Universidade de Brasília
Departamento de Estatística

Estudo dos padrões de consumo de álcool e/ou outras drogas por estudantes universitários e seus familiares durante a pandemia de COVID-19

Vítor Di Lucente Vieira Gonçalves

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

Brasília
2022

Vítor Di Lucente Vieira Gonçalves

Estudo dos padrões de consumo de álcool e/ou outras drogas por estudantes universitários e seus familiares durante a pandemia de COVID-19

Orientador: Prof. Luís Gustavo do Amaral Vinha

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

Brasília
2022

Resumo

O presente estudo tem como objetivo verificar e analisar possíveis fatores que podem ter influenciado no padrão de consumo de álcool e/ou outras drogas por estudantes universitários e seus familiares durante a pandemia de Covid-19.

Para identificar como as possíveis variáveis preditoras se comportam em relação à variável resposta, foi proposto um modelo de regressão logístíca binária e um modelo de regressão logística multinomial. Ambos os modelos apresentaram um bom ajuste porém o modelo multinomial não discrimina bem as observações

Palavras-chave: álcool; drogas; Covid-19; estudantes universitários; regressão logística; regressão logística multinomial.

Sumário

1 Introdução

A entrada na universidade é um momento muito importante para muitos jovens, pois representa uma nova fase com mudanças significativas. O cotidiano e a adaptação a essa nova realidade, proporcionam novas experiências e associadas a elas novos e distintos sentimentos, que influenciam na percepção do estudante sobre sua qualidade de vida e bem-estar (SILVA; HELENO, 2012). Esse período de transição pode influenciar ou interferir nos comportamentos individuais e/ou coletivos desses estudantes, e além disso existem vários problemas associados a exigências e dificuldades das demandas acadêmicas, sendo eles pessoais, interpessoais, sociais, e até mesmo referentes à identidade individual e/ou coletiva dos universitários (DAMASCENO et al., 2016).

Nessa nova etapa, o ambiente universitário pode permitir um fácil acesso às diversas drogas, dado que muitos dos comportamentos de risco associados ao consumo de álcool e/ou outras drogas são mais frequentes entre os estudantes universitários do que entre jovens da população geral e de faixa etária correspondente (ANDRADE; DUARTE; OLIVEIRA, 2010). Os mesmos autores também apontam que 86% dos universitários do Brasil já haviam feito uso de álcool na vida, 47% de produtos de tabaco e 49% de alguma substância ilícita. Além disso, 22% dos universitários em 2010 estavam sob risco de desenvolver dependência de álcool, 21% de derivados do tabaco e 8% de maconha.

Como um possível agravante desse contexto, o surto de COVID-19 ao redor do mundo iniciado no final de 2019, fez com que os governos de diversos países tomassem medidas para reduzir e controlar o contágio do vírus, como quarentena, isolamento e distanciamento social. Embora essas medidas terem sido aplicadas para proteger a saúde e integridade física, elas tem efeitos negativos na saúde mental em uma proporção considerável da população (RUBIN; WESSELY, 2020). Um estudo feito por Filgueiras e Stults-Kolehmainen (2020), mostrou que um mês após o decreto da quarentena, a amostra brasileira estudada demonstrou uma piora significativa em estresse, depressão e ansiedade. Além disso, vários índividuos dessa amostra reportaram altos níveis de sofrimento.

Os estudantes universitários representam uma parcela da população que sofreu muito com essas medidas restritivas. Um grande número de estudantes teve que lidar com os efeitos psicológicos causados pela quebra da rotina pessoal e suspensão do ensino presencial, o que pode desencadear em maior desconforto emocional (SILVA; ROSA, 2021). Segundo Gundim et al. (2020), a pandemia e seus fatores que interferem na rotina usual e na vida acadêmica dos estudantes e seus familiares são prejudiciais para a saúde mental dos universitários e da sociedade no geral. Como consequência dessa situação, muitos jovens e seus familiares podem ter utilizado o álcool e/ou outras drogas como ferramentas para se distrair e aceitar essa nova realidade na pandemia.

Tendo isso em vista, o objetivo deste estudo é analisar mudanças no padrão de consumo de álcool e/ou outras drogas dos estudantes universitários e suas famílias durante a pandemia, assim como identificar as características dos estudantes relacionadas a essas mudanças. Ao final desse estudo, espera-se contribuir com informações valiosas que possam auxiliar a administração da universidade na tomada de decisões e criação de programas e políticas em prol de seus estudantes.

2 Referencial Teórico

Nesta seção são apresentadas as técnicas utilizadas no presente estudo, a primeira subseção apresenta os testes utilizados para análise da associação das variáveis independentes com a variável resposta, e a segunda apresenta a regressão logística, uma vez que essa variável resposta é qualitativa.

2.1 Testes

2.1.1 Teste χ^2 de Independência

Esse teste tem como objetivo verificar se existe associação entre duas variáveis, sendo mais recomendado para variáveis qualitativas (principalmente nominais). O princípio básico deste método é comparar proporções, ou seja, as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Para esse teste, as hipóteses podem ser escritas como

$$\begin{cases} H_0 : \text{A variável X é independente da variável Y} \\ H_1 : \text{A variável X depende da variável Y.} \end{cases}$$

Este teste é baseado no cálculo dos valores esperados. Os valores esperados são os valores que seriam observados caso a hipótese nula fosse verdadeira, e são calculados pela expressão

$$e_{ij} = \frac{(total\ da\ linha\ i) \times (total\ da\ coluna\ j)}{total\ geral}\ .$$

A estatística de teste utilizada é a estatística Qui-quadrado de Pearson dada por

$$\chi^2_v = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

em que:

- e_{ij} é valor esperado na i-ésima linha e na j-ésima coluna;
- o_{ij} é valor observado na i-ésima linha e na j-ésima coluna;
- $v = (r-1)(s-1)$ representa o número de graus de liberdade, onde $r =$ é o número

total de linhas e $s =$ é o número total de colunas.

Então, sob a hipótese de H_0 ser verdadeira, a estatística do teste seguirá a distribuição χ^2_v. Para que a aproximação Qui-Quadrado seja satisfatória, é preciso que a amostra seja relativamente grande, com todos os valores esperados maiores ou iguais a 5 para cada casela ou no máximo 20% deles seja menor que 5, com todos maiores que 1.

Para verificar a intensidade de uma associação entre duas variáveis qualitativas, pode ser utilizado o coeficiente de contingência V de Cramer. Este coeficiente assume valores entre zero e um. O valor zero corresponde à ausência de associação entre as variáveis, logo valores próximos a zero correspondem a fraca associação e valores mais próximos a um correspondem a associação mais forte. Este coeficiente utiliza a estatística Qui-Quadrado para seu cálculo, da seguinte forma

$$V = \sqrt{\frac{\chi^2}{n(k-1)}},$$

em que:

- $\chi^2 =$ valor da estatística Qui-Quadrado;
- $n =$ tamanho da amostra;
- $k = min(r,\ s)$, ou seja, o mínimo entre o número de linhas e colunas.

2.1.2 Teste de Wilcoxon-Mann-Whitney

O teste de Wilcoxon-Mann-Whitney ou apenas Mann-Whitney é utilizado para comparar dois grupos independentes sem a suposição de distribuição específica. Isso ocorre pois o teste baseia-se em postos atribuídos a cada observação da variável quantitativa após serem ordenadas. O teste considera as hipóteses:

$$\begin{cases} H_0 : \text{As populações têm a mesma distribuição} \\ H_1 : \text{As populações têm distribuições distintas.} \end{cases}$$

A estatística do teste é dada por

$$W = \sum_{i=1}^{n} R(X_i),$$

em que:

- $R(X_i)$ é o posto atribuído ao i-ésimo elemento da amostra;
- n é o tamanho do grupo 1.

No entanto, quando são gerados muitos empates na atribuição dos ranks, a estatística do teste é dada por

$$W = \frac{T - E(T)}{\sqrt{V_c(T)}}$$

onde:

- $E(T) = \frac{n(N+1)}{2}$;
- $V_c(T) = \frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}$;
- R_i = posto da i-esima observação considerando os dois grupos;
- m é o tamanho do grupo 2;
- $N = n + m$.

Sob a hipótese de H_0 ser verdadeira, em grandes amostras e com muitos empates na atribuição dos ranks, W segue aproximadamente a distribuição normal padrão.

2.2 Regressão Logística

2.2.1 Análise de Regressão Logística Binária

A análise de regressão logística binária é um instrumento eficaz para verificar a relação entre duas ou mais variáveis no caso específico em que a resposta (Y) é dicotômica ou dicotomizada em "sucesso" ($Y = 1$) e "fracasso" ($Y = 0$). Com apenas uma variável explicativa, sua modelagem é feita a partir da equação

$$P(Y_i = 1|X_i) = \pi_i = \frac{e^{\alpha+\beta X_i}}{1 + e^{\alpha+\beta X_i}},$$

em que a probabilidade de sucesso da variável resposta ($Y = 1$) é função da variável explicativa X. Tal equação pode ser escrita de maneira linear aplicando a transformação *logito*

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta X_i.$$

A interpretação dos parâmetros do modelo de regressão logística está baseada na **Razão de Chances**. Se um evento ocorre com probabilidade p, a chance de ocorrência desse evento é definida como

$$\hat{chance} = \frac{p}{1-p},$$

isto é, a probabilidade de ocorrência do evento dividida pela probabilidade de não ocorrência do evento. A Razão de Chances (*odds ratio*) é o resultado da divisão das chances de ocorrência de um evento em dois grupos diferentes, ou seja, é a chance de ocorrência de um evento entre indivíduos expostos a algum fator de risco dividido pela chance de ocorrência do evento entre indivíduos não-expostos. Formalmente define-se como

$$\hat{RC} = \frac{\hat{chance}_1}{\hat{chance}_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}.$$

- Valores de $\hat{RC}$ menores que 1 indicam que a chance do evento ocorrer em indivíduos do grupo 2 é maior que a chance do evento ocorrer em indivíduos do grupo 1;
- Valores de $\hat{RC}$ próximos a 1 indicam que a chance do evento ocorrer no grupo 1 é semelhante a chance do evento ocorrer em indivíduos do grupo 2;
- Valores de $\hat{RC}$ maiores que 1 indicam que a chance do evento ocorrer em indivíduos do grupo 1 é maior que a chance do evento ocorrer em indivíduos do grupo 2.

O parâmetro β corresponde ao efeito do aumento de uma unidade de X sobre o logaritmo neperiano da chance de sucesso ($Y = 1$). Dessa forma, e^{β} tem como efeito a multiplicação na chance de $Y = 1$ para o aumento de uma unidade de X.

