 1) - k - 1$.

Estas estatísticas permitem julgar as seguintes hipóteses:

$$\begin{cases} H_0 : & \text{O modelo proposto é adequado (não há falta de ajuste)} \\ H_1 : & \text{O modelo proposto não é adequado (há falta de ajuste)} \end{cases}$$

O teste acima leva em consideração a hipótese nula de que não há falta de ajuste no modelo, ou seja, sob esta hipótese, o modelo é adequado.

4 Resultados

Os códigos utilizados para obtenção dos resultados de modelagem, avaliação da qualidade do ajuste e validação de pressupostos aqui apresentados, podem ser acessados no Apêndice A. As interpretações deste ajuste foram realizadas mediante as validações dos pressupostos, se encontrando portanto, ao final deste capítulo.

4.1 Estatística Descritiva

Em um primeiro momento foi realizada uma análise descritiva de algumas perguntas pertinentes ao estudo. Para tal, houve a criação da variável *Campus*, estabelecida de acordo com as respostas obtidas com a Unidade Acadêmica/Área Administrativa. Foram também avaliados: o gênero, estado civil e a ocorrência ou não dos eventos “Contrair covid até o momento da resposta” e “Conviver com pessoas em risco”.

A Tabela 1, apresentada abaixo, contém frequências observadas para cada nível de cuidado, respectivos totais marginais bem como suas percentagens.

Tabela 1: Distribuição de frequências absoluta e relativa (%) por Níveis de cuidado, Campus e Gênero.

	1 (N=4)	2 (N=13)	3 (N=169)	4 (N=474)	5 (N=422)	Overall (N=1082)
Campus						
João Monlevade	0 (0%)	2 (15.4%)	13 (7.7%)	37 (7.8%)	38 (9.0%)	90 (8.3%)
Mariana	2 (50.0%)	0 (0%)	34 (20.1%)	99 (20.9%)	89 (21.1%)	224 (20.7%)
Ouro Preto	2 (50.0%)	11 (84.6%)	122 (72.2%)	338 (71.3%)	295 (69.9%)	768 (71.0%)
Genero						
Masculino	4 (100%)	7 (53.8%)	63 (37.3%)	194 (40.9%)	141 (33.4%)	409 (37.8%)
Feminino	0 (0%)	6 (46.2%)	106 (62.7%)	280 (59.1%)	281 (66.6%)	673 (62.2%)

Percebe-se que, dos 1082 respondentes, 768 estavam localizados no campus da cidade de Ouro Preto. Este valor representa 71% do total, sendo que a frequência observada

de respondentes quanto aos cuidados contra a Covid-19 se concentram nos dois níveis máximos, fato observado também em todas as demais categorias de campus e gênero. Há portanto um indicativo de bons cuidados de forma geral, sendo este justificado por pouco mais de 82% em presença nas escolhas dos dois níveis mais altos.

Percebe-se ainda que houve mais mulheres dispostas a responder o questionário, representando 62,2% do total. Estas demonstraram ser mais cuidadosas de uma forma geral, com níveis mais altos de cuidado em relação ao gênero masculino, invertendo apenas quando as respostas se voltam aos considerados pouco ou nada cuidadosos.

Dos 17 respondentes representando os níveis 1 e 2, 11 são masculinos e, exclusivamente, 4 pertencem aos campus de Ouro Preto e Mariana e, dentre as 6 mulheres, todas pertencem aos campus de Ouro Preto e/ou João Monlevade, e nenhuma se considerou nada cuidadosa.

Este indicativo de nível de cuidado pode contribuir para a modelagem, aumentando as chances de resposta a níveis mais altos quando se altera o gênero masculino (aqui tomado como o de referencia) para o feminino, restando verificar se esta suposição se mantém no ajuste da regressão logística.

