Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Bruna Akemi Okabayashi Hirataka

ANÁLISE DO PERFIL DOS CANDIDATOS AO VESTIBULAR DA UEM: UMA ABORDAGEM ESTATÍSTICA UTILIZANDO MODELOS DE REGRESSÃO

Maringá – Paraná
2024

Bruna Akemi Okabayashi Hirataka

ANÁLISE DO PERFIL DOS CANDIDATOS AO VESTIBULAR DA UEM: UMA ABORDAGEM ESTATÍSTICA UTILIZANDO MODELOS DE REGRESSÃO

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof. Dr. Willian Luís de Oliveira
Coorientador: Prof. Dr. Vanderly Janeiro

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – Paraná

2024

Resumo

Este estudo investiga o perfil dos candidatos ao vestibular da Universidade Estadual de Maringá (UEM) por meio de uma abordagem estatística baseada em modelos de regressão. A análise considera dados de processos seletivos realizados entre 2015 e 2024, com ênfase nos anos de 2019 (período pré-pandemia), 2022 (transição e adaptação pós-pandemia) e 2023 (consolidação do período pós-pandemia). O objetivo principal é identificar as variáveis socioeducacionais que influenciam na aprovação dos candidatos, utilizando análise multivariada e regressão logística. Os resultados indicam que fatores como tipo de escola frequentada, renda familiar, escolaridade dos pais e local de residência impactam significativamente a probabilidade de aprovação. Foi observada uma mudança no perfil dos candidatos ao longo do tempo, especialmente nos cursos de Educação Física, Enfermagem e Medicina, que ganharam maior relevância no pós-pandemia. A Análise de Correspondência Múltipla (ACM) permitiu visualizar associações entre variáveis socioeducacionais e desempenho nos vestibulares, enquanto os Modelos Lineares Generalizados (MLG) forneceram insights sobre as chances de aprovação. As conclusões deste estudo destacam a importância de políticas educacionais que promovam maior equidade no acesso ao ensino superior. A identificação de tendências e fatores determinantes pode subsidiar a formulação de estratégias para tornar os processos seletivos mais inclusivos, beneficiando tanto as instituições de ensino quanto a sociedade como um todo.

Palavras-chave: Vestibular, Ensino Superior, Modelos de Regressão, Análise Multivariada, Socioeducacional, UEM.

Abstract

This study investigates the profile of candidates for the entrance exam at the State University of Maringá (UEM) through a statistical approach based on regression models. The analysis considers data from selection processes conducted between 2015 and 2024, with a focus on the years 2019 (pre-pandemic period), 2022 (transition and adaptation post-pandemic), and 2023 (consolidation of the post-pandemic period). The main objective is to identify the socio-educational variables that influence candidate approval, using multivariate analysis and logistic regression. The results indicate that factors such as the type of school attended, family income, parents' education level, and place of residence significantly impact the probability of approval. A change in the candidates' profile over time was observed, particularly in the Physical Education, Nursing, and Medicine courses, which gained greater relevance in the post-pandemic period. Multiple Correspondence Analysis (MCA) allowed for the visualization of associations between socio-educational variables and performance in entrance exams, while Generalized Linear Models (GLM) provided insights into approval odds. The conclusions of this study highlight the importance of educational policies that promote greater equity in access to higher education. Identifying trends and determining factors can support the formulation of strategies to make selection processes more inclusive, benefiting both educational institutions and society as a whole.

Keywords: College Entrance Exam, Higher Education, Regression Models, Multivariate Analysis, Socio-Educational, UEM.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

UEM	Universidade Estadual de Maringá
SETI	Secretaria de Estado da Ciência, Tecnologia e Ensino Superior
THE	Times Higher Education
QS	Quacquarelli Symonds
CWUR	Center for World University Rankings
SISU	Sistema de Seleção Unificada
CVU	Comissão de Vestibular Unificado
PAS	Processo de Avaliação Seriada
CCA	Centro de Ciências Agrárias
CCB	Centro de Ciências Biológicas
CCE	Centro de Ciências Exatas
CCH	Centro de Ciências Humanas
CCS	Centro de Ciências da Saúde
CSA	Centro de Sociais Aplicadas
CTC	Centro de Tecnologia.
PEIES	Programa de Ingresso ao Ensino Superior
UFSM	Universidade Federal de Santa Maria

UFMG	Universidade Federal de Minas Gerais
FECILCAM	Faculdade Estadual de Ciências e Letras de Campo Mourão
MLG	Modelos Lineares Generalizados
ACM	Análise de Correspondência Múltipla
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
PCNs	Parâmetros Curriculares Nacionais
PCESP	Propostas Curriculares da Secretaria de Educação do Estado de São Paulo
IES	Instituições de Ensino Superior
CART	Classification and Regression Trees
Fiocruz	Fundação Oswaldo Cruz
ONU	Organização das Nações Unidas
PCD	Pessoa com Deficiência
CNEC	Campanha Nacional de Escolas da Comunidade
ANOVA	Analysis of Variance (Análise de Variância)
VIF	Variance Inflation Factor (Fator de Inflação da Variância)
ROC	Receiver Operating Characteristic
Ed. Física	Educação Física
IC	Intervalo de Confiança
OR	Odds Ratio

Sumário

Capítulo 1

Introdução

Explorar o perfil dos estudantes que irão ingressar na universidade é de suma importância para adequar políticas educacionais e ajuste de currículos, proporcionando uma experiência acadêmica condizente com o que os alunos necessitam. E mais, promove a inclusão e diversidade, fazendo com que a instituição de ensino possa aplicar estratégias que assegurem uma maior representatividade e suporte a grupos sub-representados. Além disso, melhora o planejamento de infraestrutura e recursos, garantindo o atendimento de necessidades específicas dos alunos.

Até a área de marketing e captação podem ser incluídos, pois compreender o perfil dos estudantes atuais viabiliza à universidade elaborar campanhas mais eficientes, tanto para atrair novos alunos com características semelhantes quanto para diversificar o corpo discente. E por fim, temos também o auxílio na análise de desempenho e no fomento à pesquisa acadêmica, colaborando para a melhoria contínua da instituição e o desenvolvimento de conhecimentos significativos.

A análise da escolha dos cursos e do perfil dos candidatos ao longo do tempo é essencial para compreender os impactos de eventos históricos e sociais na educação. Contudo, a pandemia de COVID-19 trouxe mudanças profundas e significativas nos contextos social, econômico e, também educacional. O distanciamento social, o aumento da demanda por profissionais da saúde e as transformações no ensino, como a adoção do ensino remoto, podem ter influenciado as escolhas dos candidatos no vestibular. Neste contexto, a análise e comparação dos períodos pré e pós-pandemia permite avaliar possíveis alterações nas preferências, isto é, no perfil dos candidatos.

Em particular, destacamos os cursos de Educação Física, Enfermagem e Medicina que possuem uma conexão direta com a saúde e o bem-estar,e são áreas que ganharam ainda

mais visibilidade durante e após a pandemia. O aumento da valorização das profissões relacionadas ao cuidado, ao suporte emocional e à promoção da qualidade de vida pode ter alterado a demanda por esses cursos. Dessa forma, analisar esses cursos permite investigar se houve mudanças na percepção social sobre essas carreiras em função do contexto pandêmico.

A escolha dos anos de 2019, 2022 e 2023 para a análise permite captar diferentes momentos do cenário educacional. O ano de 2019 representa o período pré-pandemia, marcado por estabilidade. Já 2022 e 2023 refletem a fase de adaptação e recuperação pós-pandemia, permitindo identificar possíveis mudanças nas tendências de inscrição e no perfil dos candidatos ao longo do tempo. Essa abordagem comparativa contribui para uma análise mais robusta e contextualizada.

Assim, ao integrar esses três aspectos na investigação, buscamos compreender como fatores externos, como uma pandemia global, podem influenciar as escolhas acadêmicas e profissionais dos estudantes, contribuindo para reflexões sobre o futuro da educação e do mercado de trabalho.

Então podemos dizer que esse estudo de análise de perfil (características dos inscritos e dos aprovados) é de interesse geral das universidades, incluindo a própria UEM, dos alunos e da sociedade como um todo. E neste trabalho, foi realizada uma revisão da literatura, cujo objetivo foi verificar como as universidades lidam com tais informações e como a utilizam, isto é, quais técnicas são utilizadas e quais conclusões podem ser obtidas de tais análises.

Diante da contextualização apresentada, é necessária e essencial a compreensão de como as características socioeducacionais dos candidatos influenciam em seu desempenho nos processos seletivos da Universidade Estadual de Maringá. A análise dessas variáveis pode oferecer subsídios para a formulação de políticas mais equitativas, contribuindo para um acesso mais democrático ao ensino superior. E ao investigar a evolução do perfil dos vestibulandos ao longo do tempo, este estudo busca identificar tendências e possíveis impactos de eventos históricos e sociais na escolha dos cursos e na taxa de aprovação. A seguir, são delineados os objetivos que guiarão esta pesquisa, detalhando as questões centrais e as metodologias empregadas na análise dos dados.

1.1 Objetivos

1.1.1 Geral

Investigar a influência de variáveis socioeducacionais na aprovação dos candidatos aos processos seletivos da Universidade Estadual de Maringá (UEM) em três diferentes momentos: período pré-pandemia (2019), período de transição e adaptação da pandemia (2022) e período da consolidação do pós pandemia (2023), utilizando técnicas estatísticas para análise do perfil dos vestibulandos e modelagem preditiva da aprovação, comparando esses períodos.

1.1.2 Específicos

- Investigar a relação entre variáveis socioeducacionais e a aprovação nos vestibulares nos anos 2019, 2022 e 2023 (período pré, durante e pós pandemia, respectivamente) nos cursos de Educação Física, Enfermagem e Medicina;
- Explorar associações entre variáveis socioeducacionais por meio de Análise Multivariada, nos períodos analisados;
- Modelar a probabilidade de aprovação dos candidatos utilizando técnicas estatísticas, nos períodos e cursos analisados.

Capítulo 2

Revisão de Literatura

Nesta capítulo, realizamos uma ampla revisão da literatura a fim de encontrar trabalhos acerca do perfil dos aprovados em vestibulares, não somente da UEM mas em geral. Além disso, esta revisão buscou encontrar trabalhos que tratasse dos impactos da pandemia no perfil dos inscritos e aprovados.

Dela Libera (2005) apresenta detalhadamente a metodologia desenvolvida para predizer a classificação de candidatos no vestibular da Universidade Federal de Santa Maria (UFSM). O estudo baseia-se em técnicas estatísticas, incluindo o Teorema Central do Limite e distribuições Normal e Binomial, para fornecer uma estimativa de classificação aos vestibulandos com base no número de acertos nas provas. A metodologia mostrou-se eficaz ao prever com precisão a classificação dos candidatos, contribuindo para a tomada de decisão dos mesmos entre o vestibular convencional e o PEIES (Programa de Ingresso ao Ensino Superior). Os resultados obtidos indicam que a aplicação da metodologia oferece uma projeção confiável da classificação, validando sua utilidade como ferramenta de apoio para candidatos e instituições educacionais.

Calil (2007) demonstrou a Análise de Discriminante de Fisher como uma ferramenta estatística eficaz na avaliação de processos seletivos de vestibulares, proporcionando discriminação e classificação robustas dos candidatos com base nos escores individuais nas diferentes disciplinas. A Análise de Discriminante de Fisher foi aplicada com sucesso na Universidade Federal de Santa Maria (UFMS) para o vestibular de 2007, demonstrando alta precisão em cursos com amostras grandes e medianas, porém apresentou limitações em amostras pequenas devido a baixa acurácia dos parâmetros. Essa metodologia envolve a estimação de variáveis dependentes categóricas a partir de variáveis independentes métricas, utilizando métodos de resubstituição e validação cruzada para avaliar a precisão da

classificação. Estudos de caso em cursos de Medicina, Direito e Música na UFMS evidenciaram que a Análise de Discriminante de Fisher possui um alto poder de classificação em cursos com maior número de candidatos por vaga, enquanto a eficácia diminui em cursos com menor número de candidatos. Estes resultados destacam a importância do tamanho da amostra na aplicação da Análise de Discriminante de Fisher, sugerindo que futuras pesquisas devem focar na melhoria das técnicas para lidar com amostras pequenas e na integração de outras técnicas multivariadas para aumentar a precisão da classificação.

Já Lopes (2007), visa conhecer melhor o perfil dos candidatos oriundos de escolas públicas e privadas que tentaram ingressar na UFMG (Universidade Federal de Minas Gerais) em 2004. Utilizando a metodologia de Árvores de Classificação e Regressão (CART), o estudo busca identificar as características definidas no questionário socioeconômico e cultural que podem estar mais associadas com a aprovação no vestibular. Concluiu-se que o local de moradia e o conhecimento de língua estrangeira são as variáveis mais fortemente associadas com a aprovação de candidatos de escolas particulares e públicas, respectivamente. Para validar o modelo, foi utilizada a Curva ROC (Receiver Operating Characteristic), que avalia a capacidade preditiva do modelo, demonstrando que a área sob a curva ROC foi significativamente maior que 0,5, indicando a adequação do modelo. O estudo fornece subsídios para uma discussão mais ampla sobre a democratização do acesso à universidade pública brasileira e a definição de políticas públicas mais adequadas.

O estudo conduzido por Dias (2008) evidencia a complexidade dos fatores que influenciam a aprovação nos vestibulares da Universidade Federal de Minas Gerais (UFMG), chamando atenção para a importância das variáveis socioeconômicas e educacionais. A análise por meio do modelo de Árvores de Classificação e Regressão (CART) revelou que a conclusão do ensino médio em escolas públicas federais ou particulares, o conhecimento de língua estrangeira e um status socioeconômico elevado são os principais determinantes de sucesso no vestibular. Além disso, a distinção entre cursos diurnos e noturnos mostrou que, enquanto a formação educacional tem maior peso nos cursos diurnos, as variáveis socioeconômicas são mais relevantes nos cursos noturnos. Esses achados indicam que a expansão dos cursos noturnos pela UFMG está de fato promovendo uma maior inclusão social, ao atrair candidatos de classes socioeconômicas mais baixas. No entanto, a maior probabilidade de aprovação continua a ser associada a candidatos com melhores condições socioeconômicas e educacionais, sugerindo que, apesar dos avanços, ainda existem barreiras significativas para a equidade no acesso ao ensino superior. Assim, políticas educacionais focadas na melhoria da qualidade das escolas públicas estaduais e municipais, bem como na oferta de cursos preparatórios gratuitos e no fortalecimento de programas de inclusão, como as cotas raciais e sociais, são essenciais para promover uma democratiza-

ção efetiva do acesso ao ensino superior no Brasil. Os resultados deste estudo fornecem uma base sólida para a formulação de tais políticas e destacam a necessidade contínua de pesquisa e intervenção nesse campo.

Em seu estudo, Silva (2009) reconheceram a importância da aplicação das técnicas de Data Mining e da estatística multivariada para estruturar a relação entre o desempenho e as variáveis socioeducacionais dos candidatos ao vestibular de verão 2007 da FECILCAM (Faculdade Estadual de Ciências e Letras de Campo Mourão). A análise permitiu observar comportamentos e padrões, evidenciando o processo de padronização em bancos de dados multivariados. A pesquisa, que possui um enfoque preditivo, utilizou técnicas estatísticas multivariadas para classificar os candidatos ao vestibular, demonstrando a eficácia dessas técnicas na simplificação dos dados e predição de resultados com base em 19 variáveis socioeducacionais. As principais técnicas empregadas incluem Análise de Componentes Principais, Análise de Agrupamentos, Análise Discriminante, Análise Fatorial e Regressão Logística, com especial destaque para as duas últimas.

