Universidade de Brasília
Departamento de Estatística

Modelo de Equações de Estimação Generalizadas: uma aplicação em estudo sobre sintomas de depressão em universitários de Minas Gerais durante a pandemia de Covid-19

Trabalho de Conclusão de Curso apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Silvânia Suely Caribé de Araújo Andrade

Modelo de Equações de Estimação Generalizadas: uma aplicação em estudo sobre sintomas de depressão em universitários de Minas Gerais durante a pandemia de Covid-19

Orientador: Prof. Frederico Machado Almeida

Trabalho de conclusão de Curso apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Resumo

A pandemia de COVID-19 gerou uma uma crise sanitária sem precedentes com impacto prático em diversos campos do cotidiano. O distanciamento social e o fechamento de diversos estabelecimentos incluindo universidades gerou uma maior prevalência de estresse, ansiedade e depressão na população. Para avaliar os impactos psicológicos da pandemia, foi desenvolvido um estudo sobre os principais determinantes da depressão em universitários na Universidade Federal de Ouro Preto (UFOP) por meio do modelo de equações de estimação generalizadas (GEE). O modelo GEE foi utilizado para verificar a associação entre as variáveis independentes e o desfecho de depressão entre os universitários entrevistados. Os resultados obtidos indicaram que as variáveis associadas à depressão entre os universitários foram: sexo feminino, idade inferior a 34 anos, casado (ou com relacionamento), com renda familiar menor/igual a dois salários mínimos e tabagismo. O modelo GEE possui vantagens como a possibilidade de modelar diversos tipos de variáveis resposta e o ajuste aos dados é feito sob estimadores consistentes, gerando resultados confiáveis.

Palavras-chave: COVID-19, Depressão, dados longitudinais, dados correlacionados, GEE.

Abstract

The pandemic of COVID-19 has generated an unprecedented health crisis with practical impact in various areas of everyday life. Social distancing and the closure of various institutions including universities has generated a greater prevalence of stress, anxiety and depression in the population. To assess the negative psychological impacts of panndemia, the present work proposes to study the main determinants of depression in university students at the Federal University of Ouro Preto through the model of generalized estimate equations (GEE). The GEE model was used to verify the association between the independent variables and the outcome of depression among the university students interviewed. The results of indicate that variables associated with depression between universities were: female sex, age under 34 years old, married, with family income less than/equal two minimum wages and smoking. The GEE model has advantages such as the possibility of modeling different types of response variables and the adjustment to the data is made using consistent estimators, generating reliable results.

Keywords: COVID-19, Depression, longitudinal data, correlated data, GEE.

Sumário

1 Introdução

Dados da Organização Mundial da Saúde (OMS) apontam que a proporção de pessoas com transtornos de ansiedade e depressão tem crescido ao redor do mundo, principalmente em países de renda baixa. Após a descoberta do primeiro caso da COVID-19, que aconteceu no final de 2019, na cidade chinesa de Wuhan, o mundo passou a enfrentar uma crise sanitária sem precedentes (WHO, 2020b). A doença foi inicialmente caracterizada como uma pneumonia de etiologia desconhecida, até a identificação genética do vírus que causava a doença. Em maio de 2020, a OMS afirmou que medidas de isolamento social (vulgo *lockdown*) são efetivas para diminuir a transmissão exponencial do vírus da COVID-19. Como forma de desacelerar a propagação do vírus, e reduzir a taxa de mortalidade, o governo chinês decretou diversas medidas para mitigar seu agente etiológico, entre as quais se destaca o isolamento social. Tais medidas foram ratificadas pela OMS, que posteriormente decretou a COVID-19 como uma pandemia em 11 de março de 2020 quando 114 países registraram casos da doença e ocorreram mais de 4 mil óbitos em todo o mundo (WHO, 2020a).

Apesar da comprovada eficácia das medidas de prevenção adotadas, estudos apontam que seus impactos na rotina de vida das pessoas afetaram em muitas dimensões das condições de vida e de saúde inclusive a componente de saúde mental. Por exemplo, o distanciamento social implicou no fechamento dos estabelecimentos de ensino, e consequentemente na suspensão das aulas e trabalho remoto ou suspensão do trabalho, o que pode ter aumentado significativamente na taxa de incidência dos problemas psicológicos em diversos grupos sociais, inclusive os universitários, o que levou à ocorrência de uma maior prevalência de estresse, ansiedade e depressão na população (Barros et al., 2020).

Conforme Rubin e Wessely (2020), a ansiedade na comunidade aumenta quando ocorrem surtos de doenças e o isolamento contribui para a ampliação considerável dessa condição. Os autores descrevem as razões pelas quais o isolamento social potencializa a ansiedade e outras doenças crônicas não transmissíveis na comunidade. Como por exemplo: a determinação do isolamento por parte das autoridades é uma medida que indica a gravidade da situação; a imposição dessa ação diminui a confiança e a segurança das pessoas que estão sob isolamento. O isolamento indica perda de controle individual e provoca a sensação de "estar preso" nas pessoas e se torna mais exacerbada com a separação de membros da família. Os rumores fazem com que as pessoas busquem mais informações mesmo que seja em fontes menos confiáveis.

Os autores acrescentam ainda que tais efeitos podem ser cumulativos e exacerbar a ansiedade e a depressão. A presença de transtornos mentais pode se agravar ou constituir um fator de risco para a presença de doenças crônicas e virais, além de influenciar a

adoção de comportamentos deletérios relacionados à saúde como o hábito de fumar e o consumo abusivo de bebidas alcoólicas. Algumas pesquisas mostram que, estudantes que integraram o estudo no período pandêmico apresentaram níveis significativamente mais elevados de depressão, ansiedade e estresse comparativamente aos que integraram o estudo no período normal (Maia; Dias, 2020).

Com o intuito de analisar o impacto da COVID-19 na saúde mental dos universitários, vários estudos de revisão publicados recentemente permitem perceber os efeitos da quarentena na saúde dos universitários, como Maia e Dias (2020), Brooks et al. (2020), Dodd et al. (2021), Bernardelli et al. (2022), César et al. (2022), entre outros. Ambos os estudos concluíram que é inegável o impacto da pandemia de COVID-19 na vida das pessoas em todo o mundo. Entretanto, os efeitos, principalmente na saúde mental, são diferenciados conforme os grupos populacionais e suas atividades específicas. Dessa forma, entender como a pandemia afeta a saúde mental de estudantes universitários em diferentes contextos, além de contribuir com o conhecimento sobre este tema, possibilita o desenvolvimento de estratégias preventivas e de cuidado em saúde mental para os universitários em situações de grande estresse coletivo como a pandemia de COVID-19.

Em geral, estudos desta natureza se restringem em uma simples análise descritiva do conjunto de dados, ou na aplicação de alguma metodologia estatística adequada para modelar bancos de dados transversais. Alguns dos modelos popularmente utilizados para tal proposito são: o modelo de regressão linear e a regressão logística binária e multinomial. Entretanto, em muitas situações práticas é de interesse do pesquisador avaliar a saúde mental dos estudantes em diferentes pontos do tempo. Desta forma, o presente trabalho propõe estudar os principais determinantes do transtorno da depressão em universitários da UFOP por meio do modelo de equações de estimação generalizadas (GEE-Generalized Estimating Equation, a sigla em inglês).

O modelo GEE é uma escolha aquedada para situações em que a variável de resposta para cada sujeito ou elemento experimental é avaliada repetidamente em vários momentos ou sob várias condições. Estudos envolvendo observações correlacionadas são conhecidos como "estudos longitudinais" e, ocorrem comumente em aplicações relacionadas a área de saúde (Agresti, 2018).

Assim, o objetivo deste trabalho foi analisar os fatores determinantes no aumento dos problemas associados aos sintomas de depressão em egressos da UFOP durante a pandemia de COVID-19 usando o modelo GEE. Especificamente, os objetivos foram elenncados em:

- Descrever a prevalência da depressão entre os universitários em dois momentos da pandemia de COVID-19.
- Apresentar o perfil dos universitários que relataram depressão durante a pandemia.

- Conduzir uma aplicação da metodologia apresentada, no banco de dados referente ao projeto de sintomas de transtorno de ansiedade e depressão em universitários (PADu).
- Selecionar o modelo que apresentar melhor ajuste do conjunto de dados.

2 Revisão da Literatura

Estudos longitudinais são frequentes em diferentes áreas de conhecimento, com particular enfoque nas ciências da saúde e psicologia. A metodologia estatística que permite modelar estes dados desempenha um papel importante no aprimoramento de nossa compreensão sobre o desenvolvimento e persistência de certas doenças. Pois, há muita heterogeneidade natural entre os indivíduos em termos de como estes se desenvolvem e progridem. Essa heterogeneidade se deve a fatores genéticos, ambientais, sociais e comportamentais. Um desenho de estudo longitudinal permite a descoberta de características individuais que podem explicar essas diferenças inter-individuais nas mudanças dos resultados de saúde ao longo do tempo (Fitzmaurice; Laird; Ware, 2012).

O principal aspecto que caracteriza um estudo longitudinal é a repetição das medições nos mesmos indivíduos ao longo do tempo, permitindo assim o estudo direto da mudança ao longo do tempo. Estudos longitudinais têm como objetivo principal a caracterização da mudança na variável resposta ao longo do tempo e os fatores que influenciam na tal mudança. O caso mais simples consiste em verificar se houve uma mudança na variável resposta entre os instantes T0 (conhecido como linha de base) e T1 (primeira onda de avaliação), por meio do teste t-pareado, quando a variável resposta é contínua e a suposição de normalidade dos dados é verificada, ou quando o número de observações é suficientemente grande ($n \geq 30$). No entanto, quando tais critérios não são válidos, alguns testes não-paramétricos ou modelos de regressão mais gerais representam alternativas para modelar dados com tal estrutura (Twisk, 2013).

2.1 Modelos Lineares Generalizados

Antes de descrever uma abordagem mais ampla sobre o tópico dos modelos GEE, é importante apresentar uma breve síntese de uma classe de modelos de extrema importância na literatura Estatística. Tais modelos são conhecidos como modelos lineares generalizados (GLM). De formal geral, os MLG englobam uma ampla classe de modelos de regressão adequados para analisar diversos tipos de respostas univariadas, por exemplo, contínuas, binárias ou contagens (Fitzmaurice; Laird; Ware, 2012; Agresti, 2018).