Uma visão geral a respeito das outras covariáveis pode ser obtida a partir das Tabelas 2 e 3. Estas permitem vislumbrar que a grande maioria dos respondentes não havia contraído o vírus até o momento de resposta independentemente de conviverem ou não com outras pessoas que possuem fatores de risco, além de que, majoritariamente, os indivíduos definiram em sua resposta o estado civil como solteiro(a).

Assim, a partir da Tabela 2, percebeu-se que mais da metade dos respondentes convive com pessoas em risco e não haviam contraído o vírus até o momento.

Tabela 2: Distribuição de frequência absoluta dos eventos Contraiu Covid e Convive com Pessoas em Risco.

Contraiu Covid	Fator de Risco	
	Sim	Não
Sim	56	40
Não	668	318

Já a Tabela 3, percebe-se que houve a predominância de solteiros em ambos os gêneros, mas que a maioria pertence ao feminino.

Tabela 3: Distribuição de frequência absoluta por Estado civil e Gênero.

Estado Civil	Gênero	
	Masculino	Feminino
Solteiro(a)	312	536
Casado(a)/união estável	85	112
Outro	12	25

A Figura 2, apresentada a seguir, foi criada também com base na variável *Campus* e aborda conjuntamente o nível de concordância com determinados tópicos formulados no âmbito do projeto de pesquisa. Em todos os 6 considerados, houve pelo menos 62% de concordância total ou parcial em cada um dos campi. O segundo, que vincula as recomendações da OMS à eficiência contra contágio, apresentou os maiores percentuais de concordância. Por outro lado, destacaram em discordância este último, com 19% em Mariana, 20% em João Monlevade e 22% em Ouro Preto.