Garcia (2010) destaca a crescente demanda por ensino superior devido aos avanços tecnológicos e à globalização, o que pressiona as Instituições de Ensino Superior (IES) a realizarem seleções mais justas e eficazes. Garcia utiliza a técnica de Regressão Logística Múltipla para identificar variáveis socioculturais que influenciam na seleção de candidatos aos cursos de Engenharia, abordando aspectos como formação educacional, vida econômica familiar e hábitos dos candidatos. A análise revelou que apenas duas variáveis foram estatisticamente significativas: a quantidade de vezes que o candidato prestou vestibular e a renda familiar mensal. Este estudo, assim como o trabalho de Aversa (2022), busca entender o impacto das variáveis socioeducacionais nos resultados de exames de ingresso, utilizando técnicas estatísticas robustas como a Regressão Logística e a Análise de Árvores de Decisão (CART), proporcionando insights valiosos para a formulação de políticas educacionais que promovam a equidade e a eficácia no processo seletivo.

A análise da equidade no acesso ao ensino superior público no Brasil é o foco de Carvalho (2012), destacando a influência das variáveis socioeconômicas e escolares no desempenho dos candidatos ao vestibular. O estudo, fundamentado na teoria de Bourdieu sobre capital cultural, utiliza dados coletados entre 2005 e 2010 pela Universidade Federal de Sergipe (UFS) para desenvolver um modelo de Regressão Logística Binária, com o intuito de prever a probabilidade de aprovação dos candidatos ao curso de Estatística. Os resultados demonstraram uma boa precisão na discriminação entre candidatos aprovados e reprovados, com uma taxa geral de acertos de 61% e uma área sob a Curva ROC de 0,748, indicando discriminação aceitável.

Lima (2017) investigou a influência das condições socioeconômicas no sucesso no ves-

tibular da Universidade Federal de Goiás (UFG), utilizando dados de 29.226 candidatos do processo seletivo de 2013. A metodologia incluiu análise descritiva e modelos de regressão de Cox para avaliar o impacto de variáveis como escolaridade dos pais, tipo de escola, renda familiar e situação laboral no tempo até a aprovação. A análise revelou que candidatos casados e empregados têm menores chances de aprovação, enquanto a escolaridade dos pais e o tipo de escola influenciam positivamente o desempenho. A utilização de análise de componentes principais (ACP) permitiu criar um indicador de bem-estar material, reforçando a importância das condições socioeconômicas para o sucesso acadêmico. Os resultados sugerem a necessidade de políticas públicas que abordem essas desigualdades para promover maior equidade nos processos seletivos.

A pesquisa de Aversa (2022) investiga o desempenho dos candidatos ao vestibular da UNESP entre 2013 e 2017, com um foco especial nos egressos de escolas públicas. Utilizando uma abordagem metodológica que analisou a concepção educativa do vestibular e a evolução do desempenho dos candidatos, os autores identificaram as principais variáveis socioeconômicas que influenciam o desempenho, como renda, etnia, sexo e procedência escolar. Através da organização e análise dos dados dos vestibulandos, obtidos junto à Vunesp, o estudo revelou que as mudanças implementadas no vestibular em 2010, destinadas a alinhar o exame aos Parâmetros Curriculares Nacionais (PCNs) e às Propostas Curriculares da Secretaria de Educação do Estado de São Paulo (PCESP), não foram eficazes para corrigir as desigualdades de desempenho entre os candidatos de escolas públicas e privadas. A pesquisa destaca que o vestibular da UNESP ainda representa uma barreira significativa para o acesso ao ensino superior público para as camadas populares, reforçando a inacessibilidade ao território da universidade pública.

O estudo de Correa(2022) apresenta os impactos da pandemia da COVID-19 na trajetória acadêmica e no bem-estar psicológico dos estudantes de pós-graduação no Brasil. Essa pesquisa foi realizada com quase 6 mil alunos do ensino superior público e particular das mais diversas áreas do conhecimento através de um questionário online. Revelando que a pandemia causou mudanças significativas nos projetos acadêmicos de muitos estudantes, sendo que 72% tiveram que modificar suas pesquisas, e 35,27% relataram alterações substanciais. A adaptação ao ensino remoto levou os alunos a participarem de atividades como leitura de artigos (82,47%), reuniões online com orientadores (71,66%) e revisões bibliográficas (69,49%). No que diz respeito ao bem-estar, os estudantes relataram altos níveis de estresse, ansiedade e depressão, com 81,95% sentindo-se desmotivados e 78,65% apresentando dificuldades de concentração. Além disso, 61,77% tiveram crises de ansiedade e 61,59% relataram dificuldades para dormir. Apenas 33,35% buscaram ajuda psicológica, e 68,04% dos entrevistados afirmaram não ter recebido apoio emocional dos programas de pós-graduação.

A pandemia do covid-19, causada pelo SARS-CoV-2 ou Novo Coronavírus, não afetou somente a área da saúde e da epidemiologia, mas também houve impactos socias, econômicos, políticos, culturais e históricos em escala mundial sem precedentes. (Fundação Oswaldo Cruz (Fiocruz), 2024)

Lucca (2024) destaca que a análise dos fatores socioeducacionais é fundamental para entender o perfil dos candidatos aprovados nos vestibulares da UEM. E ainda, através de sua análise descritiva, pode-se destacar vários fatores que influenciam na maior probabilidade de aprovação dos candidatos, veja: embora as mulheres se inscrevam em maior número, os homens têm uma taxa de aprovação superior (14%) comparada às mulheres (10%). A probabilidade de aprovação varia entre as etnias, com pessoas amarelas tendo uma probabilidade de 11,67% de serem aprovadas, enquanto pessoas pretas têm a menor probabilidade, de apenas 6,17%. Candidatos com renda familiar superior a 20 salários mínimos têm uma probabilidade de aprovação de 10,14%. Candidatos que estudaram integralmente em escolas particulares apresentam uma probabilidade maior de serem aprovados, especialmente no curso de Medicina, onde a probabilidade é 85% maior comparada aos candidatos de escolas públicas. Candidatos nascidos fora do estado do Paraná e residentes em outras cidades que não Maringá têm menor probabilidade de aprovação. Moradores da zona rural apresentaram uma chance ligeiramente maior de aprovação comparados aos moradores da zona urbana. A escolaridade dos pais, estudantes com pais que possuem ensino superior têm maior chance de aprovação, especialmente no Centro de Tecnologia. Candidatos que estudaram em escolas particulares e frequentaram cursos pré-vestibulares têm maiores chances de aprovação. Estudar no turno matutino também está associado a maiores chances de sucesso nos vestibulares. E ainda, destacando a complexidade dos fatores socioeducacionais que influenciam o sucesso nos vestibulares da UEM e sugerem a necessidade de políticas inclusivas para apoiar candidatos de diversas origens socioeconômicas e educacionais.

Em geral, podemos concluir com os trabalhos acima que enquanto os modelos preditivos desenvolvidos nesses estudos são valiosos para a compreensão e previsão nos resultados de processos seletivos para o ingresso às universidades, temos uma lacuna significativa na análise inferencial, isto é, na análise do impacto das variáveis socioeducacionais. Esses fatores, apesar de serem reconhecidos como influentes, muitas vezes não recebem a devida atenção necessária para informar políticas públicas que poderiam promover uma democratização mais efetiva em relação ao acesso ao ensino superior, podendo abordar sobre os impactos das variáveis socioeducacionais nos candidatos ao longo do tempo, o efeito da combinação de fatores na aprovação dos candidatos no vestibular, a falta de análises qualitativas, o estudo de diferenças entre regiões e instituições e a análise de desempenho acadêmico pós-aprovação.

Além disso, nota-se uma ausência de estudos comparativos dos perfis de candidatos aos vestibulares das universidades em geral, nos períodos pré, durante e pós pandemia. Até mesmo a UEM, possui poucos trabalhos sobre tal tema, um exemplo é Lucca (2024) e com este trabalho queremos aprofundar o estudo sobre impacto das variáveis socioeducacionais utilizando uma abordagem descritiva, logística e multivariada.

Capítulo 3

Materiais e métodos

3.1 Banco de Dados

A Universidade Estadual de Maringá (UEM) é uma instituição pública de ensino superior, financiada pelo Estado do Paraná e está vinculada à Secretaria de Estado da Ciência, Tecnologia e Ensino Superior (SETI). O empenho e a qualificação de seu corpo docente têm sido reconhecidos por diversos rankings mundiais que avaliam a qualidade das universidades com base em critérios acadêmicos e científicos. Em 2024, a UEM figura nas seguintes posições: 1501+ no ranking mundial (THE), 97º na América Latina (QS World University Rankings), 1306º globalmente (CWUR), entre outros. Também é classificada em rankings de impacto social e sustentabilidade, de acordo com os Objetivos de Desenvolvimento Sustentável da ONU. Fundada em 6 de novembro de 1969 pela Lei nº 6.034, a UEM foi oficialmente criada como Fundação pelo Decreto-Lei nº 18.109 em 28 de janeiro de 1970. A Universidade foi formada pela incorporação de instituições já existentes na época, como a Faculdade Estadual de Ciências Econômicas de Maringá, a Faculdade Estadual de Direito de Maringá, e a Faculdade de Filosofia, Ciências e Letras de Maringá.

A Universidade conta com 14.605 alunos matriculados nos cursos de graduação: 13.364 no modo presencial e 1.241 no ensino à distância. Atualmente, a instituição oferece cerca de 91 cursos de graduação (entre turnos, habilitações e câmpus diferentes), 56 cursos de mestrados acadêmicos (1.391 alunos matriculados), 29 doutorados (1.227), 23 cursos de especializações (874), 18 áreas de residências médicas (114) e 26 cursos de pós-doutorado (92 alunos). O ingresso na UEM pode ser feito pelo Sistema de Seleção Unificada (SISU), adotado a partir de 2021, ou pela própria universidade: Vestibulares de Verão e Inverno, e o PAS (Processo de Avaliação Seriada).

A prova dos vestibulares da UEM é composta por uma redação (um gênero textual) e 50 questões objetivas que são do tipo múltipla escolha por somatória, com cinco alternativas numeradas de 01, 02, 04, 08 e 16. A resposta é a soma dos números das alternativas corretas. Já o PAS é o processo no qual o estudante matriculado na 1ª série do ensino médio realiza uma prova ao final de cada uma das séries do ensino médio e sua pontuação acumulada nessas provas poderá classificá-lo a uma vaga na universidade. O PAS/UEM constitui-se de três etapas: Etapa 1: Prova com peso 1, realizada ao final do primeiro ano do Ensino Médio, com conteúdos dessa série; Etapa 2: Prova com peso 2, realizada ao final do segundo ano do Ensino Médio, com conteúdos dessa série, para os alunos classificados na Etapa 1; Etapa 3: Prova com peso 2, realizada ao final do último ano do Ensino Médio, com conteúdos dessa série, para os alunos classificados na Etapa 2.

Em 2023, foram ofertadas 1170 vagas no Vestibular de Inverno de 2023, 1.211 vagas no Vestibular de Verão de 2023, 754 vagas no PAS e 656 vagas para SISU.

A Comissão Central do Vestibular Unificado (CVU) da UEM é o setor responsável pelo processamento, correção e divulgação de datas, e notas e classificação nos processos seletivos.

O banco de dados utilizado neste estudo, é resultado de uma triagem de dados, que compreende informações sobre aproximadamente 274.567 candidatos que participaram dos processos seletivos oferecidos pela Universidade Estadual de Maringá (UEM) com enfoque nos anos de 2019, 2022 e 2023, incluindo eventos como o Vestibular de Verão, Vestibular de Inverno, e o Processo de Avaliação Seriada (PAS).

Este banco de dados é composto por 36 variáveis principais, que incluem dados fornecidos pelos candidatos durante o processo de inscrição: a que centro o curso escolhido pertence, cidade e estado de nascimento, sexo, além de respostas a um questionário socioeducacional com 30 perguntas, focado em coletar informações sobre o perfil socioeconômico e educacional dos candidatos, capturando diversos aspectos relevantes para o estudo da influência de fatores socioeducacionais na aprovação dos candidatos

Adicionalmente, o banco de dados incorpora informações geradas pela CVU, como sua situação no final do processo seletivo (aprovado, reprovado, aprovado por cotas, etc.). As variáveis são codificadas para facilitar a análise, com códigos específicos para eventos, cursos, e centros acadêmicos, entre outros aspectos.

- **nu_ano:** Ano do Evento (2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024).
- **nm_evento:** Vestibular de Inverno, Vestibular de Verão e PAS-UEM.

- **st_final:** Situação final dada em: Aprovado, Aprovado negro, Aprovado PcD, Aprovado sociais, Aprovado sociais negro, Classificado, Não homologado e Reprovado.
- **lt_centro**: Indica o centro acadêmico ao qual o curso escolhido pertence, sendo eles:
  - CCA - Centro de Ciências Agrárias
  - CCB - Centro de Ciências Biológicas
  - CCE - Centro de Ciências Exatas
  - CCH - Centro de Ciências Humanas
  - CCS - Centro de Ciências da Saúde
  - CSA - Centro de Ciências Sociais Aplicadas
  - CTC - Centro de Tecnologia
- **nm_curso**: Indica o curso escolhido pelo candidato.
- **nu_opcao_cotas**: Cotista negro, Cotista negro social, Cotista PcD, Cotista social e Não cotista.
- **q.1 a q.30**: Respostas a 30 perguntas do questionário socioeducacional.

  1. Qual o seu sexo?
     1. Masculino.
     2. Feminino.
  2. Quantos anos você completará até o próximo dia 31 de dezembro?
     1. Menos de 16 anos.
     2. 16 anos.
     3. 17 anos.
     4. 18 anos.
     5. 19 anos.
     6. 20 anos.
     7. 21 anos.
     8. De 22 a 25 anos.
     9. De 26 a 30 anos.
     10. Mais de 30 anos.
  3. Qual a sua cor ou raça? (Fonte: IBGE - Censo 2010)
     1. Branca.

2. Preta.
3. Amarela.
4. Parda.
5. Indígena.

4. Qual o seu estado civil?
    1. Solteiro(a).
    2. Casado(a).
    3. Outro.
5. Você tem alguma deficiência/necessidade educativa especial?
    1. Não.
    2. Deficiência auditiva.
    3. Deficiência física.
    4. Deficiência visual total.
    5. Deficiência visual parcial.
    6. Paralisia cerebral.
    7. Deficiência múltipla.
    8. Outra.
6. Qual o Estado em que você nasceu?
    1. Paraná.
    2. Santa Catarina.
    3. Rio Grande do Sul.
    4. São Paulo.
    5. Mato Grosso.
    6. Mato Grosso do Sul.
    7. Outro.
7. Onde você reside permanentemente?
    1. Maringá.
    2. Outra cidade do Estado do Paraná situada na região noroeste.
    3. Cidade do Estado do Paraná não situada na região noroeste.
    4. Cidade do Estado de Santa Catarina.
    5. Cidade do Estado do Rio Grande do Sul.
    6. Cidade do Estado de São Paulo.
    7. Cidade do Estado do Mato Grosso.

    8. Cidade do Estado do Mato Grosso do Sul.
    9. Cidade situada em Estado não relacionado nos itens anteriores.

8. Qual a localização de sua residência?
    1. Zona urbana.
    2. Zona rural.
9. Quantas pessoas residem com você?
    1. Moro sozinho(a).
    2. Uma pessoa.
    3. Duas pessoas.
    4. Três pessoas.
    5. Quatro pessoas.
    6. Cinco pessoas.
    7. Mais de cinco pessoas.
10. Qual o nível de instrução do seu pai?
    1. Sem escolaridade.
    2. Ensino Fundamental/1º grau incompleto.
    3. Ensino Fundamental/1º grau completo.
    4. Ensino Médio/2º grau incompleto.
    5. Ensino Médio/2º grau completo.
    6. Superior incompleto.
    7. Superior completo.
    8. Pós-Graduação.
    9. Não sei informar.
11. Qual o nível de instrução de sua mãe?
    1. Sem escolaridade.
    2. Ensino Fundamental/1º grau incompleto.
    3. Ensino Fundamental/1º grau completo.
    4. Ensino Médio/2º grau incompleto.
    5. Ensino Médio/2º grau completo.
    6. Superior incompleto.
    7. Superior completo.
    8. Pós-Graduação.
    9. Não sei informar.