O modelo de regressão logística simples pode ser facilmente estendido para acomodar p variáveis explicativas. O modelo pode ser escrito como

$$P(Y_i = 1|X_{1i}, \ldots, X_{pi}) = \pi_i = \frac{e^{\alpha+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}}{1 + e^{\alpha+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}},$$

e, assim como no caso simples, pode ser escrito de maneira linear utilizando a transformação *logito*

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}.$$

Neste caso, o parâmetro β_j corresponde ao efeito do aumento de uma unidade de X_j sobre o logaritmo neperiano da chance de sucesso ($Y = 1$), mantendo as demais variáveis constantes. Dessa forma, e^{β_j} tem como efeito a multiplicação na chance de $Y = 1$ para o aumento de uma unidade de X_j, mantendo as outras variáveis constantes.

2.2.2 Análise de Regressão Logística Multinomial

A regressão logística multinomial pode ser vista como uma extensão do modelo logístico binário, em situações nas quais a variável dependente tem múltiplas categorias. Neste caso, a equação básica geral do modelo relaciona cada uma das c categorias com uma categoria de referência, que pode ser a última ou a mais frequente, escrita como

$$ln\left(\frac{\pi_j}{\pi_c}\right),\ j = 1, \ldots, c-1.$$

Essa equação representa o logaritmo neperiano da chance da resposta ser a categoria j, dado que a resposta pode ser ou j ou c. O modelo geral com uma única variável explicativa é dado por

$$ln\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j X.$$

O modelo tem $c-1$ equações, uma para cada par de categorias da variável resposta, e cada uma delas com parâmetros diferentes. Os efeitos desses parâmetros variam de acordo com a categoria pareada com a categoria de referência. Usando $c = 3$ como exemplo, com a terceira sendo a categoria de referência, as equações são

$$ln\left(\frac{\pi_1}{\pi_3}\right) = \alpha_1 + \beta_1 X$$
$$ln\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \beta_2 X.$$

As $c-1$ equações apresentadas acima também determinam os parâmetros de outros pares. Portanto, nessa formulação é necessário apenas especificar $c-1$ pares, uma vez que os demais são redundantes, como mostrado a seguir

$$\begin{aligned} ln\left(\frac{\pi_1}{\pi_2}\right) = ln\left(\frac{\pi_1/\pi_3}{\pi_2/\pi_3}\right) &= ln\left(\frac{\pi_1}{\pi_3}\right) - ln\left(\frac{\pi_2}{\pi_3)}\right) \\ &= (\alpha_1 + \beta_1 X) - (\alpha_2 + \beta_2 X) \\ &= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)\, X \end{aligned}$$

$$= \alpha_3 + \beta_3 X.$$

O modelo geral de regressão logística multinomial simples pode ser facilmente estendido para acomodar p variáveis explicativas. Ele pode ser escrito como

$$ln\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + \ldots + \beta_{jp}X_p.$$

O efeito de cada variável explicativa na variável resposta é diferente para cada um dos pares, e a menos que c e p sejam pequenos, esse modelo tem um grande número de parâmetros. A probabilidade estimada pelo modelo para cada uma das categorias é dada por

$$\pi_j = \frac{e^{\alpha_j + \beta_{j1}X_1 + \ldots + \beta_{jp}X_p}}{\sum_{h=1}^{c} e^{\alpha_h + \beta_{h1}X_1 + \ldots + \beta_{hp}X_p}}, \quad j = 1, \ldots, c.$$

Na equação, o denominador é o mesmo para todo j, logo $\sum_j \pi_j = 1$. Os parâmetros referentes a categoria de referência são todos iguais a zero. Para cada um dos pares, a intepretação dos parâmetros é feita como em modelos de regressão logística binária, condicionado ao evento de que a resposta foi uma das duas categorias no par.

2.2.3 Seleção de modelos

A seleção de modelos pode ser realizada por meio de critérios de informação. O Critério de Informação de Akaike (AIC) é um valor que pode ser utilizado para comparar e determinar qual modelo entre vários é o mais indicado para um determinado conjunto de dados. Ao ajustar modelos, é possível aumentar sua verossimilhança ao adicionar parâmetros, mas isso pode resultar em um sobre-ajuste. Por isso, no cálculo do AIC é introduzido um termo de penalização com base no número de parâmetros do modelo. O AIC é dado por

$$AIC = -2ln(\hat{L}) + 2k,$$

onde $\hat{L}$ é o valor máximo da função de verossimilhança do modelo e k é o número de parâmetros do modelo. Dado alguns modelos candidatos para os dados, o modelo preferido será o que apresentar o menor valor de AIC.

Assim como o AIC, o Critério Bayesiano de Schwarz (BIC) é um critério de seleção de modelos. No entanto, o termo de penalização é maior do que no critério anterior. O BIC é dado por

$$BIC = -2ln(\hat{L}) + kln(n),$$

onde $\hat{L}$ é o valor máximo da função de verossimilhança do modelo, k é o número de parâmetros do modelo e n é o tamanho da amostra.

Com o objetivo de facilitar e automatizar o processo de escolhas das variáveis que farão parte do modelo e visando selecionar o modelo que melhor explica a variável reposta, existem alguns métodos automáticos de seleção de variáveis, descritos abaixo. Os algoritmos de seleção utilizados neste estudo se baseiam no AIC dos modelos.

- *Forward*: neste método, inicialmente é ajustado o modelo nulo com nenhuma variável explicativa, e então cada nova variável que minimize o AIC quando comparada a inclusão de outras é selecionada para o modelo. O processo continua de forma progressiva enquanto a inserção de novas variáveis diminuir o AIC, caso contrário o processo é encerrado e assim é apresentado o modelo indicado por este método.

- *Backward*: já neste método, inicialmente é ajustado o modelo saturado com todas as possíveis variáveis explicativas. A partir disso, é retirada a variável com que sua eliminação minimize o AIC do modelo, quando comparada à eliminação das outras variáveis. O processo continua até a eliminação das variáveis restantes não resultar em uma diminuição do AIC do modelo.

- *Stepwise*: método que une os métodos de seleção *Forward* e *Backward*, já que inicialmente também é ajustado o modelo nulo, e a partir dele são incluídas e retiradas variáveis que minimizem o AIC. Neste método, após cada variável ser incluída no modelo é verificado novamente se a variável incluída no passo anterior deve permanecer ou ser retirada do modelo com base no AIC. O método prossegue até que a inserção ou eliminação de variáveis não diminua o AIC do modelo.

A seleção de modelos também pode ser realizada utilizando testes de significância. O teste de Razão de Verossimilhança pode ser utilizado para comparar 2 modelos aninhados e verificar se o modelo mais simples se ajusta bem aos dados. O teste é feito da seguinte maneira:

- Sejam $\mathcal{L}_M$ e $\mathcal{L}_S$ as verossimilhanças do modelo M e do modelo saturado S, respectivamente, então

$$G^2 = \text{Deviance} = -2[log(\mathcal{L}_M) - log(\mathcal{L}_S)].$$

- Suponha que o modelo M_0 é um modelo aninhado ao modelo M. Dado que M é adequado, a estatística para testar se o modelo mais simples é adequado aos dados é dado por

$$\begin{aligned} Q_L &= -2[log(\mathcal{L}_{M0}) - log(\mathcal{L}_M)] \\ &= -2[log(\mathcal{L}_{M0}) - log(\mathcal{L}_S)] - 2[log(\mathcal{L}_M) - log(\mathcal{L}_S)] \\ &= \text{Deviance}_0 - \text{Deviance}. \end{aligned}$$

- Assim, pode-se comparar os modelos através da comparação de Deviances. Para grandes amostras essa estatística segue distribuição aproximadamente Qui-quadrado com graus de liberdade igual a diferença entre os graus de liberdade dos resíduos nos dois modelos.

2.2.4 Diagnóstico do modelo

A análise de diagnóstico deve ser realizada para avaliar a adequação do modelo proposto. Um dos problemas que pode ser observado no ajuste de modelos é a multicolinearidade, ou seja, duas ou mais variáveis explicativas que são correlacionadas e tornam impossível distinguir seus efeitos individuais na variável resposta. O Fator de Inflação da Variância (VIF - *Variance Inflation Factor*) é uma medida comumente utilizada para verificar a existência de multicolinearidade no modelo. O VIF de uma determinada variável explicativa X_j pode ser calculado ajustando um modelo de regressão linear com X_j sendo a variável resposta, em função das demais variáveis explicativas. Tem-se então

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2},$$

onde R_j^2 é o coeficiente de determinação da regressão linear ajustada.

Porém, essa medida não é aplicável a modelos que contém preditores categóricos com mais de 2 categorias. Tem-se então o Fator de Inflação da Variância generalizado (gVIF), desenvolvido por Fox e Monette (1992). Para variáveis explicativas contínuas, o gVIF é igual ao VIF. Para reduzir o gVIF a uma medida linear, foi sugerido também pelos mesmos autores usar $gVIF^{(1/2gl)}$, onde gl é o número de variáveis indicadoras advindas da variável categórica. Valores de $(gVIF^{(1/2gl)})^2$ ou de VIF inferiores a 5 indicam ausência de multicolinearidade.

Outro problema que pode ocorrer são observações influentes, que são aquelas que alteram as estimativas quando retiradas do conjunto de dados. Existem algumas medidas que analisam a influência que as observações exercem na estimativa dos parâmetros e/ou

da resposta. Algumas delas são:

- DFBETA: corresponde a alteração padronizada na estimativa do parâmetro quando a i-ésima observação é eliminada. Valores acima de $\frac{2}{\sqrt{n}}$ caracterizam uma observação influente;
- Distância de Cook: indica o quanto os valores estimados pela regressão se alteram quando a i-ésima observação é eliminada. Valores acima de um indicam uma observação influente, porém é interessante verificar observações com valores acima de 0,5.

Além dessas medidas, o teste de Hosmer e Lemeshow é um teste de qualidade do ajuste para modelos logísticos, que indica se há uma falta de ajuste dos dados ao modelo proposto. Especificamente, esse teste verifica se as proporções de eventos observados correspondem às proporções de eventos esperados em g grupos da amostra. O teste é feito da seguinte maneira:

- As n/g observações com maiores probabilidades estimadas são colocadas no primeiro grupo, as seguintes são colocadas no segundo grupo e assim por diante;
- Para cada grupo, os valores esperados são obtidos pela soma das probabilidades estimadas para as observações daquele grupo;
- Por fim, calcula-se a estatística Qui-quadrado de Pearson, X^2, a partir da tabela $g \times 2$ de frequências observadas e esperadas e rejeita-se a hipótese de que o modelo testado exibe um bom ajuste se $X^2 > \chi^2_{g-2}$.

Como a variável resposta é binária, o resíduo simples desse modelo pode assumir apenas dois valores, $1 - \pi_i$ (se $Y_i = 1$) e $-\pi_i$ (se $Y_i = 0$), e portanto esse resíduo não é útil para identificar faltas de ajustamento do modelo. Em seu lugar, pode ser utilizado o resíduo Componente de Desvio (Deviance), dado para cada observação por

$$d_i = sgn(y_i - \hat{y}_i) \left[2y_i \; ln\left(\frac{y_i}{\hat{y}_i}\right) + 2\left(n_i - y_i\right) ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]^{1/2},$$

onde:

- sgn é a função sinal, que retorna o sinal de $y_i - \hat{y}_i$
- y_i é o número de sucessos no i-ésimo conjunto (cenário) de variáveis explicativas, $i = 1, \ldots, N$;

- n_i é o número de eventos no i-ésimo conjunto;
- $\hat{\pi}_i$ é a probabilidade estimada de sucessos no i-ésimo conjunto;
- $\hat{y}_i = n_i\hat{\pi}_i$ é o número estimado de sucessos no i-ésimo conjunto.