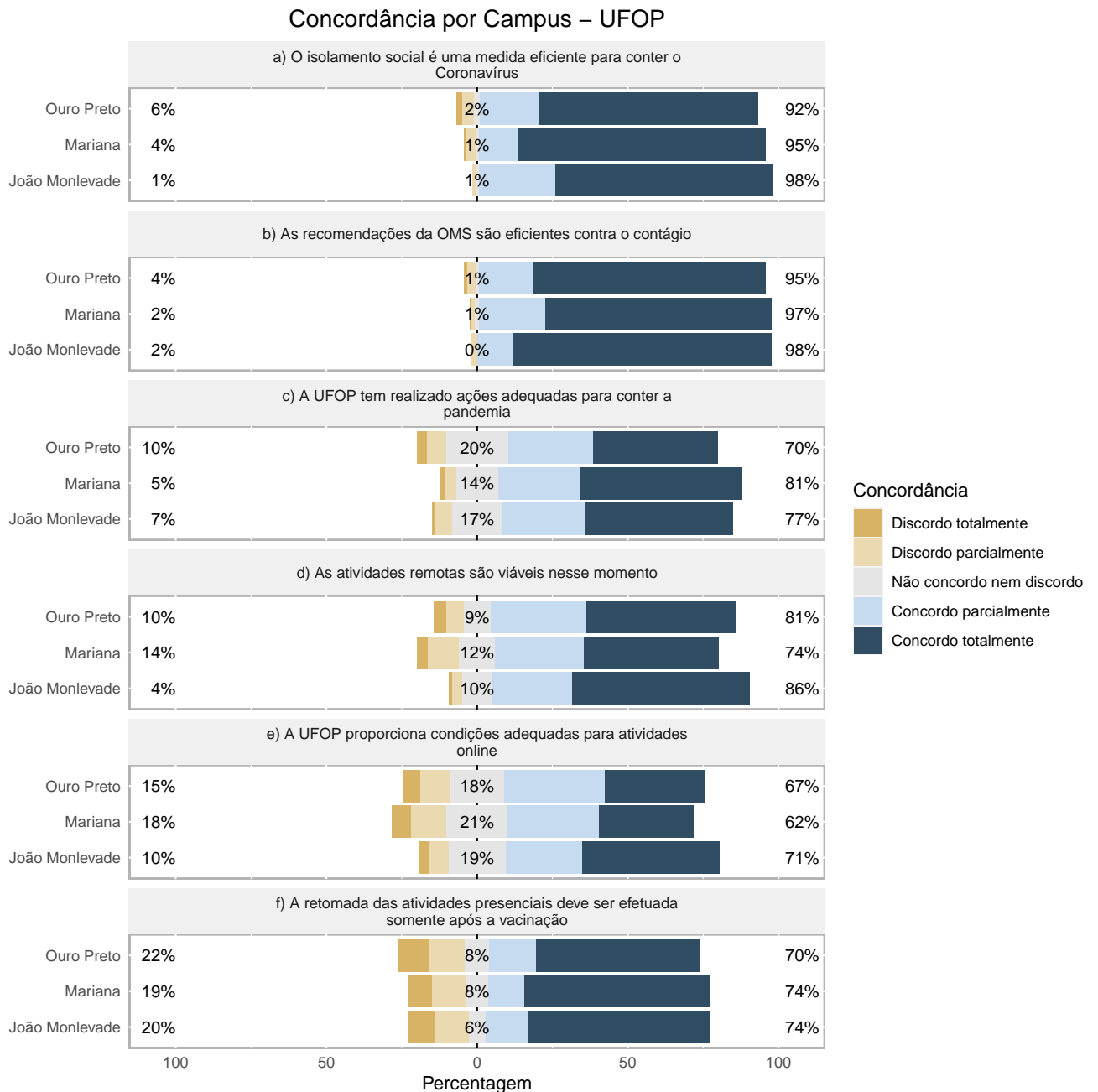


Figura 2: Distribuição de frequência relativa (%) das concordâncias contidas na Questão 31 por Campus.

4.2 Ajuste do Modelo

Em seguida foi ajustado um modelo logístico cumulativo, tendo o nível de cuidado como variável dependente em função de diferentes covariáveis.

A Tabela 4 apresenta as estimativas tanto dos coeficientes da regressão como dos interceptos, também conhecidos como pontos de corte. Estes cortes representam a divisão intervalar quanto aos níveis de cuidado, fazendo com que o gênero, estado civil e demais

covariáveis sejam utilizadas para modelar a probabilidade da variável dependente estar em cada um desses intervalos.

Tabela 4: Estimativas, erros padrões e valores- p referentes aos parâmetros de modelo logístico cumulativo ajustado.

Variável	Parâmetro	Estimativa	Erro Padrão	Valor- p
Intercepto 1	α_1	-4.80	0.54	<0.001
Intercepto 2	α_2	-3.34	0.31	<0.001
Intercepto 3	α_3	-0.75	0.22	0.001
Intercepto 4	α_4	1.31	0.22	<0.001
Genero: Feminino	X_1	0.25	0.12	0.037
Estado_civil: Casado(a)/união estável	X_{12}	0.38	0.15	0.011
Estado_civil: Outro	X_{22}	0.22	0.31	0.488
Contraiu_covid: Não	X_3	0.74	0.20	<0.001
Convive_fator_risco: Não	X_4	-0.18	0.12	0.146

Segundo a notação utilizada nesta Tabela 4, tem-se o conjunto dos interceptos $\alpha_j, j = 1, \dots, 4$. O intercepto $\alpha_1 = -4.8033$ representa o ponto de corte para os níveis de cuidado 1|2; $\alpha_2 = 1.3080$ é o ponto de corte 2|3, e assim por diante. Nesta Tabela é apresentada também a estimativa do parâmetro referente à covariável *Genero* (x_1), de modo que $x_1 = 1$ se o gênero da pessoa respondente é *Feminino*, e 0, caso contrário.

Já a covariável *Estado_civil* (x_2) corresponde a um fator com 3 categorias, e portanto com 2 graus de liberdade, sendo assim necessários 2 parâmetros, cada qual referente a uma variável *dummy* (também chamada de variável indicadora). Assim, $x_{12} = 1$ se a resposta correspondeu a *Casado(a)/união estável* e $x_{12} = 0$, caso contrário; $x_{22} = 1$ se o *Estado_civil* é *Outro* e $x_{22} = 0$, caso contrário. Assim, para a categoria de referência (*Solteiro*), tem-se tanto $x_{12} = 0$ como $x_{22} = 0$.