12. Qual a renda mensal de sua família?
    1. Até um salário mínimo.
    2. Mais de um salário mínimo e até dois salários mínimos.
    3. Mais de dois salários mínimos e até três salários mínimos.
    4. Mais de três salários mínimos e até cinco salários mínimos.
    5. Mais de cinco salários mínimos e até dez salários mínimos.
    6. Mais de dez salários mínimos e até quinze salários mínimos.
    7. Mais de quinze salários mínimos e até vinte salários mínimos.
    8. Mais de vinte salários mínimos.
13. Qual o item cuja descrição de bens mais se aproxima dos bens da sua família?
    1. Não possui casa própria nem carro ou moto.
    2. Não possui casa própria mas possui carro ou moto.
    3. Possui casa própria e carro ou moto.
    4. Possui casa própria, carro ou moto e outros imóveis urbanos.
    5. Possui casa própria, carro ou moto e caminhão.
    6. Possui casa própria, carro ou moto e propriedade rural.
    7. Possui casa própria, carro ou moto, caminhão e propriedade rural.
    8. Possui casa própria, carro ou moto, caminhão, propriedade rural e outros imóveis.
    9. Possui mais bens além dos relacionados no item anterior.
14. Qual a sua participação na vida econômica da família?
    1. Trabalho, mas recebo ajuda financeira da família ou de outras pessoas.
    2. Trabalho e sou responsável pelo meu próprio sustento.
    3. Trabalho, sou responsável pelo meu próprio sustento e contribuo parcialmente para o sustento da família ou de outras pessoas.
    4. Trabalho e sou principal responsável pelo sustento da família.
    5. Não trabalho e meus gastos são financiados pela família ou por outras pessoas.
15. Durante o curso superior, você terá que trabalhar?
    1. Sim, mas apenas nos últimos anos.
    2. Sim, desde o primeiro ano, em tempo parcial.
    3. Sim, desde o primeiro ano, em tempo integral.
    4. Não sei.

5. Não.

16. Como você realizou seus estudos de Ensino Fundamental (1º grau)?
    1. Integralmente em escola pública.
    2. Integralmente em escola particular.
    3. Maior parte em escola pública.
    4. Maior parte em escola particular.
    5. Em escolas comunitárias/CNEC.
17. Como você realizou ou está realizando o Ensino Médio (2º grau ou equivalente)?
    1. Integralmente em escola pública.
    2. Integralmente em escola particular.
    3. Maior parte em escola pública.
    4. Maior parte em escola particular.
    5. Em escolas comunitárias/CNEC.
18. Quando você concluiu ou concluirá o Ensino Médio (2º grau ou equivalente)?
    1. Há mais de quatro anos.
    2. Há quatro anos.
    3. Há três anos.
    4. Há dois anos.
    5. No ano passado.
    6. Neste ano.
    7. No próximo ano.
19. Em que turno você realizou ou está realizando o Ensino Médio (2º grau ou equivalente)?
    1. Matutino.
    2. Noturno.
    3. Maior parte no diurno.
    4. Maior parte no noturno.
    5. Vespertino.
20. Você frequentou ou frequenta curso pré-vestibular?
    1. Sim, por menos de 1 semestre.
    2. Sim, por 1 semestre.
    3. Sim, por 1 ano.

4. Sim, por mais de 1 ano.
5. Não.

21. Qual o principal motivo que o levou a frequentar curso pré-vestibular?
    1. Meu colégio não prepara adequadamente para o vestibular.
    2. Meu colégio prepara para o vestibular, mas o curso pré-vestibular ensina os "macetes".
    3. Para atualizar meus conhecimentos, porque parei de estudar há muito tempo.
    4. Meu colégio fez convênio com um curso pré-vestibular.
    5. Recebi bolsa no curso pré-vestibular.
    6. Por outro motivo.
    7. Não frequentei.
22. Quantas vezes já prestou Concurso Vestibular?
    1. Uma vez.
    2. Duas vezes.
    3. Três vezes.
    4. Quatro vezes.
    5. Cinco vezes ou mais.
    6. Nenhuma.
23. Você iniciou algum curso superior?
    1. Sim, mas não concluí.
    2. Sim, estou cursando.
    3. Sim, mas já concluí.
    4. Não.
24. Qual o principal motivo que o levou a fazer vestibular na Universidade Estadual de Maringá?
    1. É a única na cidade que oferece o curso pretendido.
    2. É a que oferece o melhor curso pretendido.
    3. É a que oferece curso pretendido em horário adequado.
    4. O curso pretendido é pouco procurado, o que facilita a classificação.
    5. É de fácil acesso (proximidade de casa, prática locomoção, etc.).
    6. Na realidade, gostaria de estudar em outra universidade.
    7. Por ser pública e gratuita, satisfazendo as condições. socioeconômicas da família.

8. Por ser pública, gratuita e de qualidade.

25. Qual o motivo que o levou a escolher o curso para o qual está se candidatando?
    1. Horário mais compatível com outras atividades.
    2. O curso prepara para uma profissão condizente com minhas aptidões.
    3. O curso prepara para uma profissão com perspectiva de boa renda financeira.
    4. O curso prepara para uma profissão com bom mercado de trabalho.
    5. Outro.
26. Como você soube da realização deste vestibular?
    1. Colégio/Cursinho.
    2. Amigos/Parentes.
    3. Jornal.
    4. TV.
    5. Rádio.
    6. Panfleto.
    7. Cartaz.
    8. Outdoor.
    9. Internet.
    10. Aplicativo.
    11. Rede Social.
    12. Website.
    13. Outro.
27. Qual o meio mais utilizado por você para acessar a internet?
    1. Celular.
    2. Tablet.
    3. Notebook.
    4. Computador de mesa.
    5. Não tenho acesso à internet.
28. Qual a forma mais frequente de acesso à internet?
    1. Wifi em casa.
    2. Dados próprios.
    3. Wifi de terceiros (escola, trabalho, centros públicos, comércio).

29. É a sua primeira graduação? (Considerar como não apenas se obteve o título em outro curso superior)
    1. Sim.
    2. Não.
30. Algum de seus pais ou responsáveis tem formação superior completa?
    1. Sim.
    2. Não.

3.2 Análise Multivariada

Segundo (JOHINSON, 2007, p. 1):

> A investigação científica é um processo de aprendizado iterativo. Objetivos relacionados à explicação de um fenômeno social ou físico devem ser especificados e, em seguida, testados por meio da coleta e análise de dados. A análise dos dados coletados por experimentação ou observação frequentemente sugere uma explicação modificada do fenômeno. Ao longo desse processo iterativo, variáveis são frequentemente adicionadas ou removidas do estudo. Consequentemente, as complexidades da maioria dos fenômenos exigem que os investigadores coletem observações sobre muitas variáveis diferentes. (...) Como os dados incluem medições simultâneas em várias variáveis, essa metodologia é chamada de análise multivariada.

Dentro do conjunto de métodos da Estatística Multivariada, temos a Análise de Correspondência Múltipla (ACM) foi adaptada para se tirar conclusões estatísticas a partir de variáveis categóricas, a primeira coisa que deve ser feita é transformar os dados quantitativos em categóricos (podendo usar de quantis estatísticos).foi adaptada para se tirar conclusões estatísticas a partir de variáveis categóricas, a primeira coisa que deve ser feita é transformar os dados quantitativos em categóricos (podendo usar de quantis estatísticos).

A interpretação da ACM geralmente se baseia nas proximidades entre pontos em um mapa de baixa dimensionalidade (normalmente em duas ou três dimensões), sendo as proximidades são relevantes somente entre pontos do mesmo conjunto (isto é, linhas com linhas, colunas com colunas). Mais especificamente, quando dois pontos de linha estão próximos entre si, isso indica que eles tendem a selecionar os mesmos níveis das variáveis nominais. Para analisar a proximidade entre variáveis, é necessário distinguir dois cenários. No primeiro, a proximidade entre níveis de variáveis nominais diferentes indica que esses

níveis frequentemente aparecem juntos nas observações. No segundo, dado que os níveis de uma mesma variável nominal não podem coexistir, a proximidade entre esses níveis deve ser interpretada de outra forma: ela indica que os grupos de observações associados a esses níveis são semelhantes entre si.

Quando o conjunto de dados está completamente representado como variáveis categóricas, é possível construir a chamada tabela disjuntiva completa. Denotamos essa tabela como X. Se I pessoas responderam a uma pesquisa com J questões de múltipla escolha com 4 respostas cada, X terá I linhas e $4J$ colunas.

E mais teoricamente, (ABDI; VALENTIN, 2007) assume-se que X é a tabela de dados completa de I observações de K variáveis categóricas. E também que a k-ésima variável tem J_k níveis (categorias) diferentes e define-se $J = \sum_{k=1}^{K} J_k$. Logo X é então uma matriz $I \times J$ com todos os coeficientes sendo 0 ou 1. Define-se a soma de todas as entradas de X como N e introduz-se $Z = XN^{-1}$. Na MCA, também existem dois vetores especiais: primeiro r, que contém as somas ao longo das linhas de Z, e c, que contém as somas ao longo das colunas de Z. Denota-se $D_r = \text{diag}(r)$ e $D_c = \text{diag}(c)$, as matrizes diagonais contendo r e c, respectivamente, como diagonal. Com essas notações, calcular uma MCA consiste essencialmente na decomposição em valores singulares da matriz:

$$M = D_r^{-1/2}(Z - rc^T)D_c^{-1/2}$$

A decomposição de M dá P, Δ e Q, tal que $M = P\Delta Q^T$, com P e Q sendo duas matrizes unitárias, e Δ é a matriz diagonal generalizada dos valores singulares (com o mesmo formato de Z). Os coeficientes positivos de Δ^2 são os autovalores de Z.

O interesse da MCA vem da maneira como as observações (linhas) e variáveis (colunas) em Z podem ser decompostas. Essa decomposição é chamada de decomposição fatorial. As coordenadas das observações no espaço fatorial são dadas por

$$F = D_r^{-1/2}P\Delta$$

As i-ésimas linhas de F representam a i-ésima observação no espaço fatorial. De maneira similar, as coordenadas das variáveis (no mesmo espaço fatorial das observações) são dadas por

$$G = D_c^{-1/2}Q\Delta$$

Portanto, os métodos que envolvem a MCA são:

- Codificação dos dados categóricos para uma matriz binária;

- Matriz de contingência expandida que representa uma tabela de frequência cruzadas entre todas as variáveis categóricas;
- Análise da inércia que decompõe a inércia total (que indica a variabilidade dos dados) em componentes correspondentes as dimensões criadas pelo método;
- Decomposição de valores singulares é aplicada á matriz de dados para ser decomposta em componentes ortogonais utilizados para mapeamento de categorias em baixa dimensão;
- Extração das dimensões principais é a fase na qual cada dimensão explica a inércia total;
- Plotagem dos resultados resulta em um gráfico bidimensional, no qual as categorias das variáveis são distribuídas de acordo com sua associação forte ou fraca, no quesito proximidade;
- Interpretação dos eixos é baseada nas variáveis e categorias que mais contribuem para as dimensões (eixos gráficos), logo as categorias que estão em uma distância maior da origem contribuem mais para a inércia do eixo correspondente.

3.3 Modelos de Regressão Logística

Os Modelos Lineares Generalizados (MLG) são uma extensão dos modelos lineares clássicos, permitindo que a variável resposta siga distribuições além da normal, e são compostos por três componentes:

- **Componente aleatória**: Está relacionada à distribuição da variável resposta. Assume-se que cada componente da variável aleatória tem distribuição na família exponencial, tomando a forma:

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

  para algumas funções específicas $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$.

- **Componente sistemática**: É constituída pelas variáveis explanatórias no formato de uma estrutura linear.
- **Função de ligação**: Realiza a conexão entre as componentes aleatória e sistemática.

Para mais detalhes, ver Cordeiro (2008) e Hosmer(2000).

A família exponencial de distribuições é um conjunto de várias distribuições conhecidas, como normal, binomial, binomial negativa, gama, Poisson, normal inversa, multinomial, beta, logarítmica, dentre outras.

Métodos de regressão têm se tornado um componente integral de qualquer análise de dados envolvendo a descrição da relação entre a variável resposta e uma ou mais variáveis explicativas. Frequentemente, esses métodos são aplicados em casos onde a variável de desfecho é discreta, considerando duas ou mais variáveis. Na última década, o modelo de regressão logística tem se tornado, em várias áreas, o método padrão de análise nessa situação. O que distingue o modelo de regressão logística do modelo de regressão linear é que a variável resposta na regressão logística é binária ou dicotômica (HOSMER; LEMESHOW; STURDIVANT, 2000).

Esse tipo de modelo estatístico é comumente utilizado para classificação e análise preditiva, pois resulta em uma probabilidade, com a variável dependente limitada entre 0 ou 1. Na regressão logística, uma transformação logit é aplicada à chance, isto é, à probabilidade de sucesso dividida pela probabilidade de fracasso. Isso também é comumente conhecido como chance logarítmica, ou logaritmo natural da chance. Logo, o modelo de regressão logística fica definido pelo uso da ligação logito em um MLG binomial.

A função de ligação logito baseia-se na função de distribuição acumulada (fda) da distribuição logística em sua forma padrão ($\mu = 0$ e $\sigma = 1$):

$$F(z) = \frac{e^z}{1 + e^z}$$

O modelo de regressão generalizado na função de ligação logito fica definido por:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}$$

Ou, na escala do preditor:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

onde:

- π_i: probabilidade de ocorrência do evento de interesse para a i-ésima observação.
- β_0: intercepto do modelo.
- $\beta_1, \beta_2, \ldots, \beta_p$: coeficientes de regressão.

- $x_{i1}, x_{i2}, \ldots, x_{ip}$: variáveis explicativas (preditoras) para a i-ésima observação.

A regressão logística é normalmente aplicada a problemas de previsão e classificação, incluindo detecção de fraudes, revisão de doenças e previsão de rotatividade.

Selecionar um conjunto de covariáveis para compor um modelo parcimonioso é uma tarefa desafiadora, principalmente devido às dificuldades combinatórias e estatísticas. O problema combinatório surge da necessidade de testar todas as possíveis combinações de covariáveis para decidir quais devem compor o modelo. Já o problema estatístico é determinar, a cada adição de um novo termo ao preditor linear, o equilíbrio entre a redução da discrepância entre o valor esperado e o valor da variável resposta, e o aumento da complexidade do modelo.

De acordo com Cordeiro(2008), outras estatísticas que servem como medidas de comparação de qualidade de ajuste do modelo e seu grau de complexidade são os critérios de informação de Akaike (AIC) e de Bayes (BIC) que para os MLG podem ser expressos como:

$$\text{AIC}_p = S_p + 2p - 2\hat{\ell}_n$$

e

$$\text{BIC}_p = S_p + p \log n - 2\hat{\ell}_n$$

com

- S_p: Soma dos erros quadrados ou alguma medida de ajuste do modelo, dependendo do contexto (pode representar a soma dos resíduos ou o erro de ajuste).
- p: Número de parâmetros no modelo.
- $\log n$: Logaritmo natural do número de observações n.
- $\hat{\ell}_n$: Verossimilhança estimada (log-verossimilhança) do modelo ajustado aos dados n.