Valores desse resíduo maiores que 3 fornecem evidências de falta de ajustamento do modelo. A partir desse resíduo, podem ser elaborados gráficos que ajudam no diagnóstico do modelo

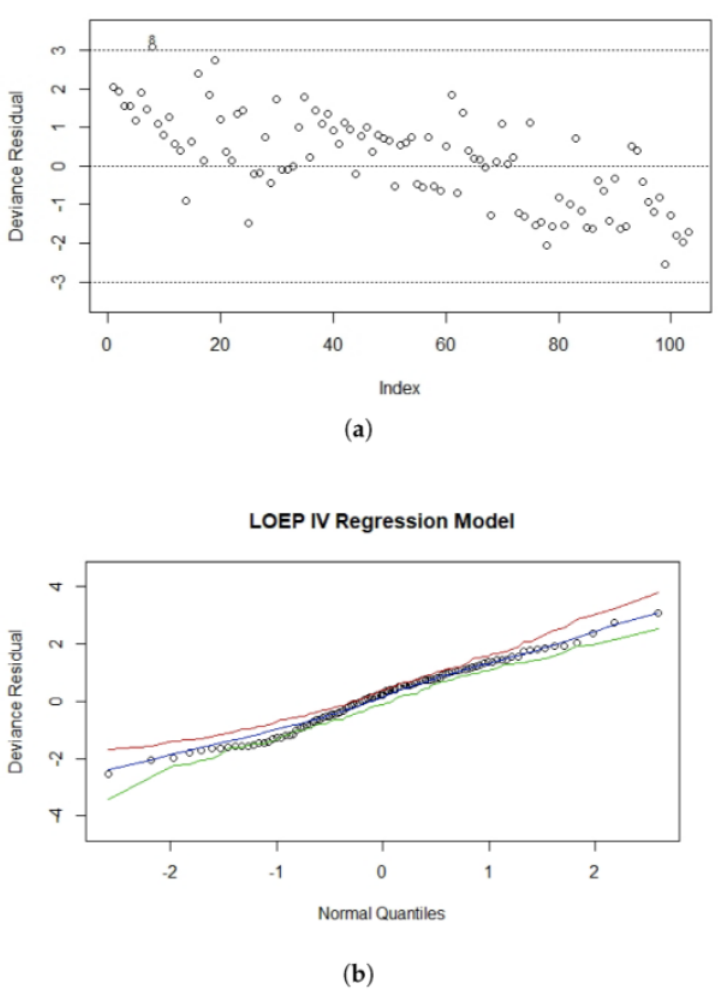

Figura 1: Exemplo de gráficos de resíduos e de envelope
Fonte: Baharith, Lamya A. et al *The Odds Exponential-Pareto IV Distribution: Regression Model and Application* (2020)

A Figura 1 (**a**) apresenta o gráfico de resíduos deviance pelos índices das observações. Para esse gráfico apontar um bom ajuste dos dados, os pontos devem apresentar comportamento aleatório, distribuídos em torno de zero e com variância constante. A Figura 1 (**b**) mostra o gráfico de envelope dos resíduos. Quando a grande maioria dos pontos se encontra dentro das bandas de confiança, esse gráfico indica que não existem indícios de afastamento da suposição de distribuição Binomial para a variável resposta.

2.2.5 Poder preditivo

A qualidade do modelo de regressão logística também pode ser avaliado a partir do seu poder preditivo. O poder preditivo pode ser expresso pelas classificações corretas dado o ajuste do modelo, ou seja, sua acurácia. Neste trabalho, o poder preditivo será avaliado a partir da Sensibilidade e Especificidade do modelo. A Sensibilidade é a probabilidade do modelo classificar corretamente os indivíduos que contém a característica de interesse, e a Especificidade é a probabilidade do modelo classificar corretamente os indivíduos que não contém a característica de interesse.

Além dessas medidas, a curva ROC é um gráfico da Sensibilidade em função de 1 menos a Especificidade para diferentes valores de π_i. Esta curva resume o poder preditivo do modelo, e quanto maior a área abaixo da curva, melhor é a performance do modelo em distinguir as categorias da variável resposta.

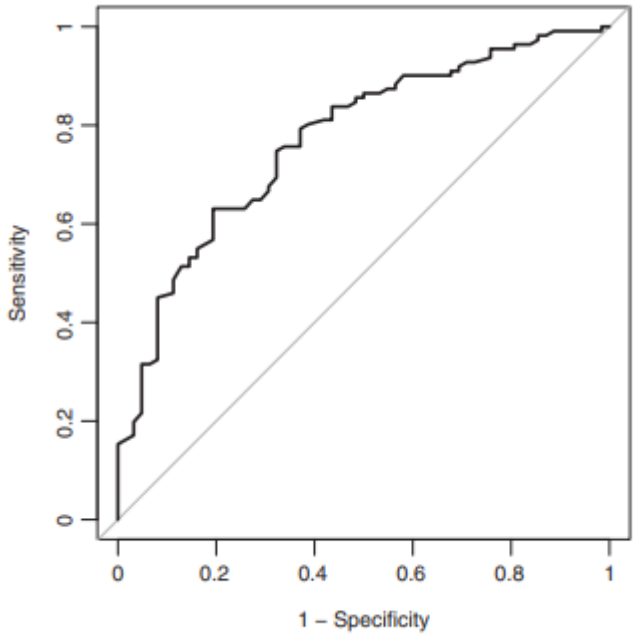

Figura 2: Exemplo de gráfico da curva ROC
Fonte: Agresti, A. *An Introduction to categorical data analysis* (2018)

3 Análises Descritivas

3.1 Caracterização da amostra

Os dados utilizados neste estudo foram coletados em uma universidade brasileira. Para obter esses dados, primeiro foi feito um levantamento em novembro de 2020 de todos os estudantes com matrícula ativa. Do total de estudantes da universidade, foi sorteada uma amostra aleatória e estratificada por unidade acadêmica, em seguida, foram enviados e-mails para os participantes sorteados, convidando-os a participarem da pesquisa através de um link anexado ao corpo do e-mail. O período de coleta de dados durou de 16/03/2021 até 19/07/2021. Como o presente estudo utiliza dados que foram coletados por meio de questionários enviados por e-mail, esses dados podem não refletir a realidade da universidade, uma vez que pode existir um viés de seleção entre aqueles que responderam. O número total de participantes foi de 1.091 estudantes universitários.

As variáveis contidas no banco de dados que serão estudadas são apresentadas na Tabela 1.

Tabela 1: Descrição e categorias das variáveis independentes

Variável	**Descrição e categorias**
Idade	Idade em anos
Identidade de Gênero	Homem Cis; Mulher Cis; Outros.
Orientação Sexual	Heterossexual; Homossexual; Bissexual; Outros.
Cor/Raça	Branca/amarela; Parda; Preta; Outros.
Estado civil	A partir dessa variável foi criada uma nova que indica a presença de um parceiro: Sim, possui parceiro (casado, união estável); Não possui parceiro (solteiro, separado,...).
Trabalho	Se o estudante trabalha: Não; Sim.
Local de Residência	Cidade Satélite; Plano Piloto/Lagos; Entorno do DF.
Nacionalidade	Se o estudante é brasileiro ou não.
Área do Curso	Os cursos foram divididos em três grandes áreas: Humanas; Biológicas; Exatas.
Reserva de Vagas	A partir dessa variável foi criada uma indicadora: Sim, se ingressou por meio de algum tipo de reserva de vaga; Não, caso contrário.
Apoio	Se o aluno recebe algum apoio da universidade: Não/Prefere não declarar; Apoio Social; Apoio Pesquisa.
Perdas financeiras	Se houveram perdas financeiras na família do estudante durante a pandemia: Não; Sim, mas sem dificuldades; Sim, com dificuldades.
Contraiu COVID-19	Não, Ninguém contraiu; Sim, um ou mais familiares; Sim, o respondente; Sim, o estudante e outro(s) familiar(es).
Óbito por COVID-19	Se alguém na família do estudante foi a óbito por COVID-19: Não; Sim, uma pessoa; Sim, mais de uma pessoa.
Atividades de Lazer	Variável com mais de uma opção de resposta. Foi criada uma variável com o número de atividades assinaladas.

A Figura 3 **A** revela que a grande maioria dos respondentes dessa pesquisa (64,16%) se identificaram como mulheres cis, e somadas aos que se identificaram como homens cis, totalizaram 1.057 respondentes (96,88%). Nota-se também, com a Figura 3 **C**, que 1.000 estudantes de graduação (91,66%) responderam não possuir parceiro e 91 (8,34%) relataram possuir parceiro. Verifica-se ainda, com a Figura 3 **B**, que a *Orientação Sexual* mais comum nesse grupo de estudantes foi a Heterossexual, com 696 das respostas (63,79%), e que a menos comum foi a categoria Homossexual, totalizando 84 respostas (7,70%). Na análise da variável *Cor/Raça*, a categoria com maior frequência foi a Branca/Amarela, com 543 estudantes (49,77%), seguida pela categoria parda, com 377 estudantes (34,56%). A categoria "Outros"corresponde a estudantes que responderam ser mestiços, manelucos, indígenas e que preferiram não declarar.