Os GLM são uma extensão de uma família de modelos de probabilidade para uma variável resposta univariada y. Tal família de distribuições de probabilidade é conhecida como *família exponencial uniparamétrica*. É importante salientar que boa parte das funções de probabilidade comumente consideradas na literatura pertencem à família exponencial. Os MLG podem fornecer previsões para a média da variável dependente a partir de valores possíveis para as variáveis independentes. Todas as distribuições que

pertencem à família exponencial uniparamétrica podem ser expressas da seguinte forma:

$$f(y,\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{2.1.1}$$

onde θ e ϕ são parâmetros escalares, $a(\cdot)$, $b(\cdot)$, também conhecidas como geradoras de cumulantes e $c(\cdot,\cdot)$ são funções reais conhecidas. As funções específicas $b(\cdot)$ e $c(\cdot,\cdot)$ associadas às distribuições da família exponencial uniparamétrica, permitem estabelecer uma distinção de um membro da família exponencial em relação ao outro. Os MLG ajudam a investigar o efeito de um conjunto de variáveis explicativas na variável resposta. As três partes que compõem tal classe de modelos são:

(*i*) Componente aleatória: diz respeito à distribuição de probabilidade da variável dependente, onde cada uma das observações y_i é independente das demais.

(*ii*) Componente fixa (determinística ou sistemática): refere-se ao preditor linear, $\eta_i = \mathbf{X}_i^\top \boldsymbol{\beta}$, para todo $i \in \{1, 2, \cdots, n\}$. O preditor linear é formado pela combinação linear entre um vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ e $X_i = (X_0, X_1, \cdots, X_n)^\top$ um vetor coluna, pertencente à matriz experimental $\mathbf{X}$, com dimensão $n \times p$ de tal forma que:

$$\eta_i = X_i^\top \boldsymbol{\beta} = \sum_{j=0}^{p} \beta_k X_{ij}, \tag{2.1.2}$$

com $X_{i0} = 1$ para todo i, sendo o vetor de 1's incluído em $\mathbf{X}_i$ para acomodar o intercepto.

(*iii*) Função de ligação: é uma função $g(\cdot)$, monótona e diferenciável, aplicada à cada componente da média da variável dependente $\mu_i = E(Y_i)$, segundo sua distribuição de probabilidade (parâmetro natural), relacionando-a à componente fixa (preditor linear) (Agresti, 2018) expressa por:

$$g(\mu_i) = \eta_i = X_i^\top \boldsymbol{\beta} = \sum_{j=0}^{p} \beta_k X_{ij} \tag{2.1.3}$$

A função de ligação $g(\cdot)$ conecta as componentes aleatória e sistemática, respectivamente. Em princípio, qualquer função de ligação $g(\cdot)$ pode ser escolhida para conectar a média da resposta y_i com o preditor linear η_i. No entanto, toda distribuição que pertence à família exponencial uniparamétrica tem uma função de ligação particular, chamada de *função de ligação canônica* (Paula, 2004; Fitzmaurice; Laird; Ware, 2012; Agresti, 2018).

Antes de introduzirmos o modelo de equações de estimação generalizadas, vamos apresentar de forma breve o conceito da função quase-verossimilhança proposta por Wedderburn (1974). Em suma, seja $Y_1, Y_2, \cdots, Y_n$ uma amostra aleatória, tal que, $\mathbb{E}(Y_i) = \mu_i$ e variância $\text{Var}(Y_i) = a(\phi)\nu(\mu_i)$, onde sem perda de generalidade podemos assumir que $a(\phi) = \phi$. A função quase-verossimilhança da observação y_i, é dada por:

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi\nu(t)} dt. \tag{2.1.4}$$

E por consequência, a função quase-verossimilhança para todas as observações da amostra será dada pela seguinte expressão:

$$Q(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{n} Q(\mu_i, y_i). \tag{2.1.5}$$

Desta forma, as estimativas para o vetor de parâmetros desconhecido serão obtidas a partir de:

$$\mathrm{U}_k = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} \nu^{-1}(\mu_i)\,(y_i - \mu_i) = 0, \tag{2.1.6}$$

com k podendo assumir valores $\{0, 1, 2, \cdots, p\}$. Entretanto, uma extensão da equação 2.1.6 será apresentada de forma mais detalhada na seção 2.3.

2.2 Modelos Baseados em Dados Longitudinais

Dados longitudinais ou dados em painel são aqueles provenientes da coleta de uma resposta específica em um conjunto de unidades amostrais, geralmente obtidas através de medidas repetidas ao longo do tempo ou em diferentes condições experimentais. Para modelagem desses dados são necessários modelos que considerem a correlação entre os dados para realizar inferências válidas. Há duas abordagens vastamente utilizadas para modelar dados longitudinais: análise de variância de medidas repetidas e modelo de equações de estimação generalizadas (GEE). A análise de variância de medidas repetidas consiste em não desprezar o pressuposto de independência entre da resposta y_i e modelando explicitamente a estrutura de correlação. Os métodos de estimação e inferência são semelhantes aos dos MLG. A modelagem multinível também é utilizada para análise de dados longitudinais em que a estrutura hierárquica do desenho de estudo é observada e o efeito dos níveis pode ser descrito por meio de parâmetros fixos (efeito de grupo) ou variáveis aleatórias (sujeitos aleatoriamente alocados nos grupos), ou ambos (modelos de efeito

misto), conforme apresentado em Dobson e Barnett (2008), o método de estimação dos MLG é a partir do método da máxima verossimilhança e envolve realizar repetidamente um cálculo de regressão ponderada, baseado em algoritmos numéricos, como é o caso do algoritmo de Newton-Raphson e escore de Fisher (Cordeiro; Demétrio, 2013).

O algoritmo é similar a um processo iterativo de Newton-Raphson, mas a característica principal é o uso da matriz de valores esperados das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança (informação), em relação aos $\boldsymbol{\beta}$, no lugar da matriz correspondente de valores observados. Essa característica foi, primeiramente, desenvolvida por Fisher (1935), para o caso da distribuição binomial com função de ligação "probit" e o processo é denominado método escore para estimação de parâmetros (Cordeiro; Demétrio, 2013). Assim, cada entrada genérica do vetor escore para o contexto dos modelos generalizados (com respostas independentes) são dadas por:

$$\mathrm{U}_k = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mathrm{Var}\,(Y_i)} \frac{\partial \mu_i}{\partial \beta_k} \text{ com } k = 0, 1, \cdots, p. \tag{2.2.1}$$

Para medidas repetidas, y_{ij} é o vetor de respostas para o indivíduo i com $E(Y_i) = \mu_i$, $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ e D_i é uma matriz diagonal contendo as quantidades, $\dfrac{\partial \mu_i}{\partial \beta_k}$, isto é, $D_i = \mathrm{diag}\{\dfrac{\partial \mu_i}{\partial \beta_k}$ com $i = 0, 1, 2, \cdots, p\}$. Sucintamente, o modelo MLG para dados longitudinais acima citado difere do modelo de Equações de Estimação Generalizadas (GEE) na sua função escore que para este último é conhecida função de quase-escore (Dobson; Barnett, 2008):

$$\mathrm{U}_k = \sum_{i=1}^{n} D_i^\top V_i^{-1} (y_i - \mu_i) = 0, \tag{2.2.2}$$

Os elementos da matriz V_i são dados por:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \phi, \tag{2.2.3}$$

onde A_i é a matriz diagonal cujos elementos são $\nu(\mu_i)$ é a função de variância, R_i (ou $R_i(\alpha)$) é a matriz de correlação para y_i e ϕ é um parâmetro para permitir a superdispersão (Dobson; Barnett, 2008). A abordagem marginal, através das GEE, concentra-se na especificação dos momentos dos dados, enquanto na abordagem condicional (usando um Modelo de Efeitos Mistos, MEM), são modeladas as respostas de indivíduos específicos, ou seja, a metodologia GEE é desenvolvida para realizar inferências marginais (modela a distribuição marginal da variável resposta) e, por consequência, não permite a realização de inferências no nível individual (modelos de dados longitudinais como o MEM) (Wakefield, 2009).

2.3 Modelo de Equações de Estimação Generalizadas

O modelo GEE envolve um conjunto de métodos de Inferência Estatística para obter estimativas mais robustas e não viesadas dos parâmetros da modelagem de regressão para dados longitudinais (Liang; Zeger, 1986). Para fazer uma descrição breve, assuma que as medidas de n indivíduos são avaliadas repetidamente ao longo do tempo. Denote por y_{ij} um valor particular da variável resposta para o i-ésimo indivíduo no j-ésimo instante de avaliação, com $j = 1, 2, ..., n_i)$. Quanto a natureza, a variável resposta pode ser contínua, binária, ou de contagem. É importante salientar que a natureza da variável dependente tem implicação importante para a especificação do modelo. No entanto, a notação não distingue entre os diferentes tipos de resposta (Fitzmaurice; Laird; Ware, 2012).