Para a covariável *Contraiu_covid* (x_3), $x_3 = 1$ se o respondente não havia contraído covid, e $x_3 = 0$ em caso contrário. Finalmente, para a covariável *Convive_fator_risco* (x_4), tem-se que $x_4 = 1$ se o respondente não convive com pessoas que possuem fator(es) de risco, e $x_4 = 0$ em caso contrário.

Dito isto, o modelo geral estimado pode ser escrito em termos dos logitos:

$$\log \left[\frac{\hat{\pi}(x)}{\hat{\pi}_{j+1}(x)} \right] = \hat{\alpha}_j + 0.2472X_1 + 0.3820X_{12} + 0.2146X_{22} + 0.7392X_3 - 0.1787X_4, \quad j = 1, \dots, 4.$$

O primeiro critério de avaliação da qualidade deste ajuste foi realizado por meio do Critério de Informação de Akaike (AIC) deste modelo em relação ao modelo nulo. Esta avaliação está apresentada na Tabela 5.

Tabela 5: Qualidade do ajuste do modelo de logitos cumulativos, considerando o Critério de Informação de Akaike (AIC).

Modelo	Desvio Residual	AIC
Nulo	2364.44	2372.44
Logitos Cumulativos	2336.45	2354.45

A partir da Tabela 5, percebe-se que houve uma redução do AIC em relação ao modelo nulo, o que indica uma melhor qualidade, favorecendo o Modelo de Logitos Cumulativos, justificando assim a complexidade adicional, em relação ao modelo mais simples. Resta saber se a qualidade deste ajuste também é justificada através do Teste de Pulkstenis-Robinson, descrito no Capítulo anterior. Este outro critério está apresentado na Tabela 6:

Tabela 6: Qualidade do ajuste do modelo de logitos cumulativos, considerando o teste de Pulkstenis-Robinson.

Teste	Estatística de Teste	g.l	Valor- p	ρ
Deviance	59.4	143	1	28.42%
Qui-Quadrado	148	143	0.368	28.42%

Levando em consideração a hipótese nula do teste de que o modelo proposto é adequado, ou seja, de que não há falta de ajuste, observa-se na Tabela 6 altos valores- p , de maneira que, ao nível de 5% de significância, não há evidências que favorecem a rejeição da hipótese nula. Conclui-se, portanto, que a Modelagem Via Logitos Cumulativos está adequada às aplicações utilizadas. Nesta Tabela, o cálculo do número de graus de liberdade foi feito mediante: $(2 \times 19 - 1)(5 - 1) - 4 - 1 = 143$.

4.3 Validação de Pressupostos

Alguns pré-requisitos básicos deste estudo foram tidos como atendidos, como a averiguação se a variável dependente é categórica ordinal e não haver dependência entre observações. A variável resposta (nível de cuidado) é claramente ordinal. No caso da independência, é razoável admitir que as respostas de um dado respondente não influenciaram

nas respostas de outros respondentes. Além disso, foram descartadas as respostas para aqueles casos em que o participante respondeu ao questionário mais de uma vez.

Além destas pressuposições básicas, também foram levados em consideração os pressupostos de ausência de multicolinearidade e o de chances proporcionais.

Para a ausência de multicolinearidade, valores de inflação da variância acima de 5 comumente são considerados como indicação de um nível alto. Assim, essa ausência foi verificada através dos valores relativamente baixos de GVIF, calculados para cada covariável, conforme a Tabela 7.

Tabela 7: Avaliação da pressuposição da ausência de multicolinearidade entre as covariáveis consideradas.