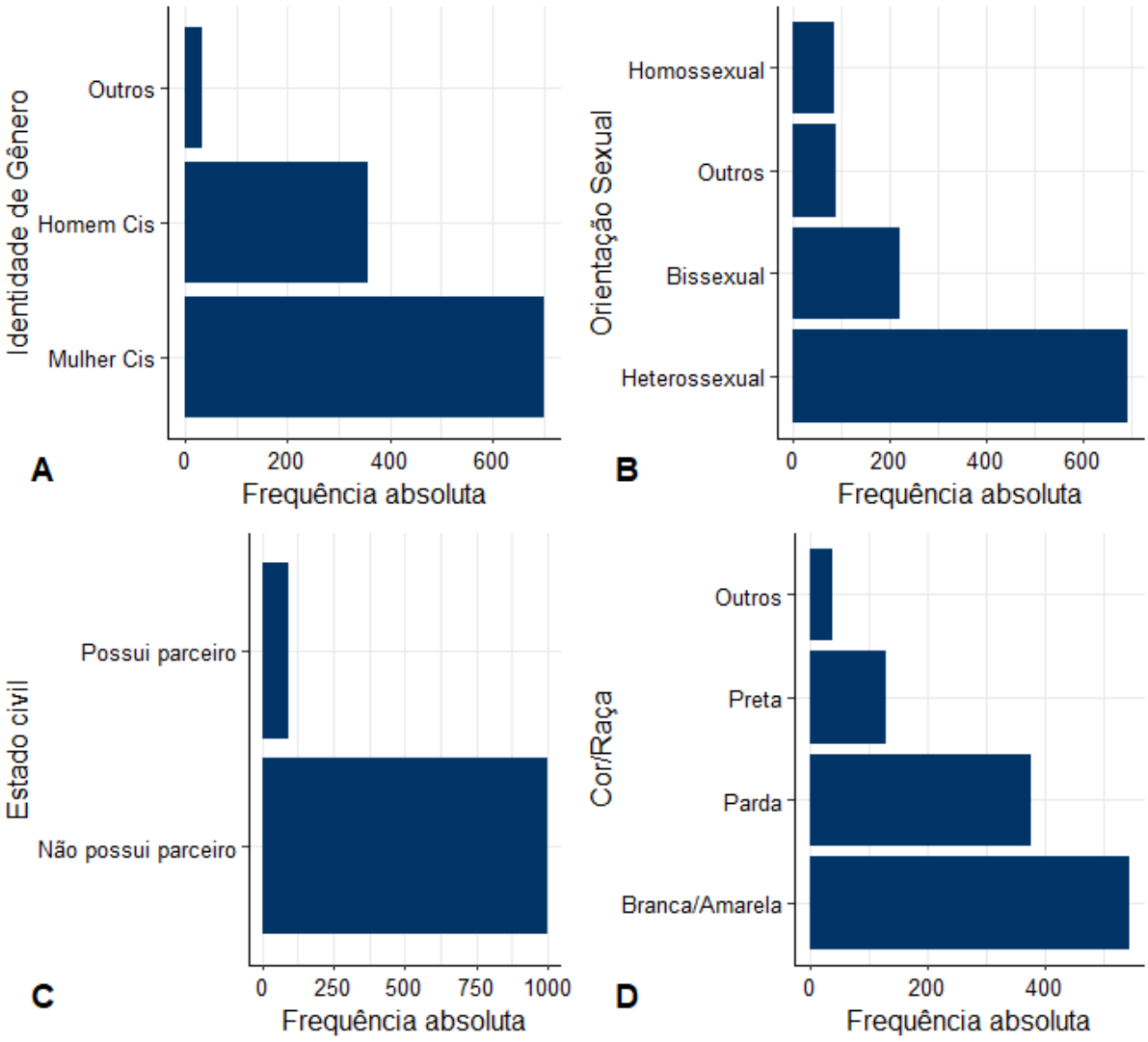

Figura 3: Distribuição das variáveis: *Identidade de Gênero*, *Orientação Sexual*, *Estado civil* e *Cor/Raça*

A Figura 4 **A** mostra a distribuição da variável *Idade*. O estudante de graduação mais novo observado na amostra possui 17 anos e o mais velho 69, ou seja, existe uma amplitude de 52 anos na variável Idade. O desvio interquartílico foi de 4 anos, ou seja, no intervalo entre 20 e 24 anos, estão os 50% dos alunos que tem idades mais centralizadas. A idade média observada foi de 23,44 anos e o coeficiente de variação, igual a 28,4%, indica que as idades estão razoavelmente dispersas. Com a Figura 4 **B**, nota-se que a maioria dos estudantes (69,38%) praticam de 2 a 3 atividades de lazer, seguidos pelos 212 estudantes que praticam 1 atividade (19.43%) e pelos 120 estudantes que praticam 4 atividades (11,0%). Apenas 2 respondentes relataram praticar 5 atividades de lazer diferentes durante a pandemia.

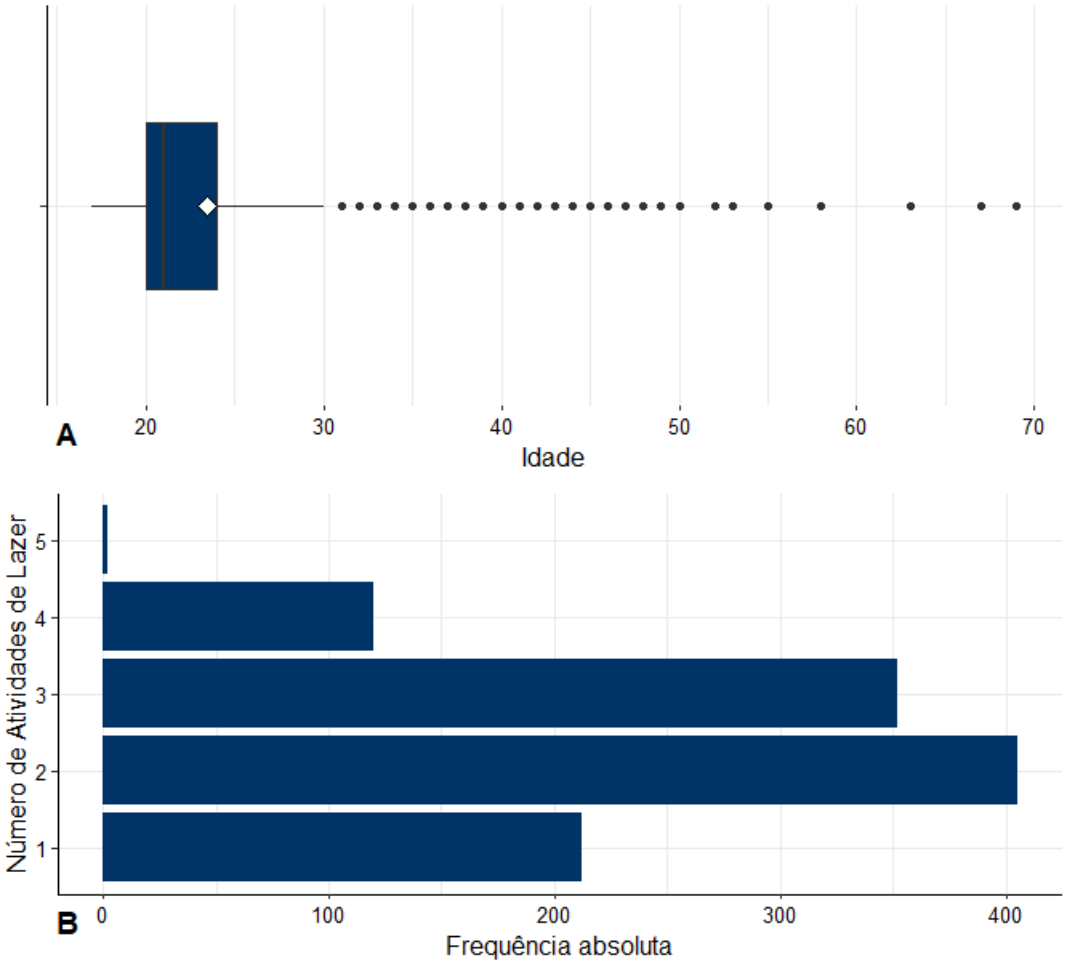

Figura 4: Distribuições da variável *Idade* e *Atividades de Lazer*

Pode-se observar com a Figura 5 **A** que a maioria dos estudantes de graduação da amostra residem em Cidades Satélites (60,68%), seguidos pelos que residem no Plano Piloto/Lagos Sul e Norte (24,56%) e por fim os alunos que moram no entorno do DF (14,76%). A Figura 5 **B** mostra que 366 dos estudantes (33,55%) trabalham integralmente ou parcialmente, e 725 estudantes não trabalham. Apenas 2,57% dos estudantes

não possuem nacionalidade brasileira.

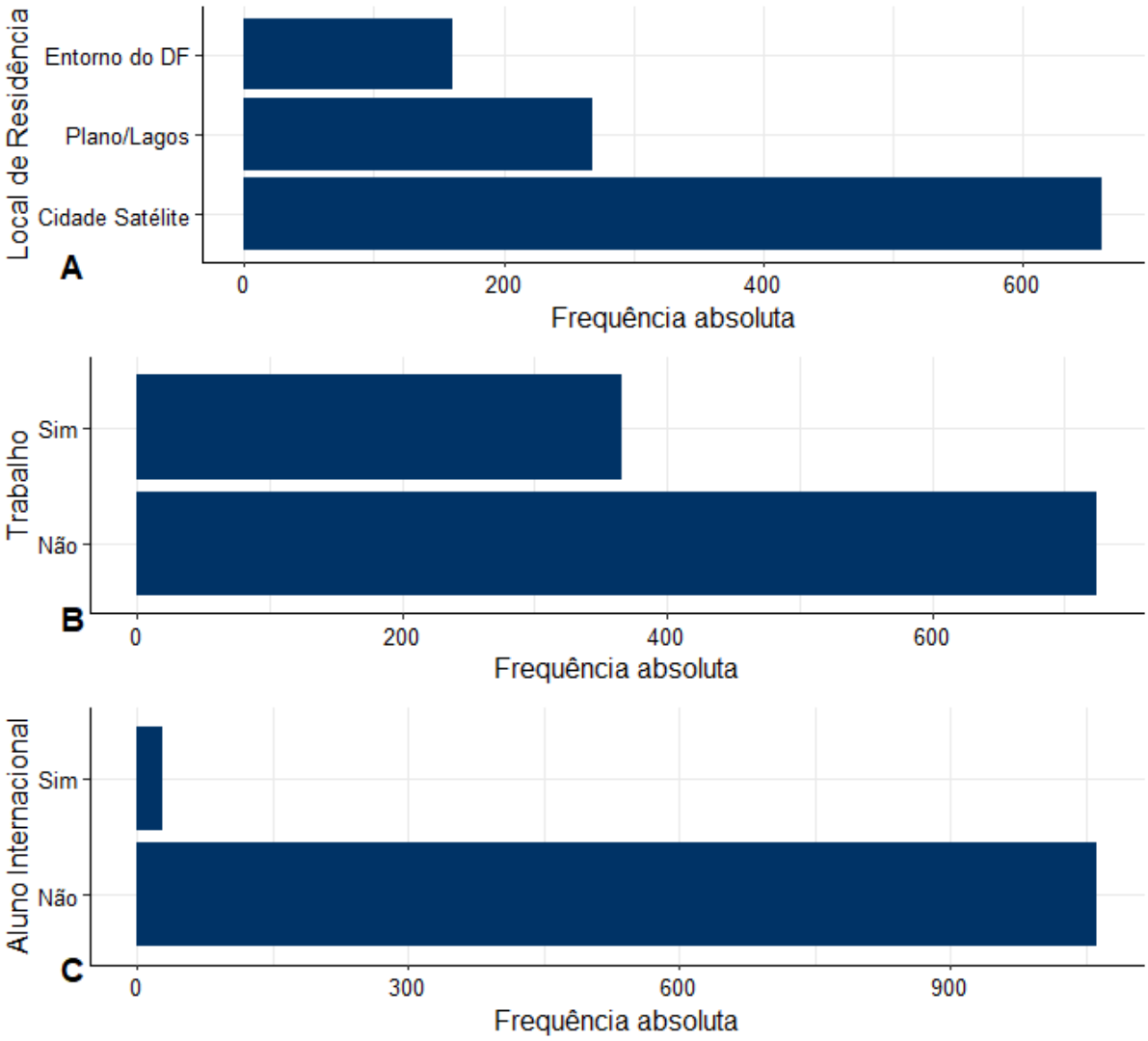

Figura 5: Distribuição das variáveis: *Local de Residência*, *Trabalho* e *Aluno Internacional*

Dos 1.091 estudantes na amostra, 491 deles (45,0%) revelaram ter entrado por algum sistema de reserva de vagas, como cotas ou transferência, visto na Figura 6 **B**. A Figura 6 **A** identifica que 822 dos estudantes (74,48%) não recebem ou preferiram não declarar receber nenhum tipo de apoio da universidade, 183 recebem algum apoio social (bolsa moradia, bolsa alimentação, etc) e 86 deles recebem algum apoio de pesquisa (bolsa pesquisa, bolsa de extensão, etc). Os alunos que fazem algum curso de Exatas foram os menos frequentes na amostra (22,91% dos estudantes), e os alunos que estão em algum curso de Humanas apresentaram a maior quantidade de respondentes (47,85%).