Diferentemente do teste t-pareado onde, para sua aplicação é necessário que se tenha o mesmo número de observações repetidas (dados balanceados), a aplicação do modelo GEE não pressupõe que os sujeitos tenham o mesmo número de medidas repetidas, ou que sejam medidos em um conjunto comum de ocasiões. Portanto, para acomodar o possível desbalanço nos dados, i.e., quando as medidas repetidas não são iguais para todas as unidades experimentais, assume-se que existem n_i medidas repetidas na variável resposta para o i-ésimo sujeito, e que assim sendo, cada y_{ij} é observado a cada instante de tempo t_{ij}. Desta forma, a variável resposta para o i-ésimo indivíduo pode ser agrupada em um vetor coluna de dimensão $n_i \times 1$. Isto é,

$$\boldsymbol{y_i} = (y_{i1}, y_{i2}, ..., y_{in_i})^\top, \quad i = 1, 2, \cdots, n. \tag{2.3.1}$$

Note que, os vetores de respostas são assumidos como sendo independentes um do outro, ou seja, $y_{ij} \perp y_{hj}$, para todo $i \neq h$, e $i, h \in \{1, 2, \cdots, n\}$. Porém, as medidas repetidas em um mesmo indivíduo não são consideradas independentes pois, existe uma estrutura de correlações interna entre elas. De igual forma, associado à cada valor da resposta de interesse y_{ij}, existe um vetor $p \times 1$ dimensional de covariáveis dado por:

$$X_{ij} = (X_{ij0}, X_{ij1}, X_{ij2}, ..., X_{ijp})^\top, \quad i = 1, 2, \cdots, n \text{ e } j = 0, 1, \cdots, p. \tag{2.3.2}$$

Desta forma, estamos assumindo que cada indivíduo tem um vetor de covariáveis X_{ij}, associado com a variável resposta y_{ij} em cada instante de tempo t_j. Até então assumimos que cada indivíduo no estudo tem um vetor de respostas repetidas, denotadas por y_i, e associada a cada uma das medidas repetidas, existe um conjunto de vetores das p covariáves que podem ser agrupadas em uma matriz experimental $\mathbf{X}_i = \Big(X_{ij}\Big)$. Segundo

a explanação apresentada na seção 2.1, o processo de modelagem para dados longitudinais é igualmente compostas por três partes, a saber (Fitzmaurice; Laird; Ware, 2012; Agresti, 2018):

(*i*) A esperança condicional ou média de cada variável resposta, $\mathbb{E}\left(Y_{ij}|X_{ij}\right) = \mu_{ij}$, que depende do vetor de covariáveis, por meio da função de ligação:

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^{\top}\boldsymbol{\beta}, \tag{2.3.3}$$

sendo $\mathbf{x}_{ij}$ um vetor fixado de X_{ij}

(*ii*) A variância condicional de cada y_{ij}, dado o vetor de covariáveis, depende da média por meio da seguinte relação:

$$\text{Var}(Y_{ij}) = \phi\nu(\mu_{ij}), \tag{2.3.4}$$

onde $\nu(\mu_{ij})$ é comumente designado por *função de variância*, i.e., uma função da média, e ϕ é um parâmetro de dispersão, podendo ser conhecido ou não. Para estudos longitudinais balanceados, um parâmetro de dispersão separado, ϕ, pode ser estimado em cada ocasião. Alternativamente, o parâmetro de dispersão pode depender dos instantes de tempo em que as medidas foram coletadas, com $\phi(t_{ij})$ sendo uma função paramétrica de t_{ij}.

(*iii*) A correlação intra-indivíduo condicional ao vetor de respostas repetidas, dadas as covariáveis, é assumida como uma função de um conjunto adicional de parâmetros de associação, α (e também depende das médias, μ_{ij}). Por exemplo, os componentes de α podem representar as correlações pareadas ou o logaritmo das razões de chances entre as respostas repetidas. A correlação intra-sujeito nas respostas é descrita com base em Fitzmaurice, Laird e Ware (2012).

As três especificações para o modelo GEE, que foram apresentadas anteriormente, constituem uma extensão dos MLG para situações envolvendo dados longitudinais. Para o presente estudo foi utilizada a modelagem com GEE, pois, os dados apresentam estrutura de agrupamento por indivíduo (mesmo universitário em dois momentos distintos). O software R (R Core Team, 2022) foi usado para manipulação do banco de dados e aplicação da modelagem com GEE.

2.4 Estimação dos parâmetros no modelo GEE

Nos modelos GEE é importante especificar corretamente a matriz de correlação de trabalho $R_i(\alpha)$ para que o $\widehat{\boldsymbol{\beta}}$ seja consistente e assintoticamente normal. Para definir

a matriz de correlação de trabalho deve ser observado o desenho do estudo e as análises exploratórias sobre as relações entre as variáveis (Agranonik, 2010; Dobson; Barnett, 2008). Esses autores sustentam que a matriz de correlação de trabalho diz respeito a correlação entre as observações de um mesmo grupo ajustadas pelas covariáveis do modelo.

A matriz de correlação de trabalho é uma matriz de variância-covariância cuja diagonal são os únicos elementos com valores diferentes de zero com o pressuposto de que as respostas para diferentes indivíduos são independentes, ou seja, $\mathbf{V} = R_i(\alpha)$, assumindo uma estrutura de correlação independente entre os indivíduos.

$$\mathbf{V} = \begin{pmatrix} V_1 & 0 & 0 \\ 0 & V_2 & 0 \\ & \vdots & \\ 0 & 0 & V_n \end{pmatrix}. \tag{2.4.1}$$

As matrizes V_i tem a mesma composição para todos os indivíduos. Para estimação dos β podem ser utilizados dois métodos quando os elementos da matriz V_i são constantes conhecidas: função de verossimilhança ou por mínimos quadrados. O estimador de máxima verossimilhança pode ser obtido do seguinte modo (Venezuela, 2003; Dobson; Barnett, 2008; Freitas, 2018):

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = X^\top R_i(\alpha)^{-1}(y_i - X\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^\top R_i^{-1}(\alpha)(y_i - X^\top R_i \boldsymbol{\beta}) = 0, \tag{2.4.2}$$

onde o $\ell(\cdot)$ é o logaritmo da função de verossimilhança. Deste modo, o vetor $\widehat{\boldsymbol{\beta}}$ pode ser obtido resolvendo numericamente a equação 2.4.2. De igual forma, a matriz de variância-covariância para $\widehat{\boldsymbol{\beta}}$, é dada por:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (X^\top R_i(\alpha)^{-1} X)^{-1} = (\sum_{i=1}^{N} X_i^\top R_i(\alpha)^{-1} X_i)^{-1}. \tag{2.4.3}$$

A matriz $R_i(\alpha)$ é desconhecida, e portanto, deve ser estimada por meio de processos iterativos como se segue:

1. Iniciar o modelo com uma matriz $R_1(\alpha)$.
2. Estimar os β e os preditores lineares $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}$ e os resíduos $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$.
3. Uma nova matriz de variância-covariância $R(\alpha)$ é calculada por meio das variâncias e covariâncias dos resíduos; e essa matriz é usada para estimar um novo β. Este processo é repetido até que ocorra a convergência do modelo.

Caso a matriz de correlação de trabalho estimada $\widehat{R_i}(\alpha)$ seja substituída pela matriz $R_i(\alpha)$ na estimativa da variância dos $\widehat{\beta}$, provavelmente essa estimativa será subestimada (Dobson; Barnett, 2008). Logo, uma estratégia para evitar essa subestimação, é gerar um estimador mais robusto conhecido como estimador sanduíche de informação que é composto por uma nova estimativa da matriz de correlação de trabalho como dado em:

$$R_s(\widehat{\beta}) = \Im^{-1} C \Im^{-1}, \tag{2.4.4}$$

$$\Im = X^\top \widehat{R_i}(\alpha)^{-1} X = \sum_{i=1}^{N} X_i^\top \widehat{R_i}(\alpha)^{-1} X_i \tag{2.4.5}$$

onde $\widehat{R}_i$ representa a i-ésima submatriz de $\widehat{R}_i(\alpha)$ e $V_s(\widehat{\beta})$ que é chamada de estimador sanduíche de informação devido ao componente $\Im$, que é a matriz de informação, a qual pode ser escrita como:

$$\Im = X^\top W X, \tag{2.4.6}$$

em que W é a matriz diagonal $n \times n$ com elementos:

$$w_{ii} = X^\top \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \tag{2.4.7}$$

Quando se trata de dados que não apresentam uma distribuição normal do vetor de respostas, um processo iterativo é utilizado para estimar o β seguindo o algoritmo abaixo descrito (Venezuela, 2003; Dobson; Barnett, 2008; Freitas, 2018):

1. Primeiramente, usa-se a matriz identidade para a matriz de correlação de trabalho $R_i(\alpha)$ e o parâmetro de superdispersão $\phi = 0$.

2. Os β são estimados pela resolução da equação:

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = X^\top R_i(\alpha)^{-1}(y_i - X\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i^\top R_i^{-1}(\alpha)(y_i - X_i\boldsymbol{\beta}) = 0. \tag{2.4.8}$$

3. Essas estimativas são usadas para calcular valores ajustados de $\mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ e também os resíduos $y_i - \widehat{\mu}_i$. A

4. Os resultados obtidos no passo 3 são usados para estimar os parâmetros de A_i, $R_i(\alpha)$ e ϕ, sendo que A_i é a matriz diagonal com elementos $\mathrm{Var}(Y_i)$

5. A equação de $U(\boldsymbol{\beta})$ é novamente resolvida para estimar $\boldsymbol{\beta}$ até obter convergência.

Salienta-se que existem diversos tipos de matrizes de correlação de trabalho e a correta seleção da mesma de acordo com alguns critérios irá gerar estimativas mais precisas para o modelo. Os modelos GEE fornecem estimativas consistentes mesmo quando a matriz de correlação de trabalho é incorretamente especificada. Entretanto, as estimativas são mais eficientes quando essa matriz é definida corretamente, mesmo em pequenos tamanhos amostrais (Wang, 2014). A seguir é apresentado um quadro com exemplos de matrizes de correlação de trabalho para modelagem GEE. Na Figura 1, é utilizado o M=4 para que a matriz fique maior e seja mais facilmente compreendida do que M=2, como é o caso do presente trabalho.