Modelos com valores baixos para AIC são considerados como representativos de um melhor ajuste, e os modelos são selecionados visando a obter um mínimo AIC. De forma semelhante, interpreta-se o BIC.

A análise de resíduos e o diagnóstico de MLG são etapas cruciais para avaliar a adequação do modelo, verificar o cumprimento de suas suposições e identificar potenciais problemas, como outliers, heterocedasticidade e falta de ajuste. As técnicas utilizadas para a

análise de resíduos e diagnóstico em MLG são semelhantes às empregadas na regressão linear clássica, com algumas adaptações. Os resíduos são calculados como a diferença entre os valores observados e os valores ajustados pelo modelo. Entre os principais tipos de resíduos, destacam-se os resíduos de Pearson, resíduos de deviance e resíduos estudantilizados.

Neste trabalho, iremos abordar sobre os *resíduos quantílicos randomizados*, de acordo com Dunn e Smyth (1996), baseiam-se no método da transformação integral da probabilidade, foram propostos como uma alternativa para avaliar a adequação de modelos estatísticos, especialmente em casos onde os resíduos tradicionais não seguem a normalidade. Essa técnica baseia-se na inversão da função de distribuição acumulada (CDF) do modelo ajustado, transformando os resíduos em uma distribuição aproximadamente normal.

Se a função de distribuição acumulada $F(y;\mu,\phi)$ for contínua, os resíduos quantílicos são definidos como:

$$r_{q,i} = \Phi^{-1}(F(y_i;\hat{\mu}_i,\hat{\phi})) \tag{3.3.1}$$

onde Φ^{-1} é a função quantílica da distribuição normal padrão.

Para variáveis discretas, os resíduos quantílicos randomizados introduzem aleatoriedade ao selecionar um valor uniforme no intervalo $(a_i, b_i]$, onde:

$$a_i = \lim_{y\to y_i^-} F(y;\hat{\mu}_i,\hat{\phi}) \quad \text{e} \quad b_i = F(y_i;\hat{\mu}_i,\hat{\phi}) \tag{3.3.2}$$

Isso evita que os resíduos fiquem concentrados em valores discretos e sigam padrões visíveis nos gráficos de resíduos.

Dobson(2002) e Hosmer, Lemeshow e Sturdivant (2013) relatam que, frequentemente, é mais fácil interpretar os efeitos dos fatores explicativos em termos de razões de chances do que dos parâmetros β. Simplificando, considere uma variável resposta com J categorias e uma variável explicativa binária x, que indica as chances de o resultado estar presente entre os indivíduos com $x = 1$ é $\pi(1)/[1-\pi(1)]$. Da mesma forma, chances de o resultado estar presente entre os indivíduos com $x = 0$ é $\pi(0)/[1-\pi(0)]$. A razão de chances, denotada por OR (Odds Ratio), é a razão entre os odds para $x = 1$ e os odds para $x = 0$, e é dada pela equação:

$$OR = \frac{\pi(1)}{[1-\pi(1)]} \cdot \frac{[1-\pi(0)]}{\pi(0)} = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} \cdot \frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}.$$

Logo, a odds ratio (OR) é uma medida de associação entre variáveis categóricas (por exemplo, exposição e desfecho), muito utilizada na regressão logística. Assim, temos:

- OR = 1: Não há associação entre a exposição e o desfecho; as odds (chances) do desfecho são iguais nos dois grupos.
- OR > 1: A exposição está associada a um aumento nas odds do desfecho.
- OR < 1: A exposição está associada a uma diminuição nas odds do desfecho.

Capítulo 4

Resultados

4.1 Análise Descritiva

Para realizar a análise descritiva utilizou-se o *software* R (RStudio Team, 2023), com especial atenção ao número de inscritos nos diferentes processos seletivos da universidade, como os Vestibulares de Inverno e Verão e o Processo de Avaliação Seriada (PAS). Embora o foco principal desse estudo sejam os anos de 2019, 2022 e 2023, uma análise preliminar foi realizada considerando o período de 2015 a 2024. Essa etapa inicial permitiu identificar possíveis tendências e mudanças ao longo dos anos nos diversos centros acadêmicos, oferecendo uma base sólida para comparações mais detalhadas e uma interpretação mais precisa das flutuações no número de inscrições nos anos de maior interesse.

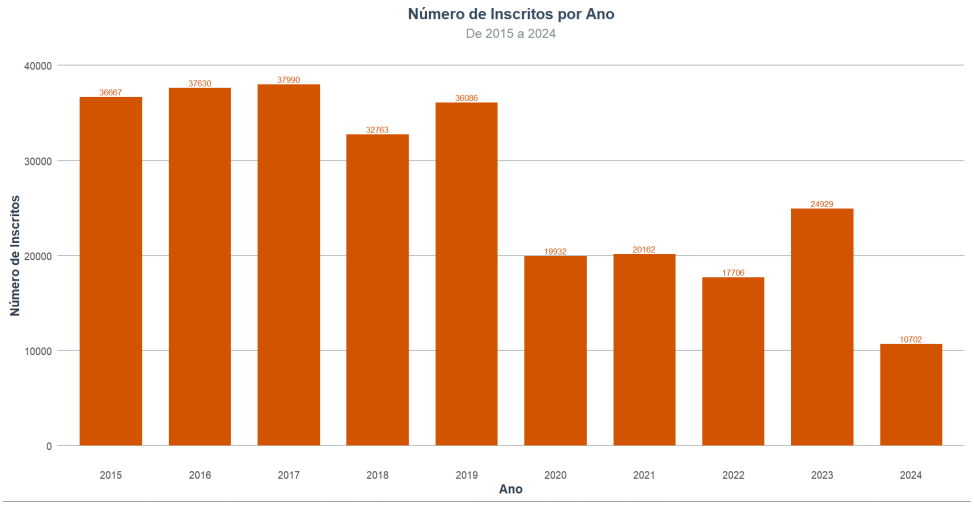


Figura 4.1.1 – Número de inscritos nos processos seletivos da UEM por ano

Com a Figura 4.1.1, pode-se verificar que entre 2015 e 2019, o número de inscritos manteve-se relativamente estável, variando entre 32.763 (2018) e 37.990 (2017). A partir de 2020, houve uma queda drástica de inscritos (representando 44,8% em relação ao ano anterior) devido a pandemia de COVID -19 e também a universidade realizou somente um vestibular por ano, o número de inscritos voltou a crescer no ano de 2023. Lembrando que temos um baixo índice de inscritos em 2024, pois no banco de dados temos somente os inscritos do Vestibular de Inverno 2024.

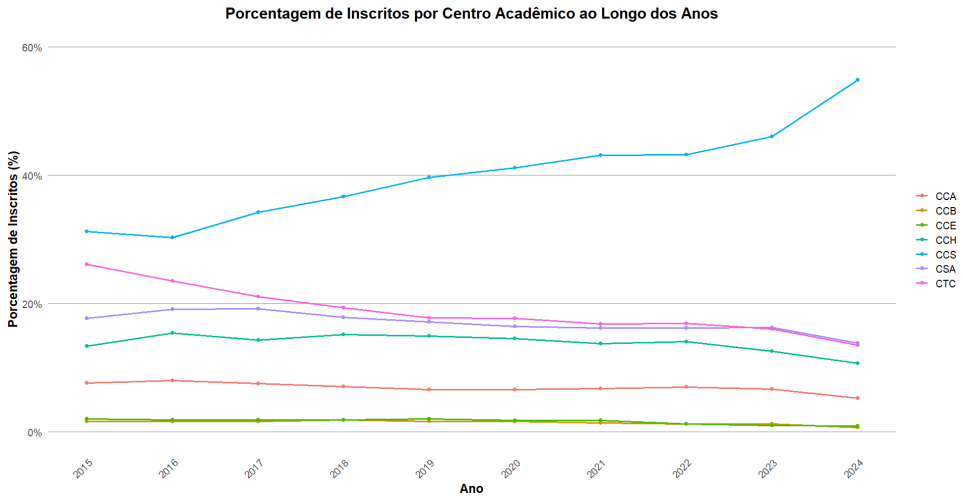


Figura 4.1.2 – Distribuição do número total de inscritos por Centro Acadêmico ao longo dos anos.

Podemos observar que na Figura 4.1.2, o Centro de Ciências da Saúde tem demonstrado o maior interesse dos candidatos ao longo dos anos, sendo que, neste ano de 2024, 56% dos candidatos optaram por cursos referentes a esse centro. Os Centro de Ciências Agrárias, Centro de Ciências Biológicas e Centro de Ciências Exatas se mantiveram constantes ao longo dos 10 anos e obtivemos uma queda de interesse dos inscritos nos Centro de Tecnologia e Centro de Ciências Sociais Aplicadas em relação último ano.

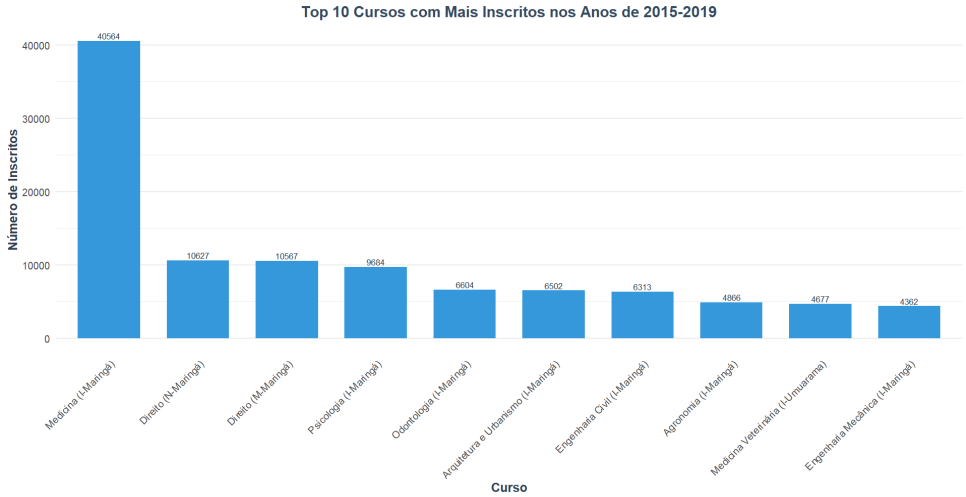


Figura 4.1.3 – Cursos com maior número de inscritos (2015-2019).

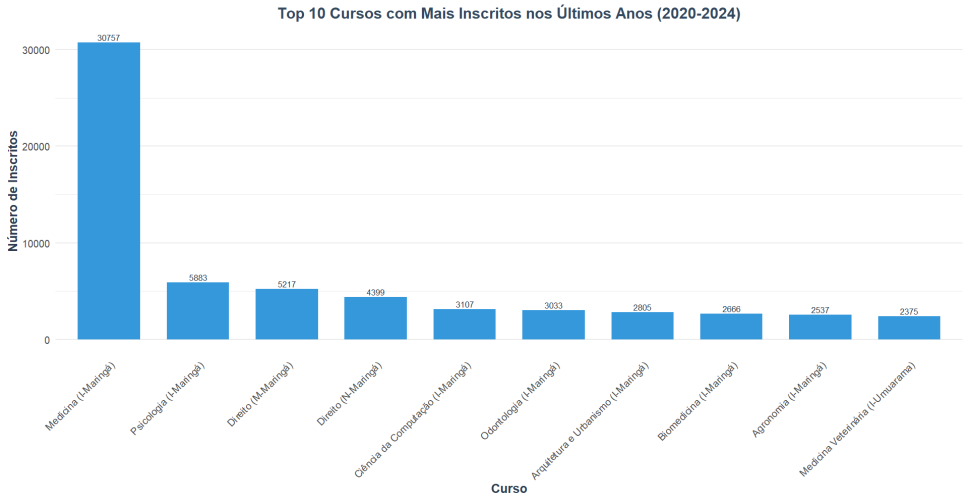


Figura 4.1.4 – Cursos com maior número de inscritos (2020-2024).

É possível observar, através das Figuras 4.1.3 e 4.1.4, que Medicina(I-Maringá) permanece sendo o curso com mais inscritos ao longos desses dez anos. No entanto, depois da pandemia, nota-se que a Engenharia Civil perdeu seu lugar para Ciência da Computação no ranking dos 10 cursos mais inscritos.

Através da análise descritiva, ao observar quais categorias são mais frequentes entre as variáveis pertencentes ao questionário socioeducacional, pode-se estabelecer um perfil para os candidatos aos processos seletivos da UEM ao longo dos anos, em particular, para os anos de 2019, 2022 e 2023, foco do estudo.

Tabela 1 – Distribuição percentual das categorias por variável.

Variável	Descrição	1	2	3	4	5	6	7	8	9	10	11	12	13
q.1	Sexo	39.95%	**60.05%**											
q.2	Idade	11.83%	6.11%	**32.36%**	22.73%	9.56%	5.14%	3.15%	5.23%	2.03%	1.84%			
q.3	Raça	**77.03%**	2.59%	4.49%	15.67%	0.21%								
q.4	Estado Civil	**96.83%**	1.56%	1.61%										
q.5	Deficiência	**88.19%**	6.15%	2.59%	0.04%	0.66%	1.54%	0.08%	0.74%					
q.6	Naturalidade	**77.32%**	1.99%	0.52%	12.77%	1.08%	1.83%	10.34%						
q.7	Residência	26.71%	**29.72%**	18.11%	7.29%	3.16%	10.34%	0.85%	1.49%	2.32%				
q.8	Localização	**93.75%**	6.25%											
q.9	Composição Familiar	2.19%	9.25%	22.27%	**34.35%**	20.59%	5.96%	5.41%						
q.10	Escolaridade do Pai	1.66%	13.58%	7.34%	10.00%	**26.31%**	6.51%	19.23%	11.83%	3.56%				
q.11	Escolaridade da Mãe	1.40%	9.54%	11.72%	7.35%	**21.74%**	6.16%	20.31%	20.45%	1.31%				
q.12	Renda Familiar	9.73%	18.97%	16.97%	**22.95%**	19.54%	7.30%	2.16%	2.37%					
q.13	Bens Possuídos	11.85%	20.77%	**45.81%**	12.30%	1.74%	4.47%	0.98%	1.48%	0.59%				
q.14	Atividade Econômica	24.10%	4.00%	3.99%	0.79%	**67.12%**								
q.15	Trabalho Durante Graduação	8.44%	22.38%	7.72%	**41.44%**	20.02%								
q.16	Ensino Fundamental	**48.24%**	37.03%	7.16%	7.49%	0.08%								
q.17	Ensino Médio	**45.70%**	45.27%	3.99%	4.99%	0.06%								
q.18	Conclusão do Ensino Médio	18.51%	3.22%	4.76%	9.11%	17.16%	**41.45%**	5.78%						
q.19	Turno do Ensino Médio	**80.48%**	4.73%	10.52%	1.87%	2.40%								
q.20	Curso Pré-Vestibular	**50.69%**	4.56%	5.43%	7.35%	32.00%								
q.21	Meio de Acesso à Internet	12.77%	8.31%	6.37%	5.66%	**34.19%**	5.72%	27.00%						
q.22	Forma de Acesso à Internet	25.53%	13.85%	9.36%	4.10%	12.27%	**34.88%**							
q.23	Primeira Graduação	21.77%	14.16%	9.25%	**54.82%**									
q.24	Escolaridade dos Pais	4.52%	9.65%	2.26%	**54.82%**	2.60%	0.57%	8.43%	24.85%					
q.25	Motivo para Escolher o Curso	2.44%	**67.62%**	6.04%	11.95%	11.94%								
q.26	Fonte de Informação sobre o Vestibular	22.44%	**46.97%**	2.48%	8.35%	5.62%	0.10%	0.18%			0.09%			1.75%
q.27	Meio de Acesso à Internet	**75.94%**	15.20%	4.02%	3.29%	1.54%								
q.28	Forma de Acesso à Internet	**77.63%**	16.29%	4.80%	1.28%									
q.29	Primeira Graduação	**92.74%**	7.26%											
q.30	Formação Superior dos Pais	**73.57%**	26.43%											

Logo o perfil apresentado pelos inscritos é: sexo feminino (60,05%); estão na faixa etária dos 17 anos (32,36%); raça branca (77,03%); solteiros (96,83%); sem deficiência/necessidade educativa(88,19 %); nascidos no Paraná(77,32%); residentes permanentes em outra cidade do Paraná situada na região noroeste (29,72%); residentes de zona urbana (93,75 %); moram com três pessoas (34,35%); os pais e mães possuem ensino médio/2ª grau completo (26,31% e 21,74%, respectivamente); renda mensal de 3 até 5 salários mínimos (22,95%); possuem casa própria e casa ou moto(45,81 %); não possui participação ativa na vida econômica, sendo os gastos financiados pela família ou outras pessoas (67,12%); não sabem se precisarão trabalhar durante o curso superior(41,44%); realizou seus estudos integralmente em escola pública tanto no Ensino Fundamental quanto Médio (48,24% e 45,7%, respectivamente); concluirá o ensino médio no ano da inscrição ao processo seletivo (41,45%); a maioria realiza o Ensino Médio no turno matutino (80,48%); frequentou curso pré-vestibular por pelo menos um semestre (50,69%); recebeu bolsa no curso pré-vestibular (34,19%); é a primeira vez que presta Concurso Vestibular (34,88%); não iniciou nenhum curso superior (54,82%); o principal motivo que o levou a fazer vestibular na Universidade Estadual de Maringá é a pouca procura do curso pretendido o que facilita a classificação (47,1%); o motivo que levou a escolher o curso para o qual está se candidatando é o preparo para uma profissão condizente com suas aptidões (67,62%); soube da realização deste vestibular através de amigos e\ou parentes (46,97%); o meio mais utilizado para acessar a internet é o celular (75,94%); a forma mais frequente de acesso à internet é wi-fi em casa (77,63%); é a primeira graduação do candidato (92,74%); e um dos pais ou responsáveis tem formação superior completa (73,57%).