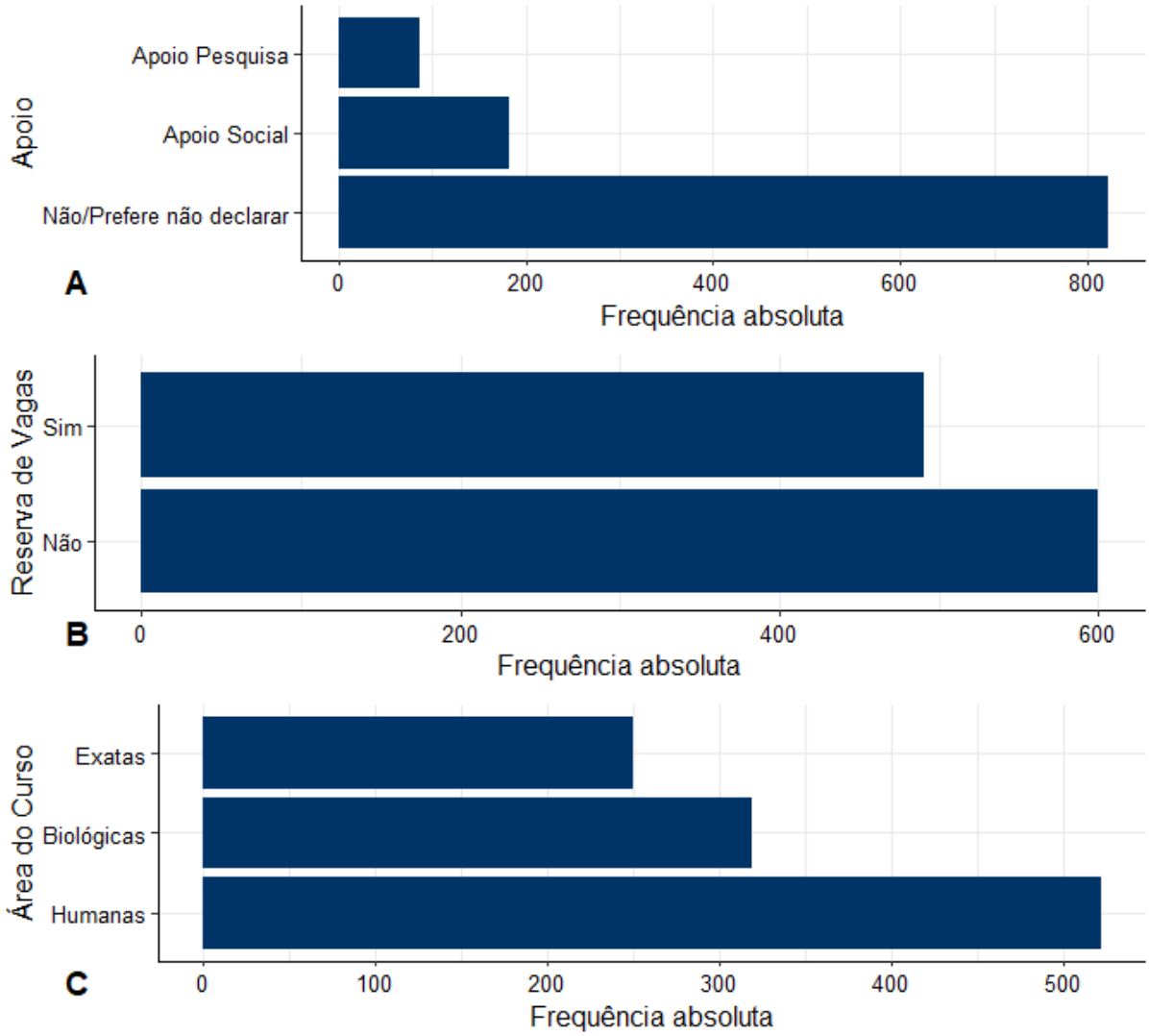

Figura 6: Distribuição das variáveis: *Apoio*,*Reserva de Vagas* e *Área do Curso*

Com a Figura 7 **A** verifica-se que 878 estudantes (80,48%) relataram ter perdas financeiras ou aumento de despesas durante a pandemia, dos quais 512 estavam admnistrando bem e 366 estavam com dificuldades para se adaptar. De todos os 1.091 respondentes, 191 disseram que ninguém da família contraiu Covid-19, 708 afirmaram que um ou mais familiares contrairam e 182 contrairam juntamente a um ou mais familiares. A quantidade de estudantes que relatou ter tido algum óbito na família por Covid-19 foi de 320 (29,33%), onde 164 revelaram a ocorrência de 1 óbito na família, e os outros 156 disseram ter mais de um óbito na família.

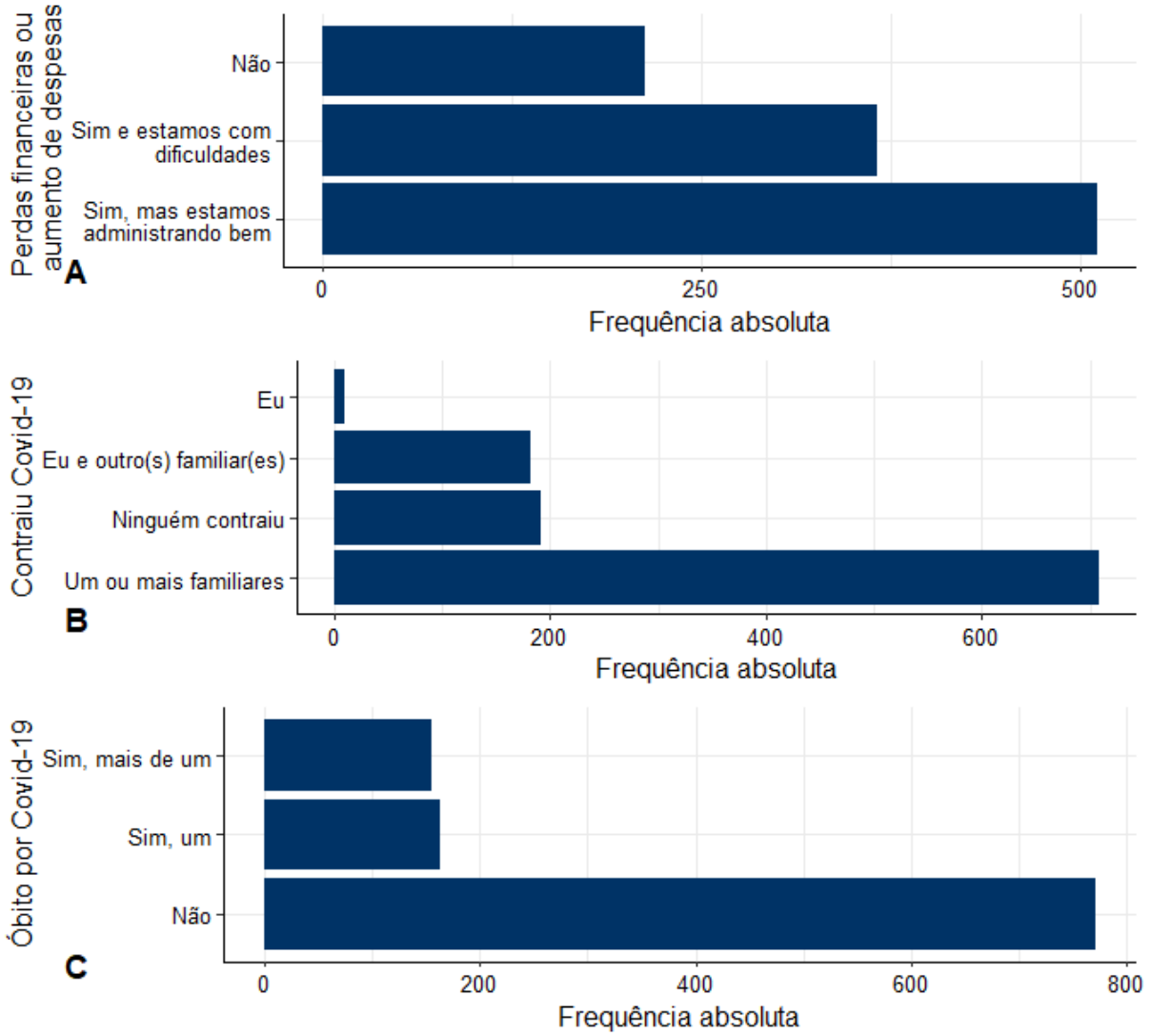

Figura 7: Distribuição das variáveis: *Perdas financeiras*, *Contraiu Covid-19* e *Óbito por Covid-19*

Observa-se na Figura 8 que a maioria dos estudantes relataram não ter havido uma mudança no consumo de álcool e/ou outras drogas por eles e seus familiares. Dentre os estudantes que identificaram mudanças nos hábitos de consumo em suas residências, a mudança mais frequente foi o aumento do consumo próprio dos respondentes, seguida pelo aumento do consumo próprio e de outras pessoas da residência. A mudança menos frequente foi a que outras pessoas na residência dos estudantes iniciaram o consumo, em que apenas 5 estudantes marcaram essa alternativa.