Figura 1: Matrizes de correlação de trabalho: estrutura, definição, exemplo e número de parâmetros

Estrutura	Definição	Exemplo ($m=4$)	Número de parâmetros
Independente	$Corr(Y_{ij},Y_{ik}) = \begin{cases} 1, se\ j=k \\ 0, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	0
Permutável	$Corr(Y_{ij},Y_{ik}) = \begin{cases} 1, se\ j=k \\ \alpha, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}$	1
AR(1)	$Corr(Y_{ij},Y_{ij+t}) = \alpha^t,$ $t=0,1,2,3$	$R(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$	1
M-dependente	$Corr(Y_{ij},Y_{ij+t}) = \begin{cases} 1, se\ t=0 \\ \alpha_t, se\ t=1,2,...,M \\ 0, se\ t>M \end{cases}$.	$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$ $M=2$	$0<M<m-1$
Não estruturada	$Corr(Y_{ij},Y_{ik}) = \begin{cases} 1, se\ j=k \\ \alpha_{jk}, se\ j \neq k \end{cases}$	$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1 & \alpha_4 & \alpha_5 \\ \alpha_2 & \alpha_4 & 1 & \alpha_6 \\ \alpha_3 & \alpha_5 & \alpha_6 & 1 \end{pmatrix}$	$m(m-1)/2$

Fonte: Agranonik (2010)
Nota: $R(\alpha)$ é a matriz de correlação entre as observações do mesmo grupo
n= número de observações

Em linhas gerais, a estrutura independente é utilizada quando as observações do mesmo sujeito são independentes. Para uma estrutura de correlação permutável, assume-se que todas as variáveis respostas (y_i) são igualmente correlacionadas dentro da mesma unidade primária de amostragem. Na estrutura de correlação usando Modelo Autorre-

gressivo de 1° ordem, digamos AR(1), os elementos externos à diagonal na matriz de correlação de trabalho diminuem com a distância entre as observações em que α_{js} (coeficiente de correlação intra-classe) é estimado como $\alpha_{js} = \alpha^{|j-s|}$ para observações nos tempos t_j e $t_s|$. Na estrutura M-dependente o número de observações dentro do grupos varia de zero até M. Quando todos os elementos da matriz de correlação de trabalho (externos à diagonal) são diferentes entre si e deste modo não há nenhum pressuposto sobre a correlação entre as observações, mas todos os vetores (y_i) possuem o mesmo comprimento, a matriz de correlação de trabalho é não estruturada (Dobson; Barnett, 2008; Wang, 2014; Freitas, 2018).

2.5 Inferência para os parâmetros do modelo GEE

Depois que computamos o estimador $\hat{\boldsymbol{\beta}}$ e a respectiva matriz de variância-covariância, o intervalo de confiança para os $\boldsymbol{\beta}$ e o teste de hipóteses são as principais ferramentas na área de inferência estatística. Para calcular o intervalo de confiança, é necessário conhecer a distribuição amostral do estimador. Já para o teste de hipóteses, considera-se a qualidade do ajuste estatístico do modelo. Essas estatísticas de qualidade de ajuste estatístico podem ser com base no valor máximo da função de verossimilhança, no valor máximo da função log-verossimilhança, no valor mínimo do critério da soma dos quadrados ou uma estatística composta baseada nos resíduos. Para o teste de hipóteses no modelo GEE busca verificar se os coeficientes estimados pelo modelo são iguais a zero (hipótese nula) (Dobson; Barnett, 2008; Wang, 2014):

$$H_0 : \beta_k = 0 \quad \text{contra} \quad H_1 : \beta_k \neq 0,$$

com $k = 0, 1, \cdots, p$. Como a normalidade do vetor $\hat{\boldsymbol{\beta}}$ é assintótica, segue que, sob H_0 verdadeira, a estatística de teste utilizada, no caso a Wald segue uma distribuição normal padrão para n suficientemente grande (Rotnitzky; Jewell, 1990):

$$W = \frac{\hat{\beta}_k}{ep(\hat{\beta}_k)}. \tag{2.5.1}$$

Ressalta-se que nos modelos GEE, a interpretação dos parâmetros é relativa à média da população e não a um indivíduo específico. No presente trabalho, foram testadas as seguintes hipóteses:

H_0: Independência entre as variáveis explicativas e a depressão autorreferida;

H_1: Não há independência entre as variáveis explicativas e a depressão autorreferida.

2.6 Métodos de Diagnóstico sob o Modelo GEE

Existem alguns métodos de diagnóstico da adequabilidade do modelo GEE, dentre esses, cita-se: envelope simulado, a distância de Cook e o Critério de Informação de Correlação (CIC) (Freitas, 2018; Agranonik, 2010).

O envelope simulado é um procedimento de execução de um gráfico de probabilidade meio-normal com envelope simulado por meio do qual os resíduos padronizados são ordenados em relação ao valor absoluto esperado da estatística de ordem, da Normal Padrão, da seguinte forma (Freitas, 2018):

$$\mathbb{E}(|Z_i|) \cong \Phi^{-1}\left(\frac{i+N-1/8}{2N+1/2}\right), \tag{2.6.1}$$

onde Φ é a distribuição acumulada da normal padrão. Salienta-se que este método pode ser utilizado ainda que os resíduos não sigam a distribuição normal padrão. Existem outros métodos de análise de adequação dos modelos GEE, como a distância de Cook, que mensura a influência de um subconjunto de observações sobre os parâmetros estimados e sobre os preditores lineares. A distância de Cook é calculada pela diferença entre as previsões para a variável resposta feitas com o modelo com a observação em análise e excluindo-a divido pelo erro quadrático médio do modelo ponderado pelo número de parâmetros estimados (Agranonik, 2010).

O Critério de Informação de Correlação (CIC) é uma das maneiras de se avaliar a estrutura de correlação de trabalho por meio da comparação das estimativas dos β obtidas via estrutura de correlação de trabalho independente. Quanto menor o valor do CIC, mais adequada é a estrutura de correlação para o desempenho do modelo. O CIC é calculado pela soma dos elemnetos da diagonal da multiplicação das matrizes hessianas dos estimadores gerados a partir de estimavas naive e robuta (Agranonik, 2010).

3 Metodologia

3.1 Conjunto de dados

O projeto PADu-Federais, constitui-se de um levantamento de informações sociodemográficas, de saúde, hábitos de vida, e vivência no ambiente acadêmico. Este estudo foi conduzido em dois momentos durante a pandemia de COVID-19 coletando informações dos mesmos indivíduos (estudo longitudinal). O PADu longitudinal busca avaliar alunos de 14 cursos da UFOP ao longo de quatro a cinco anos que eles estiveram vinculados à universidade. Foi realizado em duas etapas: a primeira no ano de ingresso, linha de base (T0) e a segunda, após dois anos (T1). Todos os universitários com idade igual ou superior a 18 anos e regularmente matriculados no primeiro semestre de 2019 em um dos 14 cursos selecionados foram convidados a participar. Os cursos eram os seguintes: Educação Física, Farmácia, Medicina e Nutrição, Arquitetura, Matemática, Engenharia Civil, Engenharia Geológica e Engenharia de Produção, Artes Cênicas, Direito , História, Jornalismo e Pedagogia. Cursos com maior número de alunos matriculados e menor taxa de evasão foram escolhidos em antecipação às possíveis perdas esperadas em um estudo longitudinal (César et al., 2022).

O levantamento da informação da linha de base foi feito por meio de preenchimento do questionário impresso e autoaplicável. Já, na segunda onda, o levantamento foi feito por meio de uma análise de respostas ao questionário PADu enviado por e-mail aos estudantes dos cursos mencionados anteriormente. As principais características avaliadas foram: sexo (0, se feminino e 1, se masculino), faixa etária (0, se inferior ou igual a 34 anos e 1, se superior a 34 anos), cor da pele (0, se preto, pardo, amarelo, indígena e outros e 1, se branco), estado civil (0, se casado, união estável, divorciado, viúvo e 1, se solteiro), grau de instrução do chefe da família-Educação (1, se analfabeto; 2, se fundamental incompleto a superior completo; 3, se superior completo), renda familiar total mensal em salários mínimos (1, se até 1; 2, se 2 até 4; 3, se mais de 4), consumo excessivo de álcool (0, se não e 1, se sim), prática de atividade física-Ativ. Física (0, se não e 1, se sim), tabagismo (0, se não e 1, se sim), índice de massa corporal (IMC), depressão (0, se não e 1, se sim).

3.2 Análise de dados

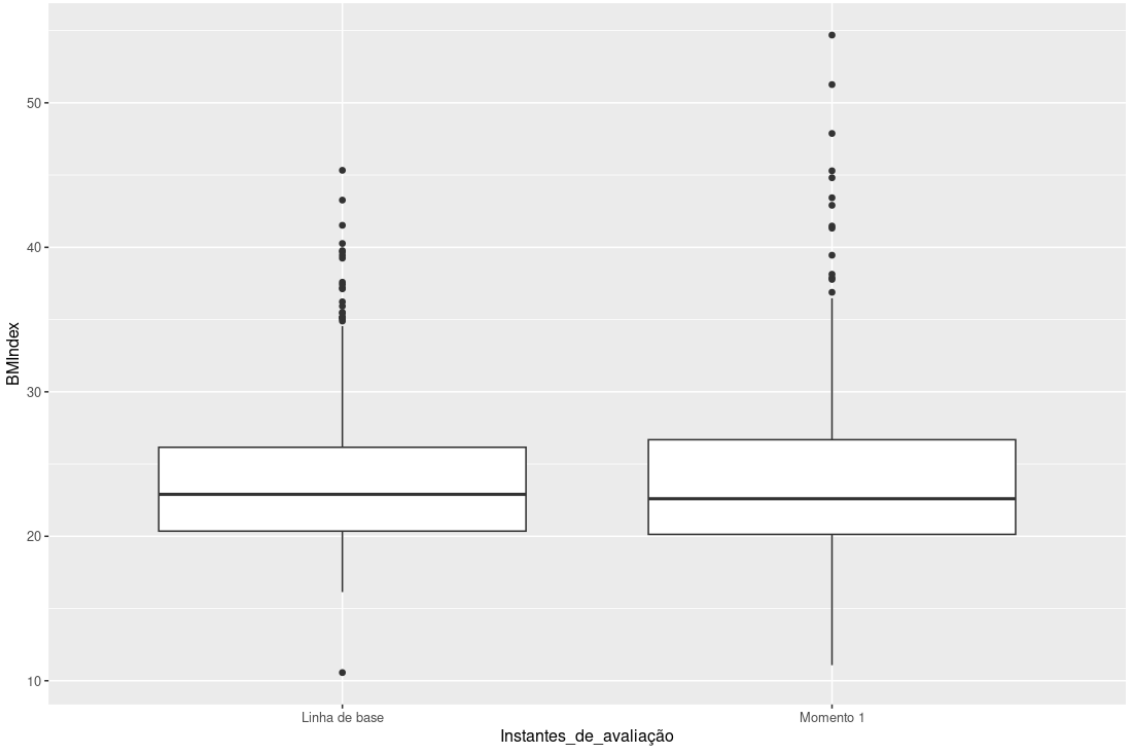
Foi realizada uma análise descritiva dos dados por meio de distribuição das frequências absolutas e relativas nos dois momentos do estudo. O modelo GEE foi utilizado para verificar a associação entre as variáveis independentes e o desfecho de depressão

entre os universitários entrevistados. O nível de significância considerado foi $p \leq 5\%$. Foi realizado o cálculo do Critério de Informação de Correlação (CIC) para decidir sobre a melhor estrutura de correlação de trabalho do modelo. Ademais, foi feita a análise gráfica dos resíduos padronizados por meio da distância de Cook (D_i) para verificar o impacto de um subconjunto de observações no modelo, considerando $D_i \geq 1$ para identificar as observações influentes independente do tamanho amostral (Agranonik, 2010; Cook, 1977).