Variável	gvif	g.l	$\frac{1}{\text{gvif}^{2(g.l)}}$
Genero	1.01	1	1.00
Estado_civil	1.02	2	1.00
Contraiu_covid	1.01	1	1.01
Convive_fator_risco	1.02	1	1.01

Por fim, para avaliar o pressuposto de chances proporcionais, foram consideradas as hipóteses abaixo, conforme descrito no Capítulo anterior:

$$\begin{cases} H_0 : \beta_j = \beta \\ H_1 : \beta_j \neq \beta \quad j = 1, \dots, c - 1 \end{cases}$$

Novamente, sob a hipótese nula H_0 , tem-se que o modelo possui chances proporcionais, e, caso rejeitada, pelo menos uma categoria será significativamente diferente de alguma outra, em termos de razões de chances.

De acordo com a Tabela 8, não houve evidências estatísticas que indicassem um julgamento a favor da rejeição da hipótese nula. Estas evidências podem ser verificadas de forma conjunta a partir do primeiro termo “omnibus”, que representa o teste global da suposição de chances proporcionais para todas as variáveis preditoras. E de forma detalhada, ao considerar as mudanças para classes não tomadas como referência. Conclui-se, portanto, que esta suposição foi atendida.

Tabela 8: Avaliação da pressuposição de chances proporcionais no modelo ajustado, considerando o Teste de Brant.

Teste Para	χ^2	g.l	Pr(> χ^2)
Omnibus	-3.955	15	1
Genero: Feminino	6.13	3	0.106
Estado_civil: Casado(a)/união estável	1.66	3	0.647
Estado_civil: Outro	-52.11	3	1
Contraiu_covid: Não	1.45	3	0.695
Convive_fator_risco: Não	2.31	3	0.512

Uma vez que todos os pressupostos foram atendidos, torna-se possível uma interpretação estatisticamente adequada e válida. Dessa forma, destacam-se os efeitos gerais das variáveis preditoras nos níveis de cuidado de acordo com a Tabela 9.

Tabela 9: Análise de deviance do Tipo II, para julgar a significância de cada covariável no modelo ajustado.

Variável	LR (Chisq)	g.l.	Valor-p
Genero	4.37	1	0.037 *
Estado_civil	6.74	2	0.034 *
Contraiu_covid	13.05	1	<0.001 ***
Convive_fator_risco	2.11	1	0.146

Esses efeitos indicam o impacto esperado destas variáveis nas mudanças de chances para níveis mais altos de cuidados e podem ser avaliados através da análise de deviance, baseada no Teste do Tipo II e na estatística de Wald. Testes do Tipo II correspondem a testes parciais onde o efeito de cada covariável é ajustado às demais covariáveis. Esta análise de deviance está apresentada na Tabela 9, na qual é possível perceber especificamente que apenas a variável *Convive_fator_risco* foi não significativa.

A covariável *Estado_Civil* possui 2 graus de liberdade pois foi a única a apresentar mais do que duas categorias (3 classes). Dessas, o parâmetro referente ao estado *Outro* foi não significativo. Ou seja, quando o grupo de referência *Solteiro(a)* é comparado ao grupo *Outro*, a mudança entre estas categorias não esteve vinculada a um aumento ou diminuição da probabilidade do resultado. Já para as demais covariáveis, que continham apenas duas categorias, os valores-*p* da Tabela 9 são os mesmos da Tabela 4. De forma a elucidar os resultados obtidos, estes serão analisados de forma detalhada na próxima seção.

4.4 Análise dos Resultados Finais

É válido ressaltar que os coeficientes da Tabela 4 são dimensionados na escala logarítmica, o que pode dificultar a interpretação. Contudo, os modelos de regressão logística podem ter seus coeficientes transformados para a escala de razões de chances, o que torna sua interpretação simples e direta. Para isso, basta exponenciar as estimativas obtidas, bem como os limites dos intervalos de confiança de Wald. Estas grandezas estão apresentadas na Tabela 10.

Tabela 10: Razão de chances e intervalos de 95% de confiança, para o modelo logístico ajustado.