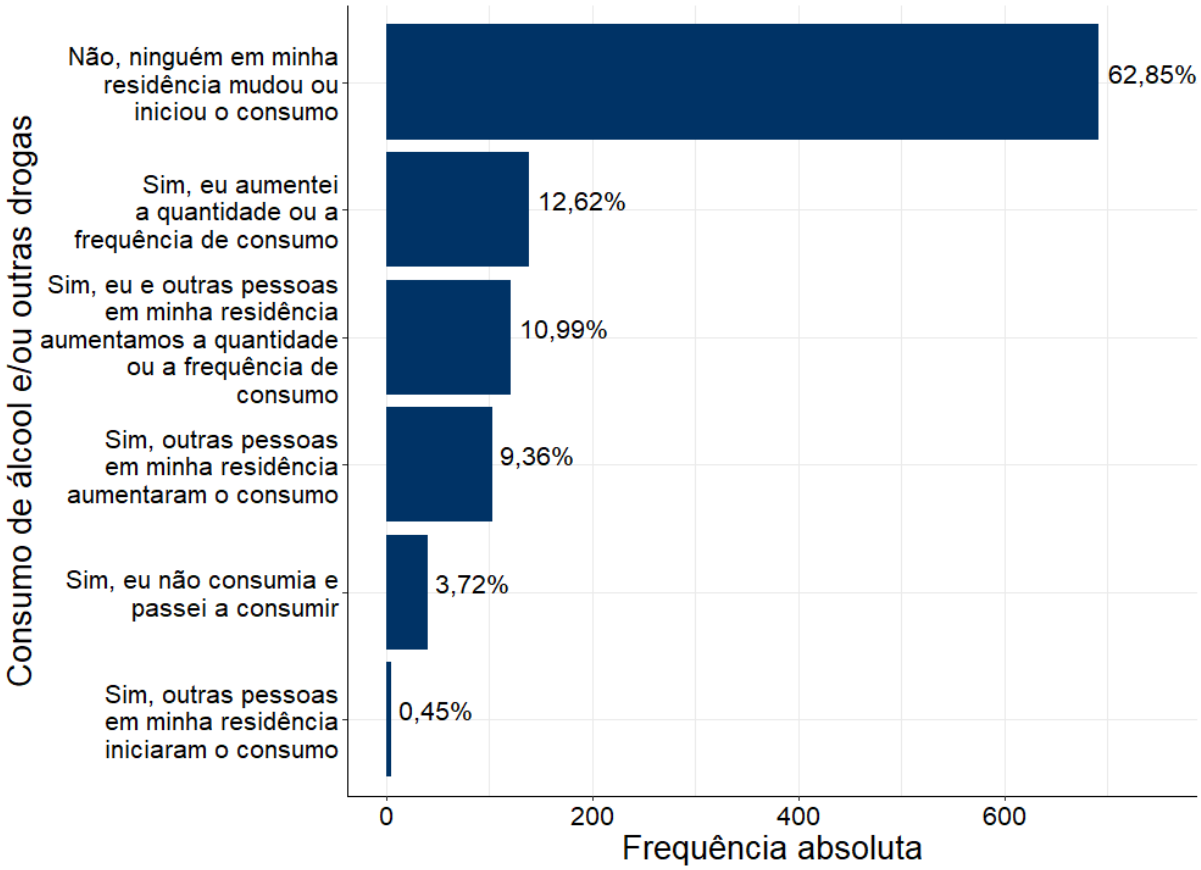

Figura 8: Distribuição da variável *Consumo de álcool e/ou outras drogas*

3.2 Estudo das associações entre as variáveis independentes e a resposta

Como foi observado na seção anterior, a variável resposta apresenta classificações com poucas observações, logo decidiu-se juntar algumas das categorias e transformá-la em uma variável com 4 categorias diferentes. Mesmo com essa nova categorização, dado o tamanho da amostra muitas caselas ficaram com poucas observações, o que inviabilza os testes e também traz resultados não satisfatórios no ajuste do modelo multinomial. A partir desse ponto, a descrição da relação entre a variável resposta e as independentes será realizada com a utilização da seguinte categorização: "Não houveram mudanças nos hábitos de consumo doméstico de álcool e/ou outras drogas durante a pandemia"; "Houveram mudanças nos hábitos de consumo doméstico de álcool e/ou outras drogas durante a pandemia". Na Tabela 2 são apresentados resultados dos testes χ^2 de independência considerando a variavel com essa nova categorização.

Tabela 2: Estatísticas χ^2, seus graus de liberdade, p-valores e V de Cramer

Variável Independente	χ^2	**gl**	**P-valor**	V **de Cramer**
Identidade de Gênero	0,556	2	0,757	0,023
Orientação Sexual	32,204	3	<0,001	0,172
Estado civil	<0,001	1	0,994	0,004
Cor/Raça	3,066	3	0,381	0,053
Local de Residência	9,600	2	0,008	0,094
Trabalho	2,000	1	0,157	0,045
Aluno Internacional	0,005	1	0,942	0,008
Apoio	6,359	2	0,042	0,076
Reserva de Vagas	0,656	1	0,418	0,026
Área do Curso	10,434	2	0,005	0,098
Perdas financeiras	3,426	2	0,180	0,056
Contraiu Covid-19	4,428	3	0,218	0,064
Óbito por Covid-19	0,204	2	0,903	0,014

Os resultados para os testes observados na Tabela 2, a um nível de significância de 5%, indicam que as variáveis significativamente associadas com o *Consumo de álcool e/ou outras drogas* foram: *Orientação Sexual*, *Local de Residência*, *Apoio* e *Área do Curso*. Embora essas variáveis tenham apresentado associação significativa, todos os V de Cramer indicam uma associação fraca.

Tabela 3: Teste de Mann-Whitney para as variáveis *Idade* e *Atividades de Lazer* por *Consumo de álcool e/ou outras drogas*

Variáveis Independentes	W	**P-valor**
Idade	139.426	0,851
Atividades de Lazer	142.963	0,349

A normalidade de ambas as variáveis foi rejeitada com p-valores muito pequenos, portanto para verificar se existe diferença na distribuição dessas variáveis pelos níveis da variável resposta, foi feito o teste de Mann-Whitney, que apresentou os resultados descritos na Tabela 3. A hipótese nula do teste de Mann-Whitney não foi rejeitada em ambas as variáveis, portanto não existem evidências estatísticas para afirmar que os indíviduos que relataram ou não presenciar mudanças nos hábitos de consumo doméstico de álcool e/ou outras drogas se diferenciam em suas distribuições de Idade e de quantidade de atividades de lazer praticadas.

4 Modelagem dos dados: Regressão Logística Binária

Nesta seção será escolhido o modelo que melhor explique o comportamento da variável resposta *Consumo de álcool e/ou outras drogas* em função das outras variáveis contidas no banco de dados. Para esta análise, a variável resposta será agrupada em 2 categorias: se o estudante ou algum familiar iniciou ou aumentou o consumo durante a pandemia e se não houve mudanças no consumo.

4.1 Seleção de variáveis

Para selecionar as variáveis que melhor explicam o padrão de consumo de álcool e/ou outras drogas durante a pandemia, foram utilizados inicialmente três algoritmos automáticos para uma primeira seleção de variáveis (*Forward*, *Backward* e *Stepwise*). Os três algoritmos utilizados apontam as mesmas variáveis preditoras: *Orientação Sexual*, *Local de Residência*, *Apoio*, *Idade*, *Trabalho*, *Estado civil* e *Área do Curso*. A Tabela 4 mostra os resultados do ajuste deste modelo.

Tabela 4: Resultados do modelo selecionado automaticamente

Coeficiente	Estimativa	Erro Padrão	Z	P-valor
Intercepto	0,343	0,349	0,981	0,327
Orientação Sexual - Bissexual	0,832	0,162	5,131	<0,001
Orientação Sexual - Homossexual	0,174	0,248	0,705	0,481
Orientação Sexual - Outros	0,416	0,237	1,756	0,079
Residência - Cidade Satélite	-0,542	0,158	-3,442	<0,001
Residência - Entorno do DF	-0,455	0,220	-2,068	0,038
Apoio - Pesquisa	-0,813	0,272	-2,994	0,003
Apoio - Social	0,102	0,177	0,575	0,565
Idade	-0,041	0,013	-3,172	0,001
Trabalho - Sim	0,244	0,144	1,691	0,091
Estado civil - Possui parceiro	0,411	0,275	1,492	0,136
Área do Curso - Biológicas	0,032	0,187	0,171	0,864
Área do Curso - Humanas	0,325	0,168	1,939	0,052

Observando os p-valores dos testes individuais de nulidade para os coeficientes, a nível de significância de 5%, nota-se que as variáveis *Trabalho* e *Estado civil* não eram significativas, e que o curso de graduação apresenta p-valor muito próximo do nível de significância adotado quando comparando humanas com exatas. Retirando as variáveis *Trabalho* e *Estado civil*, chega-se ao modelo apresentado na Tabela 5.

Tabela 5: Resultados do modelo reduzido

Coeficiente	Estimativa	Erro Padrão	Z	P-valor
Intercepto	0,142	0,321	0,445	0,657
Orientação Sexual - Bissexual	0,822	0,161	5,101	<0,001
Orientação Sexual - Homossexual	0,136	0,246	0,554	0,580
Orientação Sexual - Outros	0,380	0,235	1,613	0,107
Residência - Cidade Satélite	-0,532	0,157	-3,389	<0,001
Residência - Entorno do DF	-0,470	0,219	-2,146	0,032
Apoio - Pesquisa	-0,826	0,270	-3,059	0,002
Apoio - Social	0,077	0,176	0,437	0,662
Idade	-0,028	0,011	-2,559	0,010
Área do Curso - Biológicas	0,027	0,186	0,147	0,883
Área do Curso - Humanas	0,348	0,167	2,087	0,037

Nota-se que nesse modelo todas as variáveis apresentam pelo menos um coeficiente significativo e os coeficientes estimados não foram tão distintos aos do modelo anterior. Os testes de razão de verossimilhança exibidos na Tabela 6 revelam que o modelo saturado não é mais relevante que o modelo selecionado automaticamente e que o modelo selecionado automaticamente não é mais relevante que o modelo reduzido. A Tabela 7 mostra que os valores de AIC foram próximos entre os modelos automático e reduzido e o valor de BIC foi inferior no modelo reduzido. Como o teste de razão de verossimilhança apontou não existir diferença significativa entre o modelo automático e o reduzido, não houve mudanças significativas nos coeficientes estimados por esses dois modelos e pelo princípio da parcimônia, o modelo escolhido neste estudo será o modelo reduzido.

Tabela 6: Resultados da comparação dos modelos

Modelos	g.l.	χ^2	P-valor
Saturado x Automático	22	18,452	0,679
Automático x Reduzido	2	5,008	0,082

Tabela 7: Medidas AIC e BIC

Modelo	AIC	BIC
Automático	1394,018	1458,951
Reduzido	1395,027	1449,970

4.2 Qualidade do ajuste e diagnóstico

Nesta seção serão apresentados gráficos, medidas e testes que ajudam a identificar se o modelo é adequado. A Tabela 8 contém informações sobre a qualidade do ajuste.