4 Resultados

Neste estudo 614 universitários da UFOP responderam ao questionário do projeto de sintomas de depressão em universitários (PADu). A prevalência de depressão foi de 64.80% na linha de base (T0), enquanto que no segundo momento do estudo (T1) foi de 66.70%. O grupo de universitários entrevistados foi em sua maioria formado por mulheres (63.2%), com idade até 34 anos (93.19% na linha de base e 91.53% no momento 1), da raça/cor da pele preta, parda, amarela, indígena e outros (52.9%), e o nível educacional do chefe da família esteve no estrato Fundamental incompleto a Superior incompleto (61.73%), a faixa de renda total da família foi de dois a quatro salários mínimos (60.91% na linha de base e 60.42% no momento 1). A frequência de consumo de álcool entre os entrevistados foi 65.30% na linha de base e 62,60% no momento 1; enquanto que a prevalência de tabagismo foi 23.80% e 24.99% na linha de base e no momento 1, respectivamente. A prática de atividade física foi relatada por 63.40% dos universitários na linha de base e 60.75% no momento 1 (Tabela 1).

Figura 2: Box-plot do Índice de Massa Corporal (IMC) dos universitários nos dois momentos do estudo



A média e o desvio-padrão do Índice de Massa Corporal (IMC) entre os entrevistados foram de 23.8 (4.82) na linha de base e 23.8 (5.60) no momento 1 (Vide na Tabela 2). O IMC dos universitários estudados apresentou maior variação no momento 1 comparado à linha de base, houve mais valores extremos para o IMC no momento 1 conforme pode ser observado na Figura 2.

Tabela 1: Distribuição de frequências das variáveis de estudo na linha de base e no momento 1. PADu, 2020.

Variável	Linha de Base	Momento 1
	n(%)	n(%)
Sexo		
Feminino	389 (63.40)	389 (63.40)
Masculino	225(36.60)	225(36.60)
Faixa Etária (em anos)		
Até 34	575 (93.19)	562 (91.53)
Acima de 34	39 (6.81)	52 (8.47)
Cor da pele		
Preto, pardo, amarelo, indígena e outros	325(52.90)	325(52.90)
Branco	289 (47.10)	289 (47.10)
Estado Civil		
Casado, união estável, divorciado, viúvo	567 (92.35)	567 (92.35)
Solteiro	47 (7.65)	47 (7.65)
Nível Educacional do chefe da família		
Analfabeto	6 (0.97)	6 (0.97)
Fundamental incompleto a Superior incompleto	379 (61.73)	379 (61.73)
Superior Completo	229 (37.30)	229 (37.30)
Renda Total da Família (em Salários mínimos)		
Até 1	44 (7.17)	47 (7.65)
2 até 4	374 (60.91)	369 (60.42)
Mais de 4	196 (31.92)	200 (32.57)
Consumo de álcool		
Não	213 (34.70)	229 (37.30)
Sim	401 (65.30)	385 (62.70)
Tabagismo		
Não	468 (76.20)	461 (75.01)
Sim	146 (23.80)	153 (24.99)
Prática de Atividade Física		
Não	225 (36.60)	241 (39.25)
Sim	389 (63.40)	373 (60.75)
Sintomas de Depressão		
Não	216 (35.20)	206 (33.70)
Sim	398 (64.80)	408 (66.70)
Índice de Massa Corporal		
	Média (desvio-padrão)	**Média (desvio-padrão)**
	23.8 (4.82)	23.8 (5.60)
Total	**614 (100)**	**614 (100)**

Tabela 2: Resultados do ajuste do modelo GEE sem termos de interação. A estrutura de correlação independente foi considerada neste caso.

Fatores de risco	Estimativa	ep_{nai}	Z_{nai}	Sig_{nai}	ep_{rob}	Z_{rob}	Sig_{rob}
Intercepto	0.714	0.952	0.453	0.453	0.878	0.813	0.416
Sexo (Masculino)	-0.999	0.132	-7.594	< 0.001	0.157	-6.377	< 0.001
Idade (>34 anos)	1.008	0.267	3.772	< 0.001	0.289	3.484	< 0.001
Cor da pele (Branca)	-0.038	0.130	-0.294	0.767	0.155	-0.246	0.806
Estado Civil (Solteiro)	-0.684	0.259	-2.641	0.008	0.286	-2.388	0.017
Educação-2	-0.306	0.803	-0.381	0.703	0.636	-0.481	0.631
Educação-3	-0.312	0.812	-0.384	0.701	0.650	-0.481	0.631
Renda-2	-0.712	0.293	-2.428	0.015	0.314	-2.267	0.023
Renda-3	-0.795	0.311	-2.545	0.011	0.333	-2.379	0.017
Consumo de álcool (Sim)	-0.160	0.139	-1.149	0.251	0.154	-1.039	0.299
Tabagismo (Sim)	0.795	0.167	4.755	< 0.001	0.181	4.397	< 0.001
Ativ. Física (Sim)	-0.214	0.134	-1.596	0.110	0.139	-1.538	0.124
IMC	0.018	0.013	1.386	0.166	0.014	1.293	0.196
Tempo	0.073	0.127	0.572	0.567	0.099	0.736	0.462

`Legenda`: Renda-2 e Renda-3 denota as categorias 2 e 3 da renda total da família, Educação-2 e Educação-3 denota as categorias 2 e 3 do fator graus de instrução do chefe da família, *ep* denota o erro-padrão e *Sig* as probabilidades de significância (p-valor). De igual forma, a notação `rob` e `nai` denotam as estimativas obtidas usando os métodos robusto e naive, respectivamente.

Tabela 3: Resultados do ajuste do modelo GEE sem termos de interação. A estrutura de correlação permutável foi considerada neste caso

Fatores de risco	Estimativa	ep_{nai}	Z_{nai}	Sig_{nai}	ep_{rob}	Z_{rob}	Sig_{rob}
Intercepto	0.859	1.085	0.794	0.427	0.824	1.045	0.298
Sexo (Masculino)	-0.972	0.154	-6.319	< 0.001	0.154	-6.290	< 0.001
Idade (>34 anos)	0.765	0.290	2.621	0.009	0.256	2.967	0.003
Cor da pele (Branca)	-0.044	0.153	-0.286	0.775	0.154	-0.283	0.777
Estado Civil (Solteiro)	-0.739	0.300	-2.468	0.014	0.278	-2.660	0.008
Educação-2	-0.572	0.959	-0.599	0.549	0.594	-0.967	0.336
. Educação-3	-0.614	0.968	-0.638	0.523	0.608	-1.017	0.312
Renda-2	-0.260	0.272	-0.934	0.350	0.284	-0.895	0.359
Renda-3	-0.353	0.296	-1.170	0.242	0.301	-1.154	0.240
Consumo de álcool (Sim)	-0.044	0.141	-0.301	0.763	0.139	-0.305	0.751
Tabagismo (Sim)	0.736	0.183	4.015	< 0.001	0.168	4.382	< 0.001
Atv. Física (Sim)	-0.107	0.128	-0.819	0.413	0.124	-0.844	0.392
IMC	0.009	0.014	0.669	0.503	0.013	0.714	0.468
Tempo	0.077	0.096	0.803	0.422	0.096	0.806	0.421

`Legenda`: Renda-2 e Renda-3 denota as categorias 2 e 3 da renda total da família, Educação-2 e Educação-3 denota as categorias 2 e 3 do fator graus de instrução do chefe da família, *ep* denota o erro-padrão e *Sig* as probabilidades de significância (p-valor). De igual forma, a notação `rob` e `nai` denotam as estimativas obtidas usando os métodos robusto e naive, respectivamente.

Os ajustes via modelo GEE considerando os casos para verificar a associação dos onze fatores de risco estudados (sexo, faixa etária, raça/cor da pele, estado civil, nível educacional do chefe da família, renda total da família, consumo de álcool, tabagismo, prática de atividade física, sintomas de depressão, Índice de Massa Corporal) e o desfecho de sintomas de depressão entre os universitários entrevistados está descrita nas Tabelas 2 a 5. Por meio do teste de Wald, verificou-se que os parâmetros significativos (teste com $p - valor \leq 0.05$) para a estrutura de correlação independente, as variáveis significativas foram: sexo, faixa etária, estado civil, renda total da família e tabagismo. Dessa forma, os fatores de risco para depressão entre estudantes universitários independente do momento do estudo foram: sexo feminino, idade menor igual a 34 anos, solteiro, com renda familiar inferior a dois salários mínimos e tabagista (Tabela 3).

Para a estrutura de correlação permutável sem termos de interação, as seguintes variáveis mostraram-se associadas ao desfecho: sexo feminino, com menos de 34 anos de idade, casado/viúvo/separado e fumante (Tabela 3). Quando foram analisados os termos de interação, para a estrutura de correlação independente, as mesmas associações com o desfecho de depressão foram evidenciadas para o modelo com e sem termo de interação, acrescida da interação entre sexo masculino e prática de atividade física. Ou seja, a presença de interação entre os fatores sexo e prática de atividade física significa que a diferença entre a presença de depressão entre homens e mulheres não é a mesma nos dois momentos do estudo (Tabela 4).