Do mesmo modo, pós realizar uma triagem de dados, reduzimos o banco de dados as categorias de aprovação (Aprovado, Aprovado negro, Aprovado PcD, Aprovado sociais, Aprovado sociais negro) para observar quais categorias são mais frequentes entre as variáveis no banco de dados “Aprovação” e através dessa frequência estabelecer um perfil para os candidatos aprovados nos processos seletivos de 2015 a 2024.

O perfil dos aprovados é composto pelo: sexo feminino (55,05%); 17 anos (31,14%); cor/raça branca (77,93%); solteiro(a) (96,51%); não possui deficiência (86,09%); nascidos no Paraná (82,15%); residentes permanentes de Maringá (33,75%), residentes da zona urbana (93,65%); moram com três pessoas (20,46%); possuem pais com Ensino Médio completo (26,33%); renda familiar de mais de 1 e até 2 salários mínimos (20,44%); possui casa própria e carro ou moto (38,74%); não trabalha e é sustentado pela família (59,50%); não sabe se precisará trabalhar durante o curso (37,88%); cursou o Ensino Fundamental e Ensino Médio Integralmente em escola pública (42,19% e 27,78%, respectivamente); concluirá o Ensino Médio no ano do processo seletivo (40,90%); realizou o Ensino Médio no turno matutino (83,22%); frequentou o Curso Pré-Vestibular por menos de 1 semes-

tre (71,56%); é a primeira vez realizando o vestibular (70,30%); não iniciou nenhum curso superior anteriormente (56,68%); o principal motivo que o levou a fazer vestibular na Universidade Estadual de Maringá é a pouca procura do curso pretendido o que facilita a classificação (62,80%); o motivo que levou a escolher o curso para o qual está se candidatando é o preparo para uma profissão condizente com suas aptidões (68,66%); soube da realização deste vestibular através de amigos e/ou parentes (86,12%); o meio mais utilizado para acessar a internet é o celular (68,66%); a forma mais frequente de acesso à internet é Wifi em casa (77,63%); é a primeira graduação do candidato (92,74%); e um dos pais ou responsáveis tem formação superior completa (73,57%).

Foi realizado também o cálculo da probabilidade de aprovação utilizando tabelas de frequência, destacando a maior probabilidade de aprovação.

O resumo das probabilidades de aprovação para as variáveis analisadas evidencia as categorias com maior probabilidade de aprovação. Em relação ao sexo (q.1), os homens apresentam uma probabilidade de 11,78%, enquanto as mulheres possuem 9,60%, apesar do número de inscrito do grupo feminino ser maior. Para a variável idade (q.2), a maior probabilidade está entre os candidatos com menos de 16 anos (14,90%), seguidos pelos com mais de 30 anos (11,81%). Quanto à cor ou raça (q.3), candidatos que se identificam como pretos apresentam a maior probabilidade de aprovação (18,53%), seguidos pelos brancos (10,07%). No estado civil (q.4), casados têm maior probabilidade (11,42%), seguidos por outros estados civis (11,12%). Na variável sobre deficiência (q.5), candidatos com deficiência auditiva apresentam 14,10% de aprovação. Em relação ao estado de nascimento (q.6), candidatos do Paraná possuem 11,27%. Quanto à residência permanente (q.7), aqueles que residem em cidades do Rio Grande do Sul têm uma probabilidade de 13,64%, enquanto candidatos que moram em zona rural (q.8) possuem 11,08%. Na questão sobre pessoas em casa (q.9), famílias com mais de cinco pessoas apresentam 13,46% de aprovação. No que diz respeito à instrução do pai (q.10) e da mãe (q.11), a maior probabilidade de aprovação ocorre entre os candidatos cujos pais possuem Ensino Médio incompleto, com 12,16% e 13,10%, respectivamente. Para a renda familiar (q.12), famílias que recebem até dois salários mínimos apresentam 12,97%. Quanto à descrição de bens (q.13), candidatos que não possuem casa própria, mas possuem carro ou moto, têm 13,17%. Na participação econômica (q.14), aqueles que trabalham e recebem ajuda financeira apresentam 13,47%, enquanto o trabalho integral durante o curso (q.15) possui a maior probabilidade de aprovação entre as variáveis analisadas (16,31%). A escolaridade em escola pública no Ensino Fundamental (q.16) e no Ensino Médio (q.17) apresenta probabilidades de aprovação de 12,11% e 11,27%, respectivamente. Candidatos que concluíram o Ensino Médio no ano seguinte (q.18) possuem uma probabilidade de 14,86%, e aqueles que estudaram no turno noturno (q.19) apresentam 11,48%. Frequentar um curso pré-vestibular

por mais de um ano (q.20) eleva a probabilidade para 12,99%, enquanto não frequentar (q.21) apresenta 13,06%. Candidatos que nunca realizaram vestibulares anteriores (q.22) possuem 11,18%, enquanto os que não iniciaram um curso superior (q.23) apresentam 10,39%. Sobre o motivo de prestar vestibular na UEM (q.24), aqueles que desejam estudar em outra universidade têm 15,08%. Em relação ao motivo da escolha do curso (q.25), compatibilidade de horário tem a maior probabilidade, com 16,27%. Quanto ao meio de divulgação do vestibular (q.26), amigos e parentes representam 11,16%. Candidatos que acessam a internet por meio de computador de mesa (q.27) possuem 12,82%, enquanto aqueles que utilizam wifi em casa (q.28) apresentam 11,29%. Sobre a variável primeira graduação (q.29), responder “sim” está associado a uma probabilidade de 10,53%. Por fim, candidatos cujos pais não possuem formação superior (q.30) apresentam uma probabilidade de aprovação de 14,38%.

Para fins de análises mais profundas, destacou-se o ano de 2019, ano antes da pandemia, realizando a análise de perfil dos candidatos não-cotistas aprovados e não-aprovados neste ano e concluiu-se que ambos os grupos possuem predominância feminina, com maioria de candidatos de 17 anos de idade. A maior parte se identifica como branca e é solteira. No que diz respeito a necessidades especiais, a maioria dos candidatos de ambos os grupos não possui deficiência. Quanto ao estado de nascimento, observa-se que a maior parte é natural do Paraná. Os candidatos aprovados e reprovados também residem, majoritariamente, em Maringá, em áreas urbanas, e têm renda familiar situada entre 3 e 5 salários mínimos, além de possuírem, em sua maioria, casa própria e veículo (carro ou moto). A maior parte dos candidatos não frequentou curso pré-vestibular, estava realizando sua primeira tentativa de vestibular e ainda não havia iniciado um curso superior. O principal motivo para a escolha da UEM em ambos os grupos foi o fato de ser uma instituição pública, gratuita e de qualidade, enquanto a escolha do curso foi associada à compatibilidade com as aptidões pessoais. Além disso, a principal fonte de informação sobre o vestibular, tanto para aprovados quanto para reprovados, foi o colégio ou cursinho, e o acesso à internet predominante em ambos os grupos foi via Wi-Fi em casa.

Apesar dessas semelhanças citadas, há diferenças importantes entre os perfis de aprovados e reprovados de 2019. Os candidatos aprovados vivem, predominantemente, com três pessoas no domicílio, já os reprovados convivem com quatro pessoas. A escolaridade dos pais também difere: entre os aprovados, o pai possui ensino médio completo e a mãe, pós-graduação, enquanto entre os reprovados ambos os pais apresentam ensino médio completo. Em relação à origem escolar, os aprovados frequentaram escolas particulares no ensino fundamental e médio, ao passo que os reprovados estudaram em escolas públicas nessas etapas. Além disso, foram encontradas diferenças na participação econômica e no trabalho durante o curso: os aprovados não trabalham e estão indecisos quanto a

trabalhar durante o curso, enquanto os reprovados dependem financeiramente da família, mas muitos expressam intenção de trabalhar enquanto estudam.

Essas distinções evidenciam desigualdades que podem influenciar as chances de aprovação e apontam para a importância de considerar fatores socioeconômicos e educacionais em análises mais aprofundadas sobre o desempenho dos candidatos.

A análise do perfil dos candidatos não-cotistas em 2022 revela que tanto os aprovados quanto os reprovados apresentam predominância feminina, sendo a maioria dos aprovados com 17 anos e dos reprovados com 18 anos. A maior parte dos candidatos de ambos os grupos se identifica como branca, é solteira, não possui deficiência e nasceu no Paraná. No que diz respeito à localização, os candidatos residem majoritariamente em Maringá e em áreas urbanas.

A renda familiar dos candidatos aprovados se concentra entre um e dois salários mínimos, enquanto os reprovados têm uma renda média mais alta, entre três e cinco salários mínimos. A maioria possui casa própria e veículo (carro ou moto). Os candidatos de ambos os grupos, em sua maioria, não frequentaram curso pré-vestibular, estavam realizando sua primeira tentativa de vestibular e ainda não haviam iniciado um curso superior. O principal motivo para a escolha da Universidade Estadual de Maringá (UEM) foi o fato de ser uma instituição pública, gratuita e de qualidade. A escolha do curso, para ambos os grupos, esteve relacionada à compatibilidade com as aptidões pessoais. A principal fonte de informação sobre o vestibular foi o colégio ou cursinho, e o acesso à internet predominante era via Wi-Fi em casa.

Apesar das semelhanças, há diferenças importantes nos perfis dos candidatos aprovados e reprovados em 2022. Os candidatos aprovados vivem, predominantemente, em domicílios com quatro pessoas, enquanto os reprovados convivem com até cinco pessoas. A escolaridade dos pais também se diferencia: entre os aprovados, o pai possui ensino médio completo e a mãe possui escolaridade variável, enquanto entre os reprovados, ambos os pais frequentemente possuem apenas o ensino médio completo.

Em relação à origem escolar, os candidatos aprovados frequentaram integralmente escolas públicas tanto no ensino fundamental quanto no médio, enquanto os reprovados apresentaram maior diversidade, muitas vezes com passagem por escolas particulares. Quanto à participação econômica, os aprovados geralmente não trabalham e dependem financeiramente da família, o que lhes proporciona maior dedicação aos estudos. Já os reprovados, além de dependerem da família, demonstram uma maior necessidade de trabalhar durante o curso, o que pode dificultar o foco exclusivo nos estudos.

Em 2023, a análise do perfil dos candidatos não-cotistas aprovados e não-aprovados

revela uma predominância de candidatas do sexo feminino, com a maior parte tendo menos de 16 anos entre os aprovados e 18 anos entre os reprovados. A maioria dos candidatos de ambos os grupos se identifica como branca e solteira, e não possui deficiência. A maior parte é natural do Paraná e reside, majoritariamente, em Maringá, em áreas urbanas. Em relação à renda familiar, os candidatos aprovados possuem, predominantemente, renda entre um e dois salários mínimos, enquanto os reprovados apresentam maior dispersão, com renda mais alta em média. Ambos os grupos possuem, em sua maioria, casa própria e veículo (carro ou moto). A maioria dos candidatos não frequentou curso pré-vestibular, estava realizando sua primeira tentativa de vestibular e ainda não havia iniciado um curso superior. O principal motivo para a escolha da UEM em ambos os grupos foi o fato de ser uma instituição pública, gratuita e de qualidade, e a escolha do curso foi associada à compatibilidade com as aptidões pessoais. Tanto os aprovados quanto os reprovados tiveram o colégio ou cursinho como principal fonte de informação sobre o vestibular e acessavam a internet majoritariamente via Wi-Fi em casa.

Apesar das semelhanças, há diferenças importantes entre os perfis de aprovados e reprovados em 2023. Os aprovados tendem a viver em domicílios com mais de cinco pessoas, enquanto os reprovados moram, majoritariamente, com três pessoas. No quesito escolaridade dos pais, os aprovados possuem pais com ensino médio completo, enquanto entre os reprovados a mãe tem, frequentemente, pós-graduação. Em relação à origem escolar, os aprovados frequentaram integralmente escolas públicas no ensino fundamental e médio, enquanto os reprovados tiveram uma trajetória mais diversificada, muitas vezes envolvendo escolas particulares. No aspecto econômico e profissional, os aprovados não trabalham e dependem financeiramente da família, o que lhes permite maior foco nos estudos. Já os reprovados, além de dependerem financeiramente da família, demonstram maior necessidade de trabalhar durante o curso, indicando uma possível dificuldade em conciliar trabalho e estudos. Essas diferenças sugerem que a estabilidade socioeconômica e o foco exclusivo nos estudos podem ser fatores determinantes para a aprovação no vestibular.

Entre 2019 (pré-pandemia) e os anos de 2022 e 2023 (pós-pandemia), houve mudanças significativas no perfil dos candidatos não-cotistas aprovados. Em 2019, predominavam candidatos vindos de escolas particulares, enquanto em 2022 e 2023 a maioria dos aprovados cursou o ensino fundamental e médio integralmente em escolas públicas. Essa mudança sugere uma maior democratização do acesso ao ensino superior no cenário pós-pandemia. Quanto ao trabalho durante o curso, em 2022 muitos candidatos aprovados indicaram a necessidade de trabalhar, refletindo possíveis dificuldades econômicas herdadas da pandemia. Já em 2023, a maioria dos aprovados não trabalhava e dependia financeiramente da família, o que permitiu maior foco nos estudos. Essa transição de 2022 para 2023 pode indicar uma recuperação econômica e um ambiente mais estável, reforçando a

importância da dedicação exclusiva aos estudos para o sucesso no vestibular.

4.2 Análise Multivariada

Antes de propor os modelos estatísticos, realizou-se uma Análise de Correspondência Múltipla, para observar quais variáveis independentes são similares entre si, com o objetivo de diminuir o número de variáveis do modelo, visando obter um modelo mais parcimonioso. Resolvemos realizar uma ACM para os anos 2019, 2022 e 2023.