Tabela 8: Medidas de qualidade do ajuste do modelo reduzido

Medida	Resultado
Máximo dos DFBETAS	0,061
Máximo das Distâncias de Cook	0,014
Máximo dos gVIFs	1,025
Teste de Hosmer e Lemeshow	0,223

A Tabela 8 indica ausência de valores influentes nos dados, dado que todos os valores de DFBETAS e das Distâncias de Cook foram pequenos. Nota-se também que não existe multicolinearidade neste modelo, visto que todos os fatores de inflação da variância generalizados calculados foram inferiores a 5. Por fim, o teste de qualidade de ajuste de Hosmer e Lemeshow não rejeita a hipótese de que o modelo está bem ajustado aos dados.

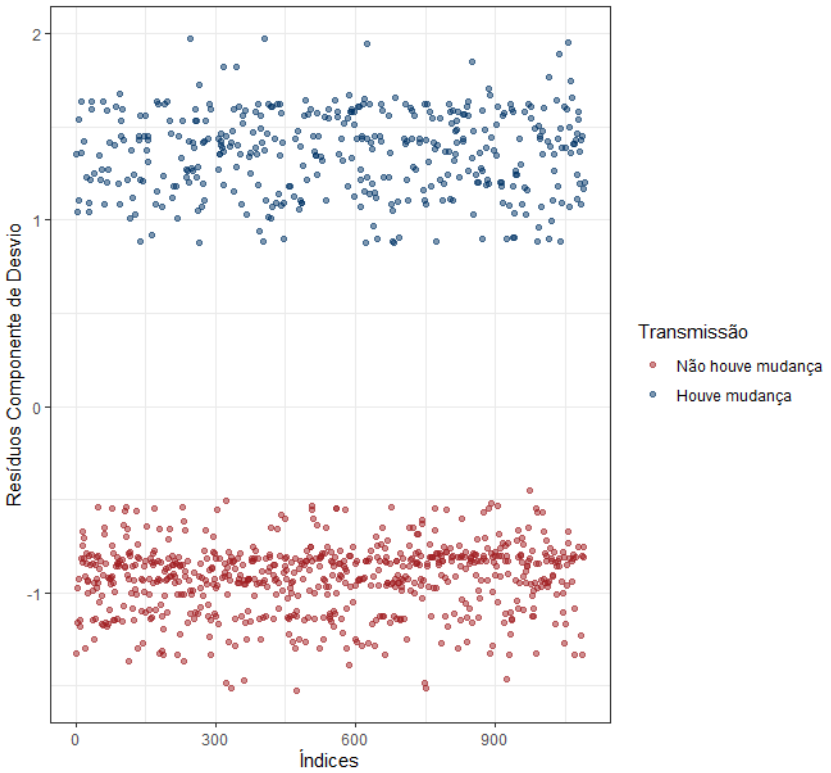

Figura 9: Gráfico dos resíduos Deviance pelo índice

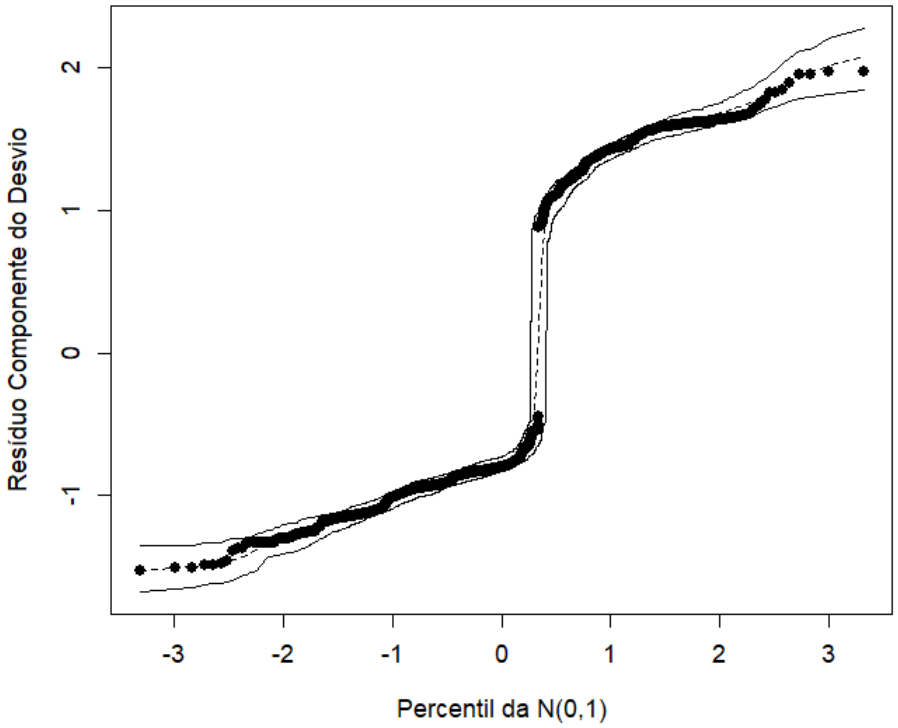

Figura 10: Gráfico de envelope dos resíduos Deviance

A Figura 9 mostra que a grande maioria dos resíduos tem módulo menor do que 2, e que aparentam ter comportamento aleatório, distribuídos em torno de zero com variância constante. A Figura 10 revela que todos os resíduos se encontram dentro da banda de confiança indicando um bom ajuste dos dados.

4.3 Poder Preditivo do Modelo

A Figura 11 indica uma boa discriminação do modelo, dado que a Sensibilidade é maior que um menos a Especificidade em todos os pontos do gráfico. A Tabela 9 contém algumas medidas que podem ser utilizadas para observar o desempenho do modelo em classificar corretamente as observações.

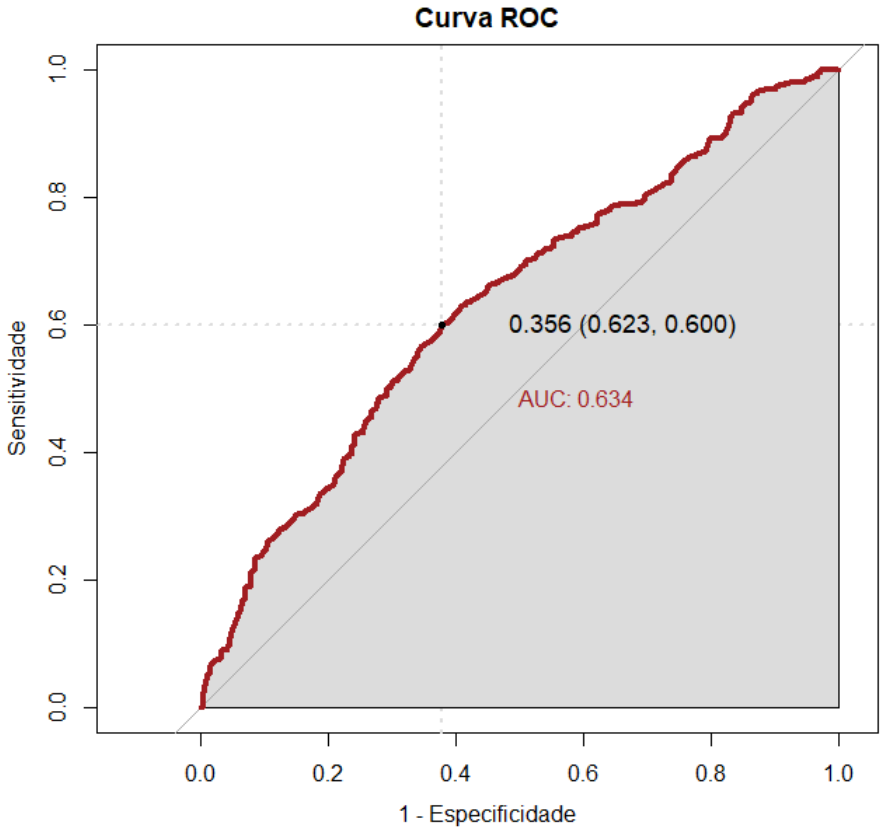

Figura 11: Curva ROC do modelo reduzido

A Tabela 9 mostra que o modelo teve uma acurácia de 61,4%, ou seja, classificou corretamente 61,4% dos estudantes dessa amostra. A Sensibilidade indica que a probabilidade do modelo classificar corretamente um indivíduo na categoria "Não houve mudança"é de 62,3%, e a probabilidade do modelo classificar corretamente um indivíduo que relatou mudanças no consumo é de quase 60%, indicada pelo valor da Especificidade.

Tabela 9: Medidas de poder preditivo

Medidas	Valores
Acurácia	0,614
IC para acurácia (95%)	[0,584 ; 0,643]
Sensibilidade	0,623
Especificidade	0,599

4.4 Interpretação dos parâmetros

Para interpretar os efeitos das variáveis preditoras na resposta, a Tabela 10 com a exponencial dos coeficientes estimados no modelo (razão de chances entre a categoria e sua referência), seus intervalos de 95% de confiança e p-valores dos testes de significância.

Tabela 10: Razões de chances, intervalos de 95% de confiança e seus p-valores para o modelo reduzido

Parâmetro	RC	IC (95%)	P-valor
Orientação Sexual			
Heterossexual	-	-	
Bissexual	2,28	[1,66 ; 3,13]	<0,001
Homossexual	1,15	[0,70 ; 1.85]	0,580
Outros	1,46	[0,92 ; 2,31]	0,107
Residência			
Plano/Lagos	-	-	
Cidade Satélite	0,59	[0,43 ; 0,80]	<0,001
Entorno do DF	0,62	[0,41 ; 0,96]	0,032
Apoio			
Não/Prefere não declarar	-	-	
Apoio Pesquisa	0,44	[0,25 ; 0,73]	0,002
Apoio Social	1,08	[0,76 ; 1,52]	0,662
Idade	0,97	[0,95 ; 0,99]	0,010
Área do Curso			
Exatas	-	-	
Biológicas	1,03	[0,71 ; 1,48]	0,883
Humanas	1,42	[1,02 ; 1,97]	0,037

Esses resultados apontam que, com base nessa amostra, a chance de uma pessoa que se declara bissexual relatar ter havido aumento no consumo de álcool e/ou outras drogas em seu ambiente doméstico é mais que 2 vezes maior do que pessoas que se declaram heterossexuais, mantendo todos os outros preditores constantes. Não existem diferenças significativas nos hábitos de consumo antes e durante a pandemia entre pessoas autodeclaradas heterossexuais, homossexuais e de outras orientações.

Considerando o Local de Residência, as chances de uma pessoa que mora ou em Cidades Satélites ou no Entorno do DF ter presenciado mudanças nos hábitos de consumo durante a pandemia é menor do que a chance de alguém que mora no Plano Piloto/Lagos, mantendo todos os outros preditores constantes. Ambas as chances são cerca de 40% menores. Mantendo todos os outros preditores constantes, a cada aumento

de um ano na idade do estudante, diminui a chance dele ou alguém de sua família ter aumentado o consumo de álcool e/ou outras drogas durante a pandemia em 3%.

Estudantes que recebem algum apoio de pesquisa da universidade tem chance de ter aumentado o consumo de álcool e/ou outras drogas 56% menor do que estudantes que não recebem apoio ou não quiseram declarar, mantendo todos os outros preditores constantes. Não foi observada diferença significativa entre as chances dos alunos que recebem algum apoio social e dos alunos que não recebem.