No modelo com interação e estrutura de correlação permutável, nota-se que as variáveis associadas inversamente com a presença de depressão foram as mesmas o modelo anterior: sexo masculino, idade acima de 34 anos, solteiro. O termo de interação sexo masculino/prática de atividade física foi estatisticamente significativo no modelo GEE com estrutura de correlação permutável, e o tabagismo manteve-se associado positivamente ao desfecho estudado (Tabela 5). Analisando o modelo GEE com estrutura de correlação independente, onde assume-se independência entre as observações para o mesmo indivíduo, e o modelo GEE com estrutura de correlação permutável (pressuposto de igual correlação entre os diferentes momentos de estudo (linha de base e momento 1) produziram estimativas muito próximas para os parâmetros do intercepto e das inclinações. O erro-padrão robusto é menor do que o naive em ambos os modelos citados.

As estatísticas de Wald para as diferenças nas estimativas dos $\hat{\beta}$, demonstradas pelo p-valor presente nas tabelas, fornecem evidência para rejeição da hipótese nula, indicando que há diferenças nas prevalências de depressão entre os universitários do sexo feminino, com idade menor de 34 anos, casados, com renda familiar menor de dois salários mínimos, fumantes. A inferência para modelos ajustados por GEEs é melhor realizada usando estatísticas de Wald com um estimador sanduíche robusto para a variância, dessa forma, a opção deste estudo é usar os estimadores com o erro-padrão robusto.

Tabela 4: Resultados do ajuste do modelo GEE com termos de interação. A estrutura de correlação independente foi considerada neste caso

Fatores de risco	Estimativa	ep_{nai}	Z_{nai}	Sig_{nai}	ep_{rob}	Z_{rob}	Sig_{rob}
Intercepto	0.438	1.035	0.424	0.672	0.972	0.451	0.652
Sexo (Masculino)	-1.399	0.219	-6.383	< 0.001	0.238	-5.875	< 0.001
Idade (> 34 anos)	1.022	0.272	3.762	< 0.001	0.300	3.402	< 0.001
Cor da pele (Branca)	-0.059	0.131	-0.452	0.652	0.156	-0.379	0.705
Estado Civil (Solteiro)	-0.690	0.262	-2.634	0.008	0.294	-2.348	0.019
Educação-2	-0.398	0.809	-0.492	0.622	0.641	-0.627	0.535
Educação-3	-0.410	0.818	-0.501	0.616	0.654	-0.627	0.531
Renda-2	-0.710	0.295	-2.408	0.016	0.317	-2.238	0.025
. Renda-3	-0.784	0.313	-2.509	0.012	0.336	-2.335	0.020
Consumo de álcool (Sim)	0.858	0.654	1.312	0.190	0.678	1.264	0.206
Tabagismo (Sim)	0.812	0.176	4.813	< 0.001	0.184	4.422	< 0.001
Ativ. Física (Sim)	-0.457	0.252	-1.816	0.070	0.272	-1.678	0.093
IMC	0.040	0.020	1.988	0.046	0.022	1.823	0.068
Tempo	0.065	0.128	0.506	0.612	0.099	0.653	0.506
Sexo×Atv. Física	0.637	0.272	2.341	0.020	0.289	2.202	0.028
Consumo de álcool× Atv. Física	-0.066	0.278	-0.237	0.812	0.287	-0.229	0.818
Consumo de álcool× IMC	-0.041	0.026	-1.607	0.108	0.026	-1.558	0.119

`Legenda`: Renda-2 e Renda-3 denota as categorias 2 e 3 da renda total da família, Educação-2 e Educação-3 denota as categorias 2 e 3 do fator graus de instrução do chefe da família, *ep* denota o erro-padrão e *Sig* as probabilidades de significância (p-valor). De igual forma, a notação `rob` e `nai` denotam as estimativas obtidas usando os métodos robusto e naive, respectivamente.

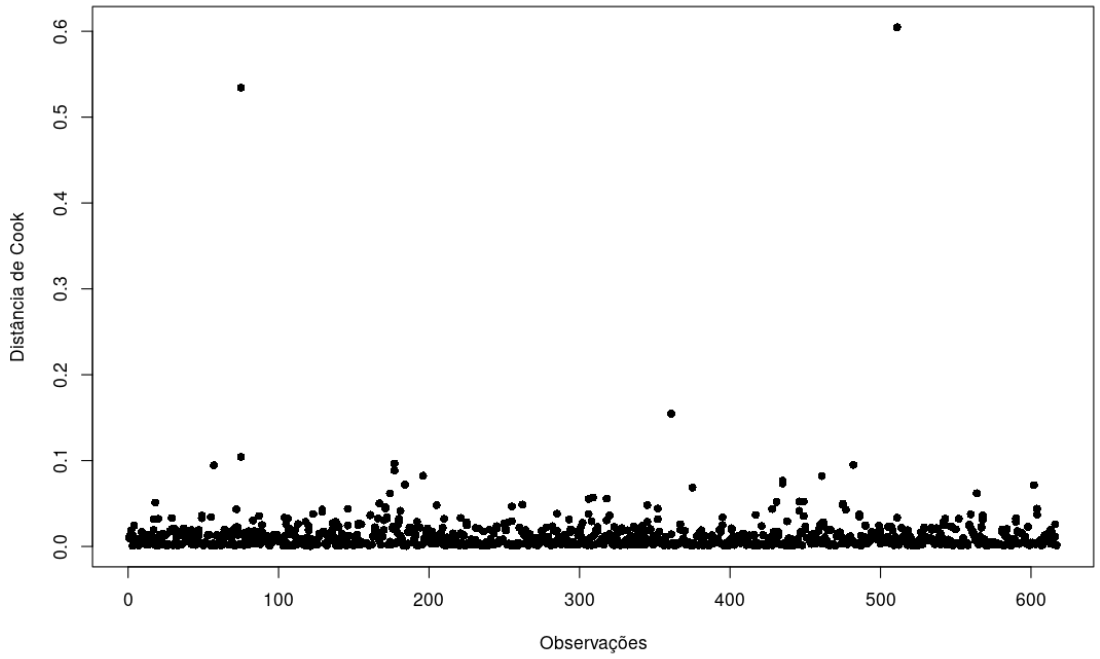
Tabela 5: Resultados do ajuste do modelo GEE com termos de interação. A estrutura de correlação permutável foi considerada neste caso

Fatores de risco	Estimativa	ep_{nai}	Z_{nai}	Sig_{nai}	ep_{rob}	Z_{rob}	Sig_{rob}
Intercepto	0.498	1.147	0.434	0.664	0.901	0.553	0.580
Sexo (Masculino)	-1.283	0.227	-5.662	< 0.001	0.221	-5.814	< 0.001
Idade (>34 anos)	-0.776	0.293	-2.648	0.008	0.264	-2.944	0.004
Cor da pele (Branca)	-0.061	0.153	-0.398	0.690	0.155	-0.395	0.692
Estado Civil (Solteiro)	-0.745	0.302	-2.467	0.014	0.283	-2.635	0.008
Educação-2	-0.641	0.958	-0.669	0.504	0.603	-1.063	0.288
Educação-3	-0.690	0.967	-0.713	0.476	0.616	-1.119	0.264
Renda-2	-0.248	0.273	-0.908	0.364	0.283	-0.876	0.382
. Renda-3	-0.344	0.298	-1.156	0.248	0.301	-1.146	0.252
Consumo de álcool (Sim)	0.952	0.630	1.510	0.132	0.610	1.561	0.118
Tabagismo (Sim)	0.748	0.184	4.061	< 0.001	0.170	4.391	< 0.001
Ativ. Física (Sim)	-0.354	0.235	-1.510	0.132	0.220	-1.610	0.108
IMC	0.034	0.020	1.646	0.100	0.019	1.738	0.082
Tempo	0.076	0.097	0.779	0.436	0.097	0.783	0.434
Sexo × Ativ. Física	0.501	0.262	1.914	0.056	0.260	1.928	0.054
Consumo de álcool×Ativ. Física	0.052	0.256	0.196	0.844	0.245	0.205	0.838
Consumo de álcool× IMC	-0.043	0.025	-1.750	0.080	0.024	-1.805	0.072

`Legenda`: Renda-2 e Renda-3 denota as categorias 2 e 3 da renda total da família, Educação-2 e Educação-3 denota as categorias 2 e 3 do fator graus de instrução do chefe da família, *ep* denota o erro-padrão e *Sig* as probabilidades de significância (p-valor). De igual forma, a notação `rob` e `nai` denotam as estimativas obtidas usando os métodos robusto e naive, respectivamente.

O Critério de Informação de Correlação (CIC), comparando os modelos sem interação (CIC= 14.3) e com interação (CIC=14.8), foi menor para os modelos sem interação. Salienta-se que esta medida é interpretada como quanto menor seu valor, melhor o desempenho do modelo a respeito da escolha da estrutura de correlação de trabalho. Desse modo, os modelos sem interação demonstraram melhor performance com estrutura de correlação independente. Para este modelo foi feita a análise dos resíduos que demonstrou a não existência de pontos influentes com distância de Cook superiores a 1 (Figura 3).

Figura 3: Gráfico da Distância de Cook para o Modelo GEE ajustado



5 Considerações Finais

A maioria dos universitários envolvidos nesse estudo foi mulheres, com idade até 34 anos, da raça/cor da pele preta, parda, amarela, indígena e outros, nível educacional do chefe da família foi intermediário e a faixa de renda total da família foi de dois a quatro salários mínimos. A frequência de consumo de álcool e a prática de atividade física entre os entrevistados foram elevadas (acima de 60% nos dois períodos de estudo). Cerca de um em cada quatro universitários no estudo relataram tabagismo. O Índice de Massa Corporal (IMC) médio entre os entrevistados foi de 23.8.

Os resultados em tela corroboram com a afirmação de que a ansiedade, depressão e sedentarismo foram as condições que mais afetaram estudantes universitários durante o primeiro período letivo que transcorreu na pandemia de COVID-19 segundo um estudo longitudinal realizado na Universidade de Dartmouth (Huckins et al., 2020). Em um estudo publicado por Dodd et al. (2021), cerca de 65.3% dos universitários apresentaram bem-estar baixo e muito baixo, respectivamente. Os fatores associados ao menor bem-estar foram: sexo (feminino), status social autodeclarado como baixo, ter sentido impacto da pandemia nos estudos e experiência negativa de aprendizado.