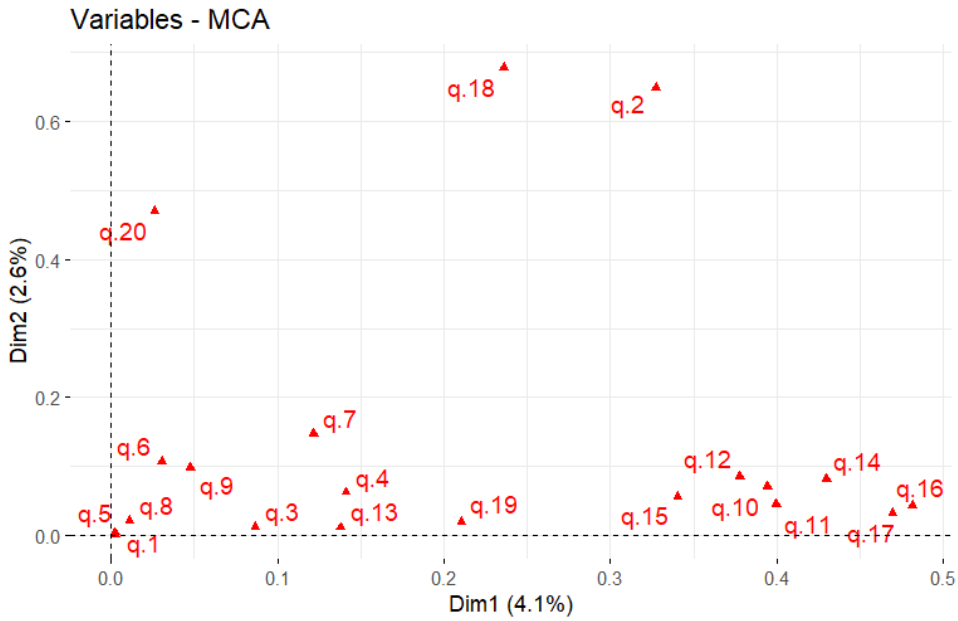


Figura 4.2.1 – Gráfico de Análise de Correspondência Múltipla (MCA) para 2019

Na Figura 4.2.1, observa-se que a Dim1 explica 4,1% da variabilidade e a Dim2 2,6%, indicando que a maior parte da variabilidade dos dados não é capturada nessas duas primeiras dimensões. As variáveis q.1, q.3, q.4, q.5, q.6, q.7, q.8, q.9, q.13, q.19 e q.20 estão próximas da origem, sugerindo que estas têm baixa influência na diferenciação das dimensões, indicando comportamento homogêneo ou baixa capacidade discriminatória. As variáveis q.2 e q.18 estão distantes da origem, sugerindo um maior poder discriminatório, possivelmente associadas a perfis específicos ou a respostas diferenciadas no questionário. No quadrante inferior direito, variáveis como q.10, q.11, q.12, q.14, q.15, q.16 e q.17 estão agrupadas, indicando que possuem correlação ou características em comum, o que pode sugerir um perfil de candidato mais homogêneo em relação a essas questões.

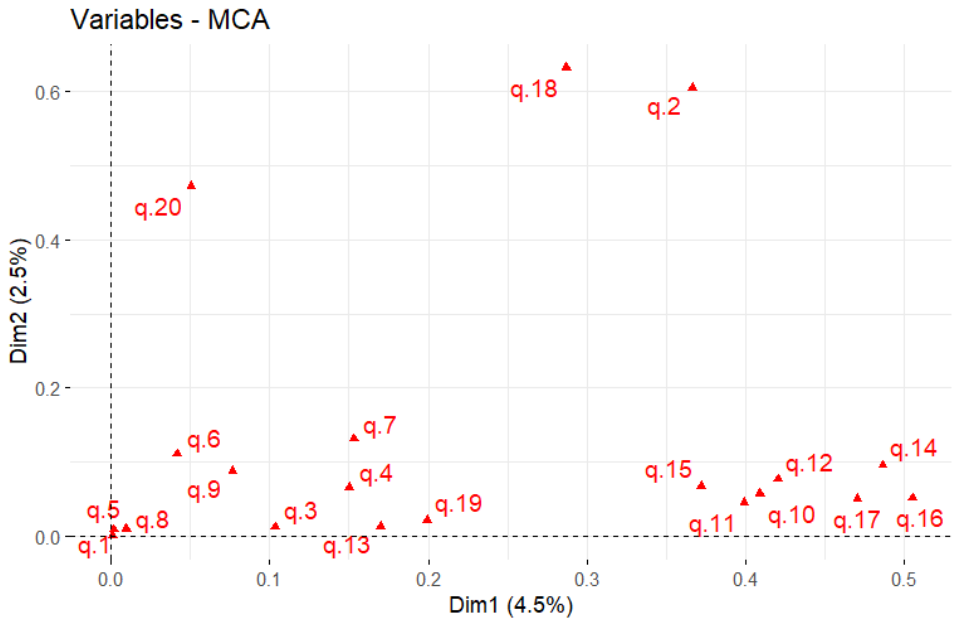


Figura 4.2.2 – Gráfico de Análise de Correspondência Múltipla (MCA) para 2022

Pra o ano de 2022, a contribuição das dimensões aumentou ligeiramente, com a Dim1 explicando 4,5% e a Dim2 2,5% da variabilidade. A distribuição das variáveis mantém o padrão observado em 2019, reforçando a estabilidade das relações entre as variáveis ao longo do tempo. As variáveis agrupadas no quadrante inferior direito (q.10, q.11, q.12, q.14, q.15, q.16, q.17) continuam mostrando correlação ou comportamento semelhante, sugerindo consistência nos perfis de candidatos ao longo do tempo.

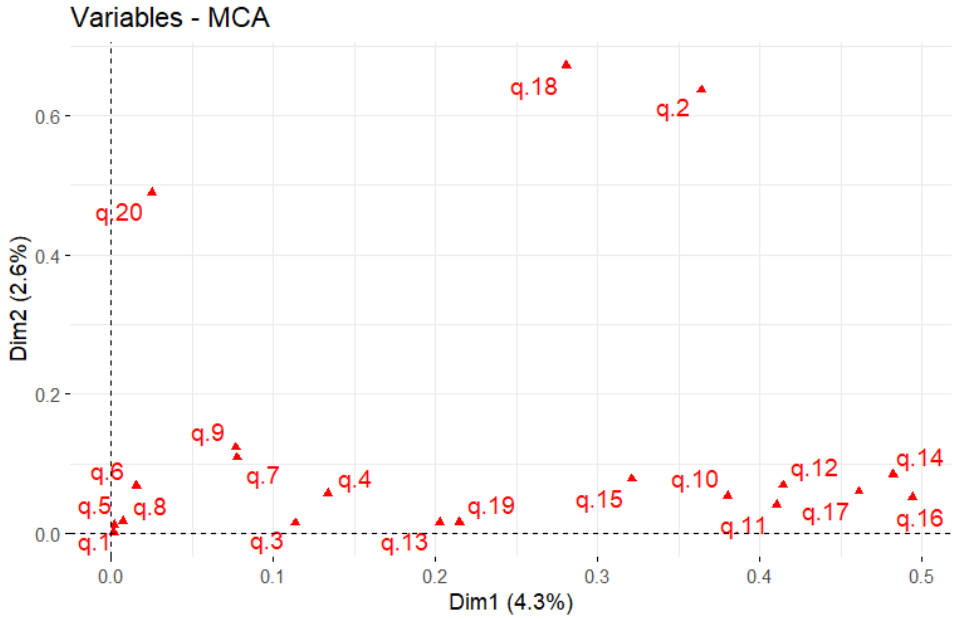


Figura 4.2.3 – Gráfico de Análise de Correspondência Múltipla (MCA) para 2023

Já em 2023, temos também uma leve alteração na contribuição das dimensões, com a Dim1 explicando 4,3% e a Dim2 2,6% da variabilidade. As variáveis q.2 e q.18 mantêm sua posição destacada, indicando que continuam sendo importantes para diferenciar os perfis. As variáveis q.10, q.11, q.12, q.14, q.15, q.16 e q.17 permanecem agrupadas, reforçando a correlação entre si e sugerindo que podem estar relacionadas a um conjunto específico de características dos candidatos.

Em resumo, para os anos de 2019, 2022 e 2023, temos uma consistência na distribuição das variáveis categóricas ao longo das duas primeiras dimensões. Com destaque as variáveis q.2 e q.18 como principais diferenciais, contribuindo significativamente para a separação dos perfis dos candidatos. Em contraste, variáveis como q.1, q.3, q.4, q.5, q.6, q.7, q.8, q.9, q.13, q.19 e q.20 permaneceram próximas à origem, indicando baixa influência na diferenciação das dimensões. E o agrupamento consistente das variáveis q.10, q.11, q.12, q.14, q.15, q.16 e q.17 reforça a estabilidade nos perfis socioeducacionais dos candidatos ao longo dos anos analisados.

Como um dos objetivos dessa análise é explorar associações entre variáveis socioeducacionais nos cursos de Ed. Física, Enfermagem e Medicina nos anos 2019, 2022 e 2023, aplicamos novamente a análise de correspondência múltipla separado por curso e período.

4.2.1 Educação Física

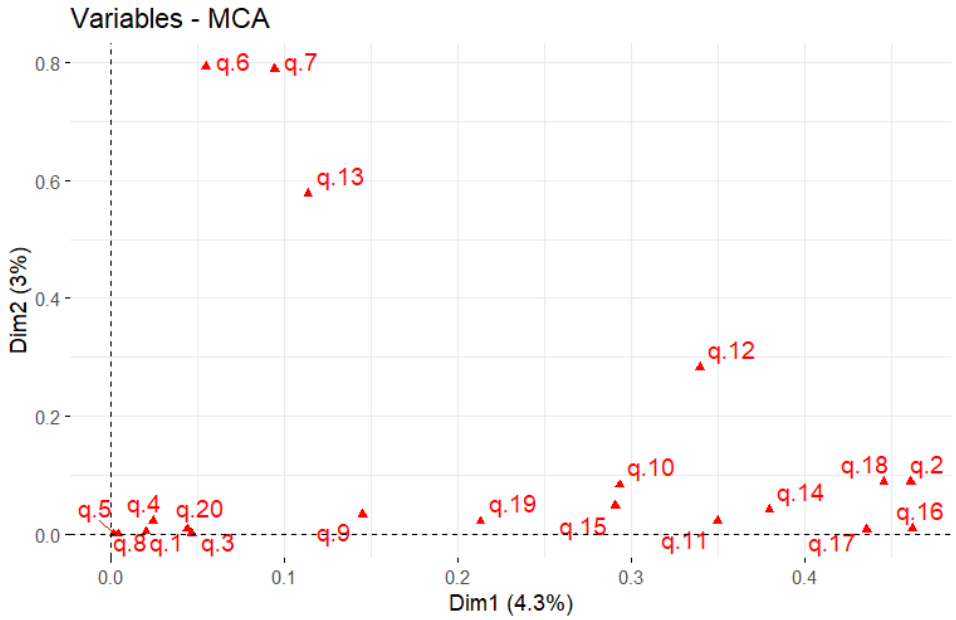


Figura 4.2.4 – Gráfico de Análise de Correspondência Múltipla (MCA) para Ed.Física em 2019

No gráfico 4.2.4, os eixos representam as dimensões extraídas pela MCA. A primeira dimensão (Dim1) explica 4,3% da variabilidade total dos dados, enquanto a segunda dimensão (Dim2) explica 3%. Esses valores indicam que as duas primeiras dimensões capturam uma parte relativamente pequena da variabilidade presente nos dados, o que sugere que as associações entre as variáveis são dispersas e não concentradas em poucos fatores principais.

Cada ponto no gráfico representa uma variável categórica incluída na análise. A proximidade entre variáveis indica que elas compartilham perfis similares e possuem associações mais fortes dentro do espaço definido pelas dimensões. Por outro lado, variáveis mais afastadas apresentam padrões distintos de resposta no conjunto de dados analisado. No gráfico, observa-se que as variáveis q.6, q.7 e q.13 possuem uma forte influência na Dim2, o que sugere que essas variáveis desempenham um papel importante na variação explicada por essa dimensão. Já as variáveis q.16, q.18 e q.2 estão mais associadas à Dim1, indicando que essa dimensão captura diferenças relacionadas a esses fatores.

Além disso, algumas variáveis, como q.1, q.3, q.4, q.5, q.8, q.9 e q.20, estão posicionadas próximas ao centro do gráfico. Isso sugere que sua contribuição para a variabilidade total é menor, ou seja, elas possuem um efeito menos pronunciado na estrutura latente dos dados. Em outras palavras, essas variáveis não apresentam grande diferenciação nos perfis analisados.

A interpretação desse gráfico pode ser contextualizada dependendo da natureza das variáveis analisadas.Além disso, a proximidade entre algumas variáveis pode indicar que elas tendem a se manifestar juntas em um mesmo grupo de indivíduos, como acontece com: q.1 e q.4, q.5 e q.8 e, q.3 e q.20.

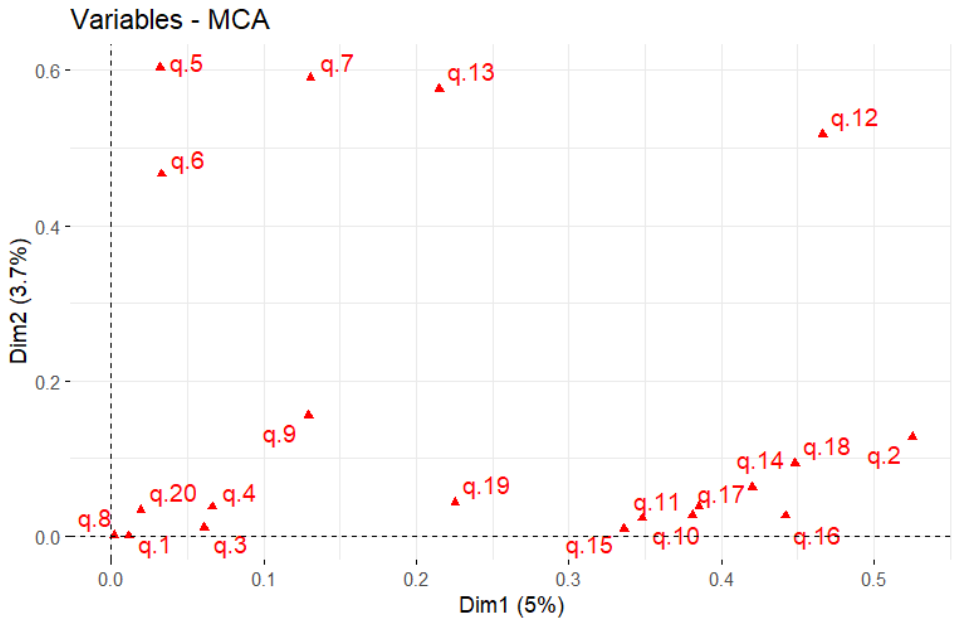


Figura 4.2.5 – Gráfico de Análise de Correspondência Múltipla (MCA) para Ed.Física em 2022

Neste gráfico, a estrutura da distribuição das variáveis é semelhante à de 2019, mas com pequenas variações nos eixos.Dim1 explica 5% da variabilidade e Dim2 explica 3,7%, o que indica leve alteração na importância dos componentes em relação a 2019. A variável q.12 tem um posicionamento diferenciado e distante, indicando uma característica forte nesse eixo, e q.1 e q.8, continuam com pouca influência nos componentes principais.