Não foi observada diferença significativa na chance de presenciar mudanças nos hábitos de consumo entre os estudantes de cursos de Exatas e de Biológicas. No entanto, ao comparar os alunos da área de Exatas com alunos da área de Humanas, notou-se que os alunos de Humanas tem uma chance 42% maior de ter aumentado ou ter tido alguém de sua família aumentado o consumo de álcool e/ou outras drogas durante a pandemia, mantendo todos os outros preditores constantes.

5 Modelagem dos dados: Regressão Logística Multinomial

A categorização inicial da variável resposta deste estudo continha 6 categorias, e duas delas continham poucas observações o que acarretaria em um erro na estimação dos parâmetros. Portanto as categorias foram agrupadas em 4: "Não, ninguém iniciou ou aumentou o consumo", "Sim, eu inicei ou aumentei o consumo", "Sim, outras pessoas iniciaram ou aumentaram o consumo"e "Sim, eu e outras pessoas iniciamos ou aumentamos o consumo". Nesta seção será escolhido o modelo que melhor explique o comportamento desta variável em função das outras variáveis contidas no banco de dados.

5.1 Seleção de variáveis

Para identificar as variáveis que melhor explicam o padrão de consumo de álcool e/ou outras drogas durante a pandemia, inicialmente foram selecionadas as variáveis que apresentaram associações significativas com a variável resposta na análise exploratória, então foram ajustados modelos univariados a fim de testar se os efeitos dessas variáveis eram individualmente significativos. As variáveis que apresentaram efeitos significativos foram: *Idade*, *Orientação Sexual*, *Estado civil*, *Trabalho*, *Residência*, *Área do Curso*, *Apoio* e *Contraiu Covid-19*.

No entanto, nesse modelo as variáveis *Trabalho*, *Área do Curso* e *Contraiu Covid-19* não aparentaram ser significativas, e então foi verificada a possibilidade de reduzir esse modelo. As Tabelas 11 e 12 mostram alguns resultados que indicam que ao retirar as variáveis *Área do Curso* e *Contraiu Covid-19*, os testes de razão de verossimilhança não apontaram diferenças entre o modelo inicial com o modelo reduzido, e nem entre o modelo inicial com o modelo saturado. Os valores de AIC e BIC foram inferiores no modelo reduzido e os coeficientes estimados pelos dois modelos não apresentaram mudanças significativas, portanto, o modelo reduzido será o escolhido.

Tabela 11: Resultados da comparação dos modelos

Modelos	g.l.	χ^2	P-valor
Saturado x Inicial	57	46,455	0,839
Inicial x Reduzido	15	23,092	0,082

Tabela 12: Medidas AIC e BIC

Modelo	AIC	BIC
Inicial	2266.858	2506,61
Reduzido	2259.950	2424,78

A Tabela 13 contém os resultados do modelo reduzido.

Tabela 13: Resultados do modelo reduzido

Coeficiente	Sim, eu	Sim, outras pessoas	Sim, eu e outras pessoas
Intercepto	-0,918*	-0,227	-0,398
Idade	-0,019	-0,067*	-0,052*
Orientação Sexual - Bissexual	1,143*	0,651*	0,634*
Orientação Sexual - Homossexual	0,556	-0,574	0,182
Orientação Sexual - Outros	0,398	0,718*	0,195
Estado civil - Possui parceiro	-0,547	0,833	1,068*
Residência - Cidade Satélite	-0,522*	-0,432	-0,716*
Residência - Entorno do DF	-0,618*	-0,245	-0,608
Trabalho - Sim	0,344	-0,119	0,479*
Apoio - Pesquisa	-1,132*	-1,331*	-0,209
Apoio - Social	0,181	0,377	-0,495

5.2 Qualidade do ajuste e Poder preditivo do modelo

O teste de qualidade de ajuste de Hosmer e Lemeshow para modelos multinomiais desenvolvido por Fagerland, Hosmer e Bofin (2008), não rejeitou a hipótese de que o modelo está bem ajustado aos dados, e nenhum dos resíduos deviance desse modelo apresentou valores elevados que poderiam indicar inadequação do modelo. Além disso, todos os fatores de inflação generalizados calculados foram inferiores a 5, e portanto não existe multicolinearidade neste modelo. A acurácia do modelo foi igual a 63,43%. Ao que tudo indica, o modelo se adequou bem aos dados e explica bem o comportamento da variável *Consumo de álcool e/ou outras drogas*, porém a matriz de confusão e a tabela com medidas de Sensibilidade e Especificidade abaixo revelam que esse não é o caso.

Tabela 14: Matriz de confusão

	Categoria obervada			
Categoria Estimada	Não, ninguém	Sim, eu	Sim, outras pessoas	Sim, eu e outras pessoas
Não, ninguém	685 (62,79%)	167 (15,31%)	104 (9,53%)	113 (10,36%)
Sim, eu	4 (0,40%)	6 (0,55%)	4 (0,40%)	5 (0,46%)
Sim, outras pessoas	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Sim, eu e outras pessoas	0 (0%)	2 (0,18%)	0 (0%)	1 (0,09%)

Tabela 15: Sensibilidade e Especificidade do modelo

Medidas	Não, ninguém	Sim, eu	Sim, outras pessoas	Sim, eu e outras pessoas
Sensibilidade	0,994	0,034	0,000	0,008
Especificidade	0,045	0,986	1.000	0,998

A matriz de confusão contida na Tabela 14 indica o número de observações estimados e observados para cada categoria, onde os percentuais correspondem ao percentual com relação ao total geral. Essa matriz mostra que o modelo estimou a grande maioria das observações (98%) como "Não ninguém iniciou ou aumentou o consumo", e nenhuma como "Sim, outras pessoas iniciaram ou aumentaram o consumo". O motivo da acurácia ter sido alto é que os estudantes que não relataram haver mudanças nos seus hábitos de consumo durante a pandemia compunham 63,15% da amostra em estudo, e grande parte deles foi classificado nessa categoria. Ao observar os valores da Sensibilidade e Especificidade do modelo dispostos na Tabela 15, nota-se que a Sensibilidade para a categoria "Não ninguém"é extremamente alta e sua Especificidade extremamente baixa, e o com-

portamento oposto é observado nas outras categorias. Isso significa que o modelo está basicamente classificando todas as observações como "Não ninguém", e ele acerta muito por conta da grande porcentagem de pessoas dessa categoria no banco de dados. O modelo multinomial não discrimina bem os dados, muito provavelmente pelo tamanho da amostra nas observações fora da categoria "Não, ninguém". Portanto, não será feita a interpretação dos parâmetros.

6 Conclusão

Com base nos resultados obtidos neste estudo, pôde-se observar que o modelo de regressão logística binária se ajustou razoavelmente bem aos dados, com uma acurácia de 61,4%. Tem-se então que, com base nessa amostra, a mudança no padrão de consumo de álcool e/ou outras drogas pode ser explicado pelas variáveis: *Orientação sexual*, onde os estudantes que se declararam bissexuais tem maior chance de ter presenciado aumento no consumo doméstico durante a pandemia; *Residência*, onde estudantes que moram no Plano Piloto/Lagos tem mais chances; *Apoio*, onde os estudantes que recebem algum tipo de apoio pesquisa da universidade exibiram menor chance; *Idade*, onde quanto maior a idade do estudante, menor é a chance; *Área do Curso*, onde estudantes de Humanas apresentam maiores chances quando comparados com estudantes da área de Exatas.

O modelo de regressão logística multinomial a princípio, aparentava ter se ajustado bem aos dados, por conta dos valores, medidas de diagnóstico e acurácia calculados. Porém, notou-se que o modelo estimou 98% das observações como "Não, ninguém iniciou ou aumentou o consumo", e como essa era a categoria mais presente na amostra, a acurácia do modelo foi alta. Portanto, o modelo multinomial não discrimina bem os dados.

Vale ressaltar que embora tenha sido feito um planejamento amostral para o levantamento dos dados, o questionário foi enviado por e-mail aos estudantes e respondê-lo não era obrigatório. Isso pode ter acarretado em um viés em que apenas os alunos que se interessaram pelo tema do questionário o responderam, e que também alguns estudantes podem não ter visto o e-mail. Portanto não é recomendado estender os resultados encontrados neste estudo a todos os estudantes universitários.

Referências

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

ANDRADE, A. G. d.; DUARTE, P. d. C. A. V.; OLIVEIRA, L. G. d. Levantamento nacional sobre o uso de Álcool, tabaco e outras drogas entre universitários das 27 capitais brasileiras. Publicação elaborada pela Secretaria Nacional de Políticas sobre Drogas (Senad) - Observatório Brasileiro de Informações sobre Drogas (Obid) em parceria com o Grupo Interdisciplinar de Estudos de Álcool e Drogas – GREA/IPQ-HC/FMUSP, 2010.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. [S.l.]: São Paulo: Saraiva, 2009.

CONOVER, W. J. *Practical nonparametric statistics*. 3. ed. New York, NY [u.a.]: Wiley, 1999. (Wiley series in probability and statistics). ISBN 0471160687.

DAMASCENO, R. O. et al. Uso de álcool, tabaco e outras drogas e qualidade de vida de estudantes universitários. *Revista baiana de enfermagem, 01 Julho 2016, Vol.30(3)*, 2016.

DINNO, A. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, v. 15, n. 1, p. 292–300, 2015. Disponível em: ⟨https://doi.org/10.1177/1536867X1501500117⟩.

FILGUEIRAS, A.; STULTS-KOLEHMAINEN, M. Factors linked to changes in mental health outcomes among brazilians in quarantine due to covid-19. *MedRxiv (preprint), 2020.*, 2020.

GUNDIM, V. A. et al. Saúde mental de estudantes universitários durante a pandemia de covid-19. *Revista baiana de enfermagem, 01 November 2020, Vol.35*, 2020.

LEMESHOW, S.; STURDIVANT, R.; HOSMER, D. *Applied Logistic Regression*. Wiley, 2013. (Wiley Series in Probability and Statistics). ISBN 9781118548394. Disponível em: ⟨https://books.google.com.br/books?id=wGO5h0Upk9gC⟩.

RUBIN, G. J.; WESSELY, S. The psychological effects of quarantining a city. *BMJ*, BMJ Publishing Group Ltd, v. 368, 2020. Disponível em: ⟨https://www.bmj.com/content/368/bmj.m313⟩.

SILVA, E. C.; HELENO, M. G. V. Qualidade de vida e bem-estar subjetivo de estudantes universitários. *Revista psicologia e saúde. 2012;4(1):69-76*, 2012.

SILVA, S. M. d.; ROSA, A. R. O impacto da covid-19 na saúde mental dos estudantes e o papel das instituições de ensino como fator de promoção e proteção. *Revista Prâksis, 01 May 2021, Vol.2, pp.189-206*, 2021.