Característica	OR	95% CI	Valor-p
Genero			0.037
Masculino	-	-	
Feminino	1.280	1.02, 1.61	
Estado_civil			0.034
Solteiro(a)	-	-	
Casado(a)/união estável	1.465	1.09, 1.97	
Outro	1.239	0.68, 2.29	
Contraiu_covid			<0.001
Sim	-	-	
Não	2.094	1.40, 3.13	
Convive_fator_risco			0.150
Sim	-	-	
Não	0.836	0.66, 1.06	

OR = Odds Ratio, CI = Confidence Interval

Dessa forma, de acordo com os intervalos de confiança e os valores-*p* apresentados na Tabela 10, merecem destaque os seguintes resultados:

- **Gênero:** Tomando o gênero masculino como referência e mantendo todas as demais variáveis constantes, uma mudança de categoria para o feminino corresponde a um aumento de 28% na chance de o respondente ter maiores níveis de cuidado contra o vírus, considerando quaisquer dois níveis consecutivos de cuidado. Por exemplo, este aumento ocorre na comparação ‘Nada Cuidadoso(a)’ versus ‘Pouco Cuidado(a)’, ou ainda na comparação ‘Muito Cuidadoso(a)’ versus ‘Extremamente Cuidadoso(a)’.
- **Estado Civil:** Ao nível de 5% de significância, pode-se afirmar que a mudança do estado civil solteiro(a) para o casado(a)/união estável propicia um aumento de cerca de 47% nas escolhas de níveis de cuidado mais elevados.

- **Contraiu Covid:** Para respondentes que não haviam contraído covid até o momento de resposta, sua chance de transição do nível 'Muito Cuidadoso(a)' para a o nível máximo 'Extremamente Cuidadoso(a)' é 2.09 vezes maior quando comparados aos que já contraíram.
- **Convive com Fator(es) de Risco:** A mudança de categoria entre quem convive com pessoas que possuem fator(es) de risco para aqueles que não convivem, não foi estatisticamente significativa, considerando um nível de confiança de 95%.

5 Considerações finais

Com o intuito de ilustrar o ajuste de um modelo de regressão logística ordinal, foi desenvolvido este estudo, avaliando os níveis de cuidado em relação à Covid-19 de participantes de um questionário aplicado no âmbito da Universidade Federal de Ouro Preto.

Entre os modelos propostos nesse contexto de avaliação, optou-se pela modelagem via logitos cumulativos com chances proporcionais, uma vez que os pressupostos necessários foram atendidos. Para avaliar a qualidade desse ajuste, utilizou-se o Critério de Informação de Akaike (AIC), que evidenciou uma melhora em relação ao modelo nulo. Além disso, aplicou-se o Teste de Pulkstenis-Robinson, que por sua vez indicou não haver evidências que favorecessem a rejeição da hipótese de que o modelo utilizado foi adequado.

Aqui, optou-se pela conversão dos resultados em razões de chances, o que permitiu uma interpretação mais clara e direta, viabilizando uma compreensão mais precisa das associações que se estabelecem entre as variáveis examinadas.

Por fim, os resultados evidenciaram que a mudança do gênero referencial masculino para o feminino aumentou em 28% a chance de os respondentes apresentarem níveis mais elevados de cuidado contra o vírus. Além disso, a mudança do estado civil de solteiro(a) para casado(a) ou em união estável aumentou em 47% essas escolhas. E, ainda, participantes que não haviam contraído a Covid-19 apresentaram uma chance 2,09 vezes maior de relatarem níveis mais elevados de cuidado em comparação aos que já haviam contraído.

Este estudo de caso evidenciou como modelos de regressão logística ordinal podem ser extremamente úteis na análise de dados de questionário.