A estrutura geral do MCA de 2022 mantém semelhanças com 2019, com algumas variações nas posições das variáveis. Consideramos que devido sua proximidade o par de variáveis q.1 e q.8, q.3 e q.4, e q.10 e q.17 estão associadas.

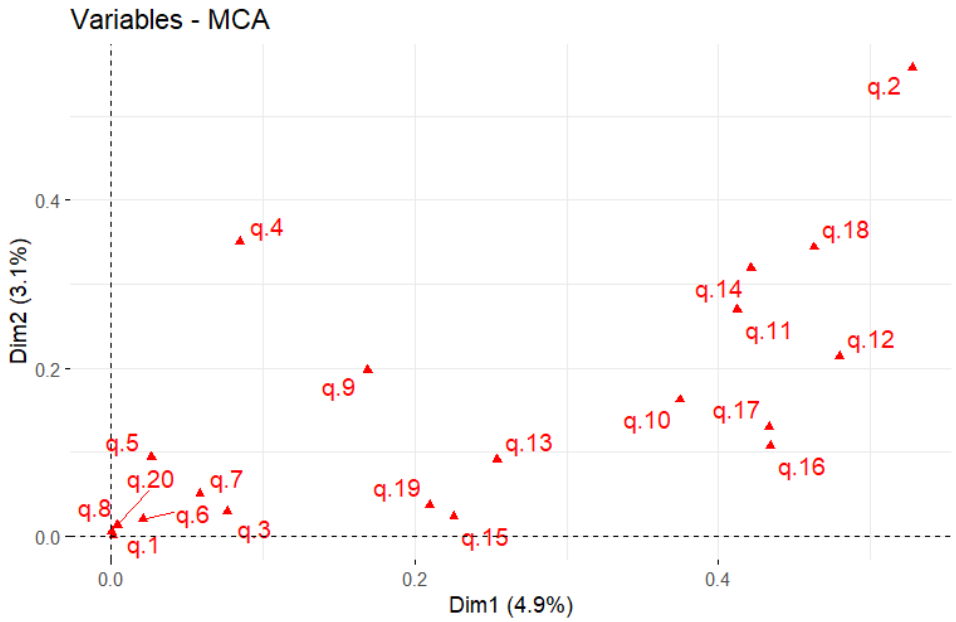


Figura 4.2.6 – Gráfico de Análise de Correspondência Múltipla (MCA) para Ed.Física em 2023

Em 2023, nota-se uma estrutura um pouco mais dispersa em comparação com os anos anteriores, com Dim1 explicando 4,9% e Dim2 3,1% da variabilidade. Algumas variáveis que estavam mais agrupadas nos anos anteriores estão um pouco mais distribuídas ao longo do eixo Dim1. E variáveis similares são: q.1 e q.8, q.1 e q.20, q.8 e q.20 e, q.16 e q.17.

4.2.2 Enfermagem

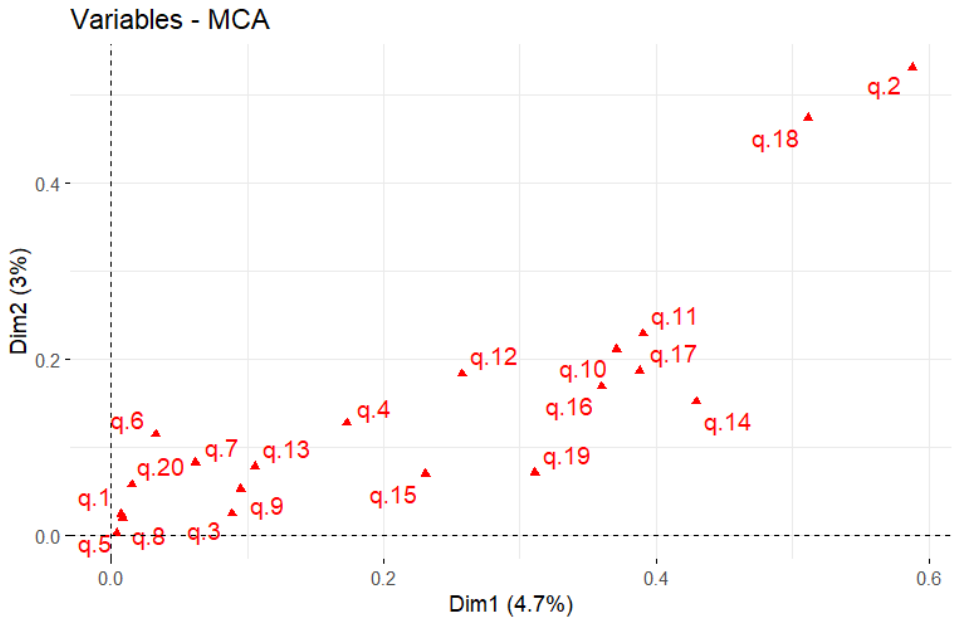


Figura 4.2.7 – Gráfico de Análise de Correspondência Múltipla (MCA) para Enfermagem em 2019

Através da Figura 4.2.7 acima, temos que Dim1 explica 5,6% da variabilidade e Dim2 explica 4,1%, indicando que a dispersão das variáveis nesses eixos captura parte da estrutura dos dados. As variáveis que apresentam indícios de associação pela sua proximidade são q.1 e q.8.

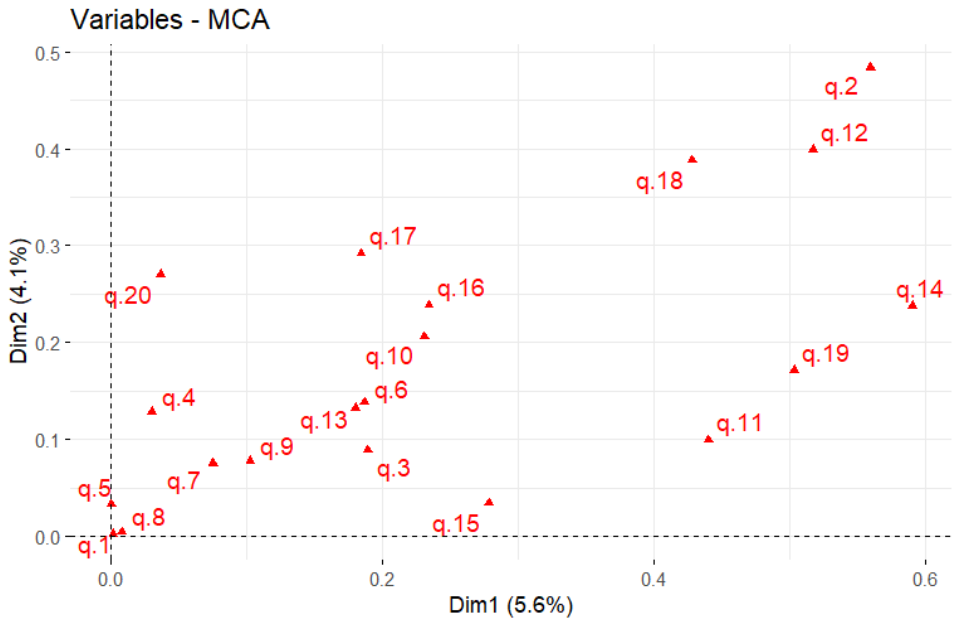


Figura 4.2.8 – Gráfico de Análise de Correspondência Múltipla (MCA) para Enfermagem em 2022

Em 2022, temos que Dim1 explica 5,7% da variabilidade e Dim2 explica 4,1%, valores próximos aos de 2019. As varáveis associadas são: q.1 e q.8, q.6 e q.13.

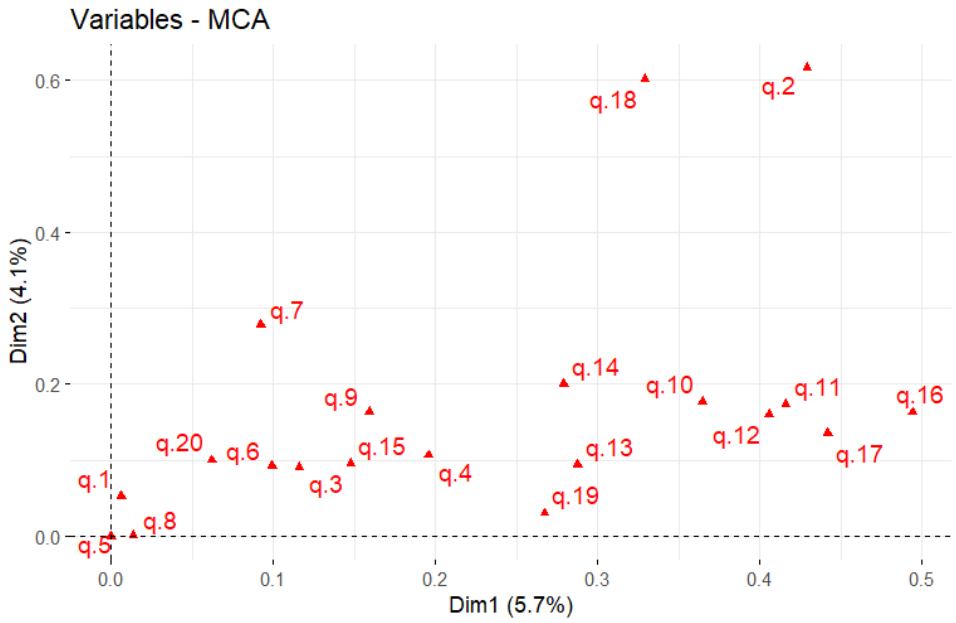


Figura 4.2.9 – Gráfico de Análise de Correspondência Múltipla (MCA) para Enfermagem em 2023

Já o gráfico de 2023 (Figura 4.2.9) apresenta uma estrutura ligeiramente diferente, com Dim1 explicando 4,7% e Dim2 explicando 3% da variabilidade, um pouco menor em relação

aos anos anteriores. As variáveis parecem mais alinhadas em um padrão diagonal, indicando uma estrutura de respostas mais ordenada. Consideramos as variáveis associadas entre si: q.5 e q.8 e q11 e q.12.

Pequenas variações na posição das variáveis indicam que houve mudanças sutis ao longo do tempo, mas as associações gerais se mantêm coerentes.

4.2.3 Medicina

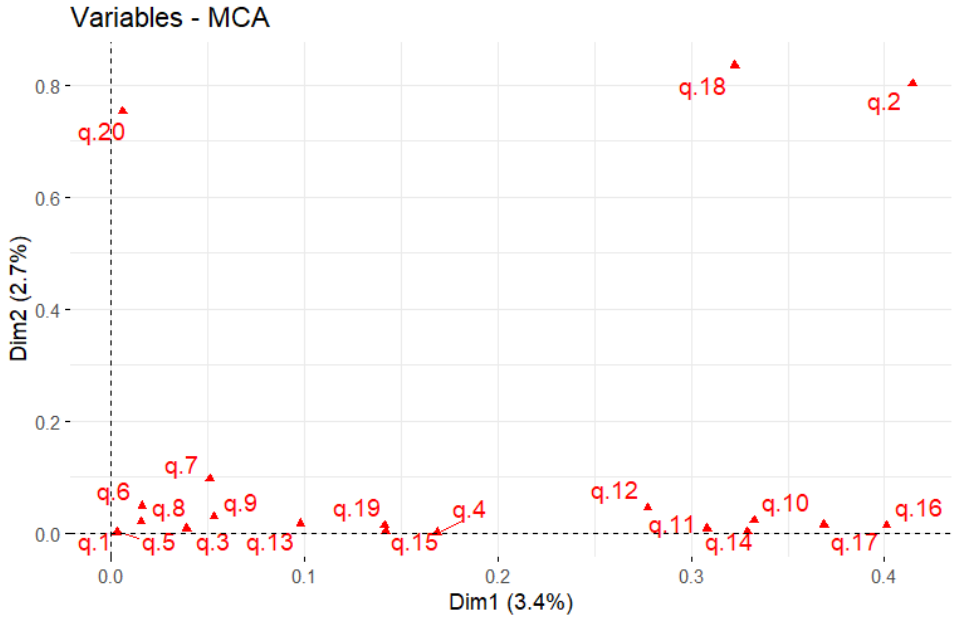


Figura 4.2.10 – Gráfico de Análise de Correspondência Múltipla (MCA) para Medicina em 2019

O gráfico da Figura 4.2.10, captura a variabilidade dos dados em 4,1% e 2,6% para as dimensões Dim1 e Dim2, respectivamente. Há indícios de que as q.1 e q.5, q.6 e q.8, q.10 e q.14, e q.15 e q.19 possuem similaridade.

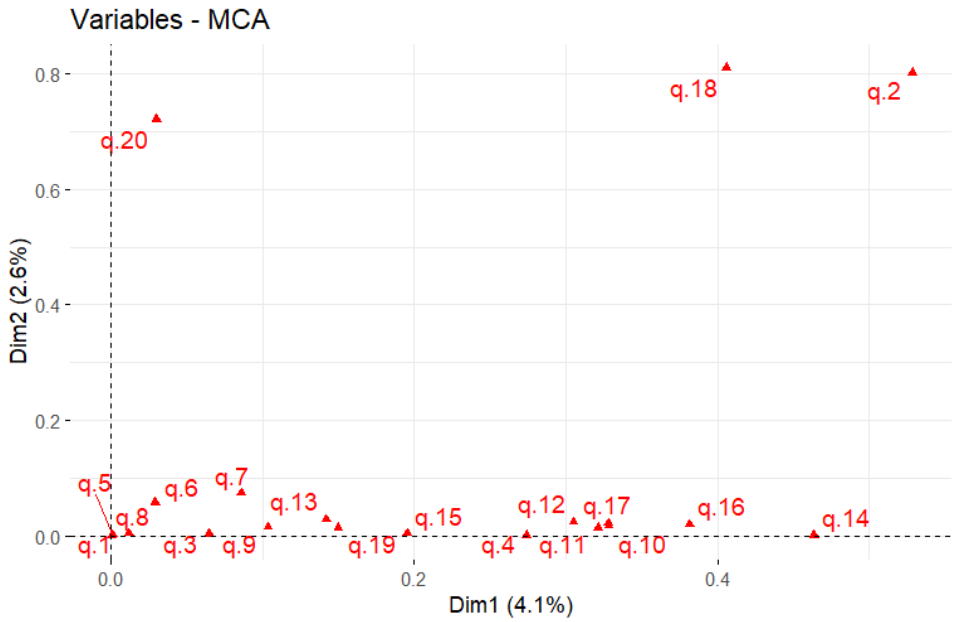


Figura 4.2.11 – Gráfico de Análise de Correspondência Múltipla (MCA) para Medicina em 2022

Em 2022, os padrões gerais se mantêm semelhantes aos de 2019, mas com pequenas variações na distribuição das variáveis. Dim1 explica 3,9% da variabilidade e Dim2 explica 2,7%, levemente menores do que em 2019. As q.1, q.5 e q.8, q.10, q.11 e q1.7, e q.13 e q.19.