Uma revisão de literatura com o objetivo de conhecer o impacto psicológico do isolamento social devido à pandemia de COVID-19 mostrou que os sintomas psicológicos mais prevalentes foram distúrbios emocionais, depressão, estresse, humor deprimido, raiva, irritabilidade, insônia, ansiedade, sintomas de estresse pós-traumático e exaustão emocional. Os fatores estressores associados a estes impactos pscicológicos na pandemia foram: duração do isolamento, medo da infecção, sentimento de frustração e tédio pela mudança brusca na rotina, escassez ou inadequação do aporte de suprimentos (alimentos, àgua, roupas...), informações inadequadas sobre a realidade da pandemia (Brooks et al., 2020).

Para a análise de dados longitudinais, quando a suposição de normalidade do conjunto de dados da variável resposta não é atendida, os métodos que utilizam a verossimilhança para inferência não estão disponíveis. Assim, o método de análise de dados longitudinais mais indicado são as equações de estimação generalizadas (GEE) (Liang; Zeger, 1986). Os dados do estudo PADu foram analisados no presente texto utilizando modelagem GEE.

Observou-se que por meio da modelagem GEE as variáveis associadas à depressão entre os universitários foram: a) associação inversa: sexo masculino, idade acima de 34 anos, solteiro, com renda familiar acima de dois salários mínimos. Ainda no presente estudo; b) associação positiva: tabagismo.

A pandemia de Covid-19 acarretou impactos de amplo espectro na rotina de vida e gerou consequências prolongadas nos campos da saúde, da economia e social (Haleem;

Javaid; Vaishya, 2020). É notório que qualquer grande epidemia trará impactos negativos tanto para os indivíduos quanto para a sociedade (Duan; Zhu, 2020).

Um axioma dos GLM é que as observações devem ser independentes (AGRESTI, 2015) e isso limita sua aplicação em um cenário de dados longitudinais. Os dados longitudinais são aqueles conjuntos que possuem uma variável resposta (y_{it}) e covariáveis explicativas (x_{it}) para cada uma das observações (i = l,...,K) no banco de dados medidas em mais de um momento no tempo (t = 1,...,n) (Liang; Zeger, 1986). Esses autores argumentam que quando se trata da análise de dados com medidas repetidas, a correlação entre esses valores para uma dada observação deve ser considerada. Para isso, no modelo de estimativa dados longitudinais usando regressão, os autores desenvolveram um método baseado nos GLM, apresentando uma classe de equações de estimação (equações estimação generalizadas (GEE)). As GEE fornecem estimativas mais robustas para os parâmetros de regressão quando há uma estrutura de dependência entre as observações (Liang; Zeger, 1986).

Além do argumento exposto acima, como vantagens do uso de modelagem GEE cita-se a versatilidade de tipos de variáveis como desfecho, geração de estimadores consistentes, escolha de matriz de correlação mais adequada aos dados. Como limitações do presente estudo, o relato de depressão foi autodeclarado, não dispondo de diagnóstico médico para tal. A pergunta sobre consumo de álcool pode não ter sido sensível o suficiente para captar a ingestão entre os universitários. Não foi utilizado um método de seleção de variáveis para compor o modelo, pois não havia outras variáveis disponíveis no banco de dados, ou seja, todas as variáveis do banco foram empregadas na modelagem.

É salutar destacar que a seleção adequada da matriz de correlação não é crucial, quando o estimador é robusto em relação às variações na escolha dessa matriz. No entanto, uma escolha inadequada pode diminuir a eficiência do estimador. Na prática, a decisão pode ser influenciada pelo número de parâmetros de correlação a serem estimados. Por exemplo, o uso de uma grande matriz de correlação não estruturada pode resultar em estimativas instáveis ou dificuldades de convergência nos cálculos (Dobson; Barnett, 2008).

Considerando o que foi observado no presente estudo, o método de modelagem de Equações de Estimação Generalizadas demonstrou ser consistente com os dados e gerou informações que podem orientar medidas de saúde mental para grupos de universitários, pois contribuiu para identificação de características que implicam em uma associação com sintomas autoreferidos de depressão em contexto pandêmico.

Referências

Agranonik, M. *Equações de Estimação Generalizadas (GEE): Aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007)*. 2010.

Agresti, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018.

Barros, M. B. d. A. et al. Relato de tristeza/depressão, nervosismo/ansiedade e problemas de sono na população adulta brasileira durante a pandemia de covid-19. *Epidemiologia e Serviços de Saúde*, SciELO Public Health, v. 29, p. e2020427, 2020.

Bernardelli, L. V. et al. A ansiedade no meio universitário e sua relação com as habilidades sociais. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, SciELO Brasil, v. 27, p. 49–67, 2022.

Brooks, S. K. et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The lancet*, Elsevier, v. 395, n. 10227, p. 912–920, 2020.

César, P. d. S. et al. Dietary practices of university students according to the dietary guidelines for the brazilian population: Padu study. *Revista de Nutrição*, SciELO Brasil, v. 35, 2022.

Cook, R. Deletion of influential observations in linear regression. *Technometrics*, v. 19, n. 1, p. 15–18, 1977.

Cordeiro, G. M.; Demétrio, C. G. *Modelos Lineares Generalizados e Extensões*. Geneva, CH, 2013. Disponível em: ⟨https://www.ufjf.br/clecio_ferreira/files/2013/05/Livro-Gauss-e-Clarice.pdf⟩.

Dobson, A. J.; Barnett, A. G. *An introduction to generalized linear models*. [S.l.]: CRC/Taylor & Francis, 2008.

Dodd, R. H. et al. Psychological wellbeing and academic experience of university students in australia during covid-19. *International Journal of Environmental Research and Public Health*, MDPI, v. 18, n. 3, p. 866, 2021.

Duan, L.; Zhu, G. Psychological interventions for people affected by the covid-19 epidemic. *Lancet Psychiatry*, Lancet, v. 7, n. 4, p. 300–302, 2020.

Fitzmaurice, G. M.; Laird, N. M.; Ware, J. H. *Applied longitudinal analysis*. [S.l.]: John Wiley & Sons, 2012. v. 998.

Freitas, J. V. d. B. *Modelagem de dados com medidas repetidas via Equações de Estimação Generalizadas*. 2018.

Haleem, A.; Javaid, M.; Vaishya, R. Effects of covid-19 pandemic in daily life. *Curr Med Res Pract*, v. 10, n. 2, p. 78–79, 2020.

Huckins, J. et al. Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *J Med Internet Res*, JMIR, v. 22, n. 6, p. e20185, 2020.

Liang, K.-Y.; Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, Oxford University Press, v. 73, n. 1, p. 13–22, 1986.

Maia, B. R.; Dias, P. C. Anxiety, depression and stress in university students: the impact of covid-19. *Estudos de Psicologia (Campinas)*, SciELO Brasil, v. 37, 2020.

Paula, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: ⟨https://www.R-project.org/⟩.

Rotnitzky, A.; Jewell, N. P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, v. 77, p. 485–497, 1990.

Rubin, G. J.; Wessely, S. The psychological effects of quarantining a city. *Bmj*, British Medical Journal Publishing Group, v. 368, 2020.

Twisk, J. W. *Applied longitudinal data analysis for epidemiology: a practical guide*. [S.l.]: cambridge university press, 2013.

Venezuela, M. *Modelos Lineares Generalizados para Análise de Dados com Medidas Repetidas*. 2003.

Wakefield, J. Course. In: *Stat/Biostat 571 Statistical Methodology: Regression Models for Dependent Data.* [S.l.: s.n.], 2009.

Wang, M. Generalized estimating equations in longitudinal data analysis: A review and recent developments. *Advances in Statistics*, v. 2014, p. 1–11, 12 2014.

Wedderburn, R. W. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, Oxford University Press, v. 61, n. 3, p. 439–447, 1974.

WHO. Coronavirus disease (covid-19): Herd immunity, lockdowns and covid-19. 2020. Disponível em: URL ⟨www.who.int/news-room/questions-and-answers/item/herd-immunity-lockdowns-and-covid-19⟩. Acesso em: 21 dez. 2022.

WHO. Mental health and psychosocial considerations during the covid-19 outbreak, 18 march 2020. 2020. Disponível em: URL ⟨www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10(-)4⟩. Acesso em: 21 dez. 2022.