Referências

- Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed., Wiley.
- Agresti, A. (2013). *Categorical Data Analysis*. 3rd Ed., John Wiley & Sons Inc., Hoboken.
- Akaike (1974). Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19: 716723.
- Bacaër, N. (2011). Verhulst and the logistic equation (1838). In: *A Short History of Mathematical Population Dynamics*. Springer, London, 35-39.
- Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, 39(227), 357365.
- Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, 7(4), 327339.
- Bliss, C.I. (1934a). The Method of Probits. *Science*, 79, 38-39.
- Bliss, C.I. (1934b). The Method of Probits. *Science*, 79, 409-410.
- Brant R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46(4): 1171-8.
- Cox, D.R. (1966). Some procedures connected with the logistic qualitative response curve. In *Research Papers in Probability and Statistics (Festschrift for J. Neyman)*, edited by F.N. David (1966). Wiley.
- Cox, D.R. (1970). *Analysis of Binary Data*. London: Methuen.
- Finney, D.J. (1947). *Probit Analysis*. (1st ed.) Cambridge University Press, Cambridge, UK.
- Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messeng. Math.*, v.41, 155-160.
- Fox, John, and Georges Monette. 1992. Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, 87(417): 17883.
- Giolo, S.R. (2017). *Introdução à análise de dados categóricos com aplicações*. 1 Ed., Blucher.
- McCullagh, P. (1980). Regression models for ordinal data. *J.R. Stat. Soc. B*, 42: 109-142.

McKelvey, R. D. , & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1), 103-120.

Nelder, J., Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135 (3), 370-384.

Peterson BL, Hanrrel FE (1990). Partial proportional odds models for ordinal response variables. *Appl Stat*, 36: 205-217.

Pulkstenis, E. and Robinson, T.J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, 23: 999-1014.

Raymond Pearl and Reed, L.J. (1920) On the Rate of Growth of the Population of the USA Since 1790 and Its Mathematical Representation. *Proceeding of the National Academy of Science*, 6, 275-288.

Walker S., and Duncan, D. (1967). Estimation of the probability of an event as a function of several independent variables, *Biometrika*, v.549, 167-179.

APÊNDICE A – Códigos Utilizados

Este apêndice tem por finalidade apresentar os códigos utilizados em Linguagem R para as etapas de modelagem, verificação de ajuste e validação dos pressupostos.

0 - Alterações Necessárias para Aplicação e Validação

```
dados[] <- lapply(X = dados, FUN = factor)
dados$lvl_cuidado <- factor(dados$lvl_cuidado, levels = c(1:5), ordered = T)
```

1 - Ajuste do Modelo Logito Cumulativo com Chances Proporcionais

```
modelo_nulo <- MASS::polr(lvl_cuidado ~ 1, data = dados, Hess = T)

modelo_ordinal <- MASS::polr(lvl_cuidado ~ Genero + Estado_civil +
                             Contrainu_covid + Convive_fator_risco,
                             data = dados,
                             Hess = T)
```

2 - Qualidade do Ajuste do Modelo

```
# AIC (Pode também ser visto utilizando summary())
-2*stats::logLik(modelo_nulo)[1] + 2*modelo_nulo$edf
-2*stats::logLik(modelo_ordinal)[1] + 2*modelo_ord$edf

# Teste de Pulkroben
q_deviance <- gofcat::pulkroben(model = modelo_ordinal, test = "deviance")
q_chisq <- gofcat::pulkroben(model = modelo_ordinal, test = "chisq")
rbind(q_deviance, q_chisq)[,1:4]
```

3 - Pressupostos

```
# Ausência de Multicolinearidade
car::vif(modelo_ordinal)

# Chances Proporcionais
brant::brant(model = modelo_ordinal)
```

4 - Efeitos Gerais e Específicos

```
# Efeitos Gerais
car::Anova(modelo_ordinal,
            type = 2,
            test.statistic = "Wald")

# Efeitos Especificos
lmtest::coefstest(modelo_ordinal)
```

5 - Intervalo de Confiança e Razões de Chances

```
modelo_ordinal %>%
  gtsummary::tbl_regression(exponentiate = T) %>%
  gtsummary::add_global_p() %>%
  gtsummary::bold_labels()
```