Comandos no R

```
install.packages("dplyr")
library(dplyr)
library(foreign)
library(gee)
library(geepack)
library(geeasy)

load("/cloud/project/LongStructure (3).Rdata")

View(LongitudData)

class(LongitudData$times)

#-------------------------------------------------------------
#  Análise Descritiva
#-------------------------------------------------------------

class(LongitudData$ageCat)
LongitudData <- LongitudData  %>%
mutate(ageCat = recode(ageCat, "0"="1",
"1"="0"))



bancoTempo1<- LongitudData %>%
filter(times==1)
bancoTempo2<-LongitudData %>%
filter(times==2)

#Descritivos no tempo 1
#Idade

t_ageCat1 <- table(bancoTempo1$ageCat)
t_ageCat1
prop.table(t_ageCat1)

#Raça/cor da pele
t_skinCol1 <- table(bancoTempo1$skinCol)
```

```
t_skinCol1
prop.table(t_skinCol1)

#Sexo
t_sex1<-table(bancoTempo1$sex)
t_sex1
prop.table(t_sex1)

#Estado Civil
t_marStatus1 <- table(bancoTempo1$marStatus)
t_marStatus1
prop.table(t_marStatus1)

t_educaCat1<-table(bancoTempo1$eduCat)
t_educaCat1
prop.table(t_educaCat1)



#Renda
t_income1 <- table(bancoTempo1$income)
t_income1
prop.table(t_income1)

#Consumo de Álcool
t_alchool1 <- table(bancoTempo1$alchool)
t_alchool1
prop.table(t_alchool1)

#Hábito de Fumar
t_smoke1 <- table(bancoTempo1$smoke)
t_smoke1
prop.table(t_smoke1)

#Atividade Física

t_physAtv1 <- table(bancoTempo1$physAtv)
t_physAtv1
prop.table(t_physAtv1)
```

```
#Média do Índice de Massa Corporal (IMC)

t_bmi <- mean(bancoTempo1$BMIndex)
t_bmi
#desvio padrão do BMI
sd_bmi <- sd(bancoTempo1$BMIndex)
sd_bmi


#Depressão

t_depress1<- table(bancoTempo1$depress)
t_depress1
prop.table(t_depress1)

#Descritivos no tempo 2
#Idade
t_ageCat2 <- table(bancoTempo2$ageCat)
t_ageCat2
prop.table(t_ageCat2)



#Raça/cor da pele
t_skinCol2 <- table(bancoTempo2$skinCol)
t_skinCol2
prop.table(t_skinCol2)

#Sexo
t_sex2 <- table(bancoTempo2$sex)
t_sex2
prop.table(t_sex2)

#Estado Civil
t_marStatus2 <- table(bancoTempo2$marStatus)
t_marStatus2
prop.table(t_marStatus2)

#Educação do chefe de família
```

```
t_educaCat2<-table(bancoTempo2$eduCat)
t_educaCat2
prop.table(t_educaCat2)

#Renda
t_income2 <- table(bancoTempo2$income)
t_income2
prop.table(t_income2)

#Consumo de Álcool
t_alchool2 <- table(bancoTempo2$alchool)
t_alchool2
prop.table(t_alchool2)

#Hábito de Fumar
t_smoke2 <- table(bancoTempo2$smoke)
t_smoke2
prop.table(t_smoke2)

#Atividade Física

t_physAtv2 <- table(bancoTempo2$physAtv)
t_physAtv2
prop.table(t_physAtv2)

#Depressão

t_depress2<- table(bancoTempo2$depress)
t_depress2
prop.table(t_depress2)



t_depress1<- table(bancoTempo1$depress)
t_depress1
prop.table(t_depress1)

#Média do Índice de Massa Corporal (IMC)

t_bmi2 <- mean(bancoTempo2$BMIndex)
```

```
t_bmi2
#desvio padrão do BMI
sd_bmi2 <- sd(bancoTempo2$BMIndex)
sd_bmi2

t_bmi

sd_bmi


#-------------------------------------------------------------
#  Modelagem GEE
#-------------------------------------------------------------

gee.fit<- geeglm(depress ~ sex+skinCol+marStatus+physAtv+smoke+alchool+
ageCat+ relevel(income, ref = "1")+relevel(eduCat, ref = "1")+
BMIndex+times,data = LongitudData,family = binomial(link = "logit"),
id=id, std.err = "san.se", waves=times,corstr ="independence")

summary(gee.fit)

gee.fit.str<-gee(depress ~ sex+skinCol+marStatus+physAtv+smoke+alchool+
ageCat+relevel(income, ref = "1")+relevel(eduCat, ref="1")
+BMIndex+times+times,data = LongitudData,family = binomial(link = "logit"),
id=id,corstr ="independence")

summary(gee.fit.str)


gee.fit.permut<- geeglm(depress ~ sex+skinCol+marStatus+physAtv+smoke+
alchool+ ageCat+relevel(income, ref = "1")+relevel(eduCat,
ref="1")+BMIndex+times,data=LongitudData,family = binomial(link = "logit"),
id=id,std.err = "san.se", waves=times,corstr ="exchangeable")

summary(gee.fit.permut)

gee.fit.permut.gee<- gee(depress ~ sex+skinCol+marStatus+physAtv+smoke+
alchool+ ageCat+relevel(income, ref="1")+relevel(eduCat,
ref="1")+BMIndex+times,data = LongitudData,family = binomial(link = "logit"),
```

```
id=id,corstr ="exchangeable")

summary(gee.fit.permut.gee)

gee.fit.intera<- gee(depress ~ sex+skinCol+marStatus+physAtv+smoke+
alchool+ageCat+relevel(income, ref="1")+relevel(eduCat,
ref="1")+BMIndex+BMIndex+sex:physAtv+alchool:physAtv+alchool:BMIndex+times,
data=LongitudData,family = binomial(link="logit"), id=id,corstr= "independence")

summary(gee.fit.intera)

gee.fit.intera.glm<- geeglm(depress ~ sex+skinCol+marStatus+physAtv+
smoke+alchool+ ageCat+relevel(income,ref="1")+relevel(eduCat,ref="1")+
BMIndex+BMIndex+sex:physAtv+alchool:physAtv+alchool:BMIndex+times,
data =LongitudData, family = binomial(link = "logit"),
id=id,corstr = "independence")

summary(gee.fit.intera.glm)

#comparar estimativa robusta e não robusta
#RESULTADOS NA MESMA tabela: estrutura independente (robusto e o naive),
calcular os valores de p
#manualmente#acrescenta uma calcula com o valor p. Fazer também para o valor robusto
#se o valor z for negativo, esquece o sinal e pega a
#usar o valor z e pegar a cauda superior
#Testas qualidade do ajuste apenas dos modelos com interação.
#Calculo do valor de p pela normal pnorm
2*pnorm( 0.582,0.1, lower.tail = FALSE)
#[1] 0.63

0.582^2
# 0.339
2*pnorm( 0.582,0,1, lower.tail = FALSE)
#[1] 0.561
2*pnorm( -5.894,0,1, lower.tail = FALSE)

pchisq(34.74,1,lower.tail = FALSE)
#[1] 3.77e-09
```

```
gee.fit.intera.permut<- gee(depress ~ sex+skinCol+marStatus+physAtv+smoke+
alchool+ ageCat+relevel(income, ref="1")+relevel(eduCat,
ref="1")+BMIndex+BMIndex+sex:physAtv+alchool:physAtv+alchool:BMIndex+
times,data=LongitudData,family = binomial(link="logit"),
id=id,corstr="exchangeable")

summary(gee.fit.intera.permut)

gee.fit.intera.permut.glm<-geeglm(depress ~ sex+skinCol+marStatus+
physAtv+smoke+alchool+ageCat+relevel(income,ref="1")+
relevel(eduCat,ref="1")+ BMIndex+BMIndex+sex:physAtv+alchool:physAtv+
alchool:BMIndex+times,data=LongitudData,family=binomial(link="logit"),
id=id,corstr="exchangeable")

summary(gee.fit.intera.permut.glm)
summary(gee.fit.permut)
summary(gee.fit.permut.gee)
summary(gee.fit.str)

##Cálculo do CIC para escolha do melhor modelo
#Modelo sem interação
hessian.ind<-solve(gee.fit.str$naive)
robust.ex<-gee.fit.permut.gee$robust # Obtém o estimador robusto
para matriz de covariância
#considerando a estrutura de correlação permutável
cic<-sum(diag(hessian.ind%*%robust.ex)) # Calcula o valor do CIC para
a estrutura de correlação permutável
cic
#CIC=14.8


#Modelo com interação
hessian.ind.intera<-solve(gee.fit.intera$naive)
robust.ex.intera<-gee.fit.intera.permut$robust
cic<-sum(diag(hessian.ind.intera%*%robust.ex.intera)) # Calcula o
valor do CIC para a estrutura de correlação permutável
cic
#CIC=17.3
```

```
##Análise de resíduos

X<- model.matrix(as.formula(paste("~ ", gee.fit.str$call$formula[3])), LongitudData)
y <- gee.fit.str$y
beta <- coef(gee.fit.str)
R <- gee.fit.str$work
mi <- fitted(gee.fit.str)
individuo <- gee.fit.str$id

repet <- dim(gee.fit.str$work)[1]
ue <- gee.fit.str$nobs/repet

N <- nrow(X)
p <- ncol(X)

#Matriz C <- A * Delta
#Ligação canônica -> Delta=Identidade
A <- diag(mi*(1-mi),N)
C <- A

#Matriz Omega - variância e covariância de y
Omega <- matrix(0,N,N)
invOmega <- matrix(0,N,N)
l <- 1
while (l<N)
{
  Omega[l:(l+repet-1),l:(l+repet-1)] <- sqrt(A[l:(l+repet-1),
  l:(l+repet-1)])%*%R%*%sqrt(A[l:(l+repet-1),l:(l+repet-1)])
  invOmega[l:(l+repet-1),l:(l+repet-1)] <-solve(Omega[l:(l+repet-1),l:(l+repet-1)])
  l <- l+repet
}


#Matrizes H e W
W <- C%*%invOmega%*%C
H <- solve(t(X)%*%W%*%X)
raizW <- matrix(0,N,N)
i <- 1
l <- 1
```

```
while (l<N)
{
auto<-eigen(W[l:(l+repet-1),l:(l+repet-1)])
  raizW[l:(l+repet-1),l:(l+repet-1)] <-
  auto$vectors%*%sqrt(diag(auto$values))%*%t(auto$vectors)
  l <- l+repet
  i <- i+1
  }
H <- raizW%*%X%*%H%*%t(X)%*%raizW
h <- diag(H)


#Ponto Alavanca por UE
hue<-as.vector(rep(0,ue))
haux <- matrix(h,ue,repet,byrow=T)
for (i in 1:ue)
  hue[i] <- sum(haux[i,])/repet

#Resíduo Padronizado
rsd <- as.vector(rep(0,N))
part.rsd <- raizW%*%solve(C)%*%(y-mi)
for (l in 1:N)
{
  e <- as.vector(rep(0,N))
  e[l] <- 1
  rsd[l] <- t(e)%*%part.rsd/sqrt(1-h[l])
}

#Distância de Cook
cd <- as.vector(rep(0,N))
for (l in 1:N)
{
  cd[l] <- (rsd[l])^2*h[l]/((1-h[l]))
}

#------------------------------------------------------------
#  Construção de Gráficos de Diagnóstico
#------------------------------------------------------------
# Para  identificar os pontos que  mais  se destacam  em algum
```

```
# gráfico, use o comando  identify(...) colocando em n o
# número de pontos que se destacaram.
#-------------------------------------------------------------

#Distância de Cook

plot(individuo,cd,xlab="Observações", ylab="Distância de Cook", pch=16)
if (identifica[3]>0) identify(individuo,labels=labelsGraf,cd,n=identifica[3])


### Ponto de Alavanca

plot(hue,xlab="Observações", ylab="H por Unidade Experimental", pch=16)
abline(cut,0,lty=2)
```