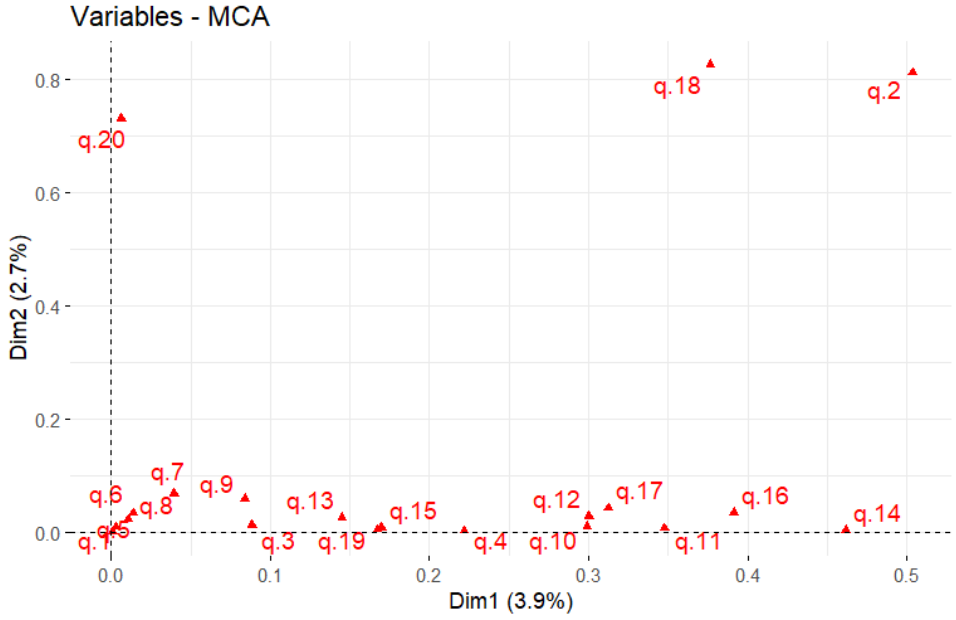


Figura 4.2.12 – Gráfico de Análise de Correspondência Múltipla (MCA) para Medicina em 2023

O gráfico de 2023 apresenta uma estrutura muito semelhante aos anos anteriores, com Dim1 explicando 3,4% e Dim2 explicando 2,7% da variabilidade. Com variáveis q.1 e q.5, q.6 e q.8, q.10 e q.12, q.15 e q.19 associadas.

Desta forma, vamos partir para a modelagem dos modelos de regressão logística, onde MCA será utilizado para definir um dos modelos. Esse modelo é composto por todas as variáveis do qual retiramos as variáveis consideradas associadas.

4.3 Modelos de Regressão Logística

Em relação ao banco de dados utilizado, houve redução de seu volume ao ser restringido para os anos de 2015 a 2023, resultando em 152.507 candidatos não-cotistas investigados. Além disso, na análise socioeducacional, das 30 questões disponíveis, apenas 20 foram consideradas, uma vez que foram as que se mantiveram consistentes ao longo desse período. Logo,para este modelo as variáveis consideradas foram:

1. Qual o seu sexo?
2. Quantos anos você completará até o próximo dia 31 de dezembro?
3. Qual a sua cor ou etnia? (Fonte: IBGE - Censo 2010)
4. Qual o seu estado civil?
5. Você tem alguma deficiência/necessidade educativa especial? (Caso necessite de atendimento especial para a realização das provas, deve solicitar no formulário anterior, em campo específico.)
6. Qual o Estado em que você nasceu?
7. Onde você reside permanentemente?
8. Qual a localização de sua residência?
9. Quantas pessoas residem com você?
10. Qual o nível de instrução do seu pai?
11. Qual o nível de instrução de sua mãe?
12. Qual a renda mensal de sua família?
13. Qual o item cuja descrição de bens mais se aproxima dos bens da sua família?

14. Qual a sua participação na vida econômica da família?
15. Durante o curso superior, você terá que trabalhar?
16. Como você realizou seus estudos de Ensino Fundamental (1º grau)?
17. Como você realizou ou está realizando o Ensino Médio (2º grau ou equivalente)?
18. Quando você concluiu ou concluirá o Ensino Médio (2º grau ou equivalente)?
19. Em que turno você realizou ou está realizando o Ensino Médio (2º grau ou equivalente)?
20. Você frequentou ou frequenta curso pré-vestibular?

Além disso, as variáveis relacionadas a cotas sociais e raciais também foram excluídas já que foram implantadas no ano de 2009 e 2019, respectivamente. (Universidade Estadual de Maringá, 2017) (Agência Estadual de Notícias do Paraná, 2009)

Inicialmente foi ajustado um modelo de regressão logística com todas as variáveis para identificar os principais fatores socioeducacionais associados à aprovação nos processos seletivos da UEM. Em sequência, houve uma avaliação do modelo inicial, que resultou na necessidade de um modelo mais parcimonioso. Logo, foi aplicado o algoritmo **stepAIC**, é uma função do pacote MASS no R que realiza a seleção de variáveis em modelos estatísticos permitindo adicionar e remover variáveis a cada passo, buscando o menor AIC possível. Esse processo resultou em um modelo final otimizado composto por 18 questões (onde foram excluídas as variáveis: "estado civil"e "nível de instrução do seu pai") mais significativas.

A partir do ajuste do modelo iniciamos a análise de resíduos, a partir da Figura 4.3.1 apresentada.

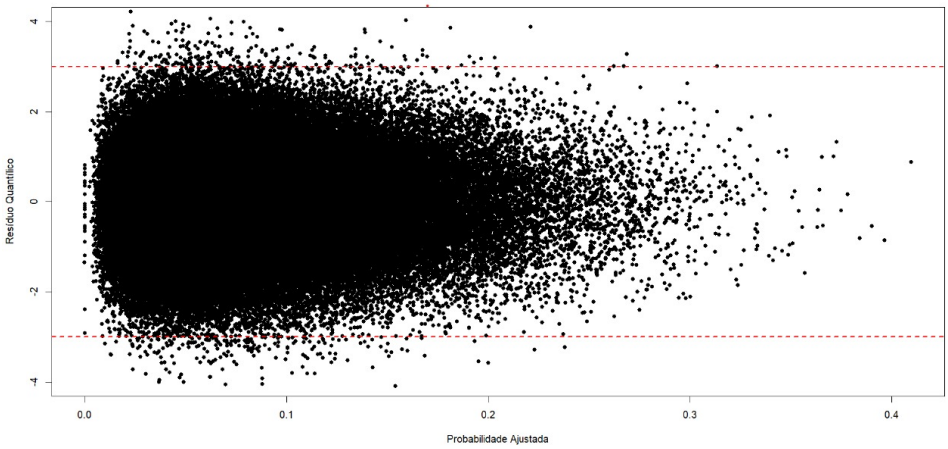


Figura 4.3.1 – Gráfico de Resíduo Quantílico vs. Probabilidade Ajustada

No entanto, como observado na Figura 4.3.1, nota-se a presença de heterocedasticidade, caracterizada pela dispersão não-uniforme dos resíduos ao longo dos valores de probabilidade ajustada , indicando que a variabilidade dos erros não é constante, o que compromete a confiabilidade das estimativas dos parâmetros do modelo. Isso pode causar inferências errôneas sobre a significância estatística das variáveis independentes, levando à inclusão de variáveis sem importância e à exclusão do que realmente faz a diferença.

Como o objetivo é analisar os perfis dos candidatos dos cursos Ed. Física, Enfermagem e Medicina nos anos 2019 (pré-pandemia), 2022 (ano de transição e adaptação) e 2023 (consolidação da pós pandemia), repetimos a análise anterior propondo cinco modelos de regressão logística (denominados m1, m2, m3, m4 e m5) para cada cenário (curso e ano):

- **modelo 1 (m1):** modelo ajustado com todas as variáveis explicativas.
- **modelo 2 (m2):** tomando o modelo 1, este modelo foi ajustado mantendo somente as variáveis explicativas significativas ($\alpha = 5\%$).
- **modelo 3 (m3):** este modelo possui as variáveis explicativas que apresentaram diferenças entre aprovados e não-aprovados.
- **modelo 4 (m4):** modelo ajustado com as variáveis explicativas utilizando MCA, ou seja, foram excluídas do modelo as variáveis relacionadas ou associadas.
- **modelo 5 (m5):** modelo ajustado resultado do m1 com stepAIC.

4.3.1 Educação Física

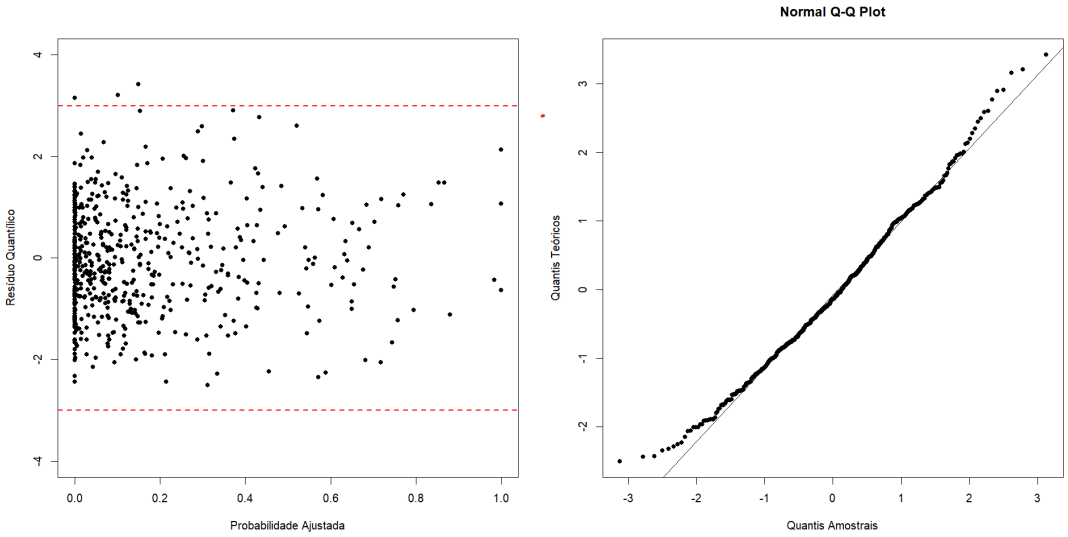
4.3.1.1 2019

Os modelos de regressão para o curso de Ed.Física em 2019 serão apresentados a seguir, sendo aprovação (sim ou não) a variável de interesse:

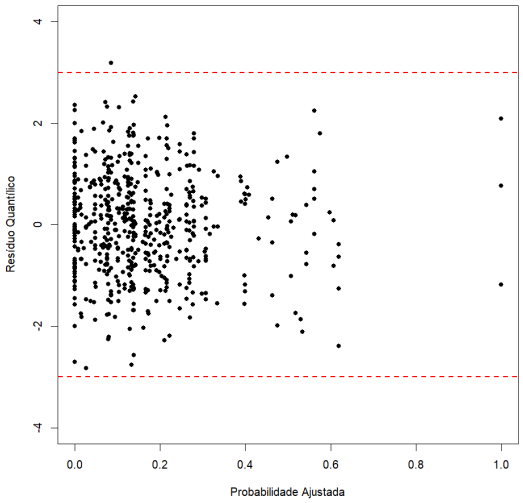
- **m1:** esse modelo é composto por todas as variáveis preditoras (q.1 a q.20);
- **m2:** é constituído por q.7, q.11, q.12 e q.17;
- **m3:** é formado por q.7, q.9 e q.12;
- **m4:** esse modelo é composto pelas q.1, q.2, q.3, q.5, q.6, q.7, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.16, q.17, q.18 e q.19. Foram excluídas (q.4, q.8 e q.20) do modelo uma das variáveis que estão associadas entre si, como vista na Figura 4.2.2;
- **m5:** o modelo de regressão resultante do stepAIC possui q.3, q.7, q.11, q.12 e q.17 como variáveis preditoras.

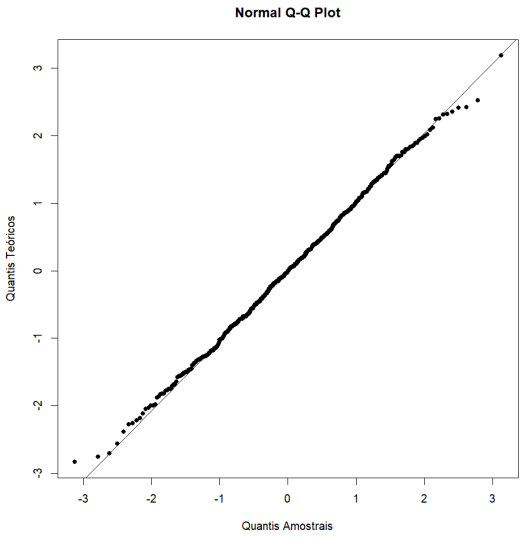
Os gráficos dos resíduos desses modelos serão apresentados, respectivamente, abaixo:

m1

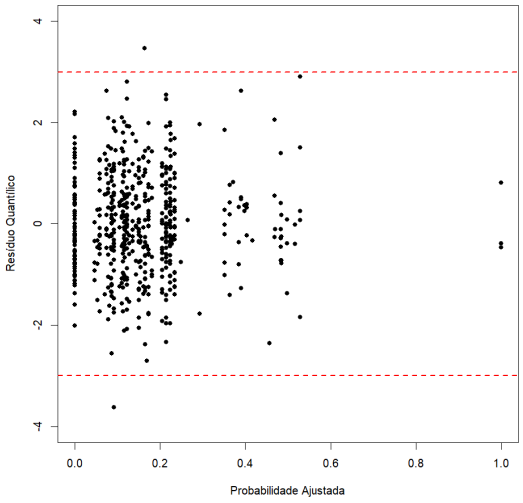


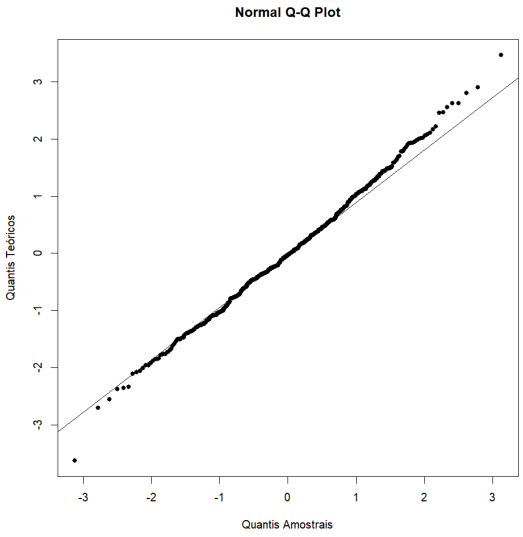
m2


Resíduo Quantílico
Probabilidade Ajustada

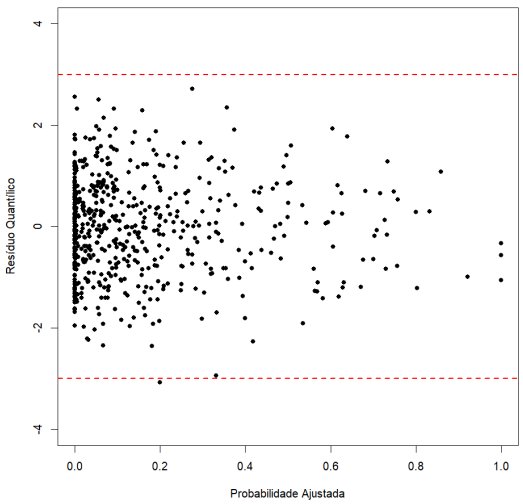

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

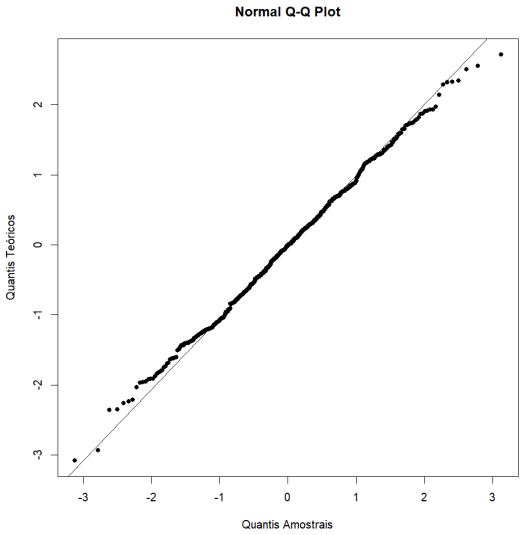
m3


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m4


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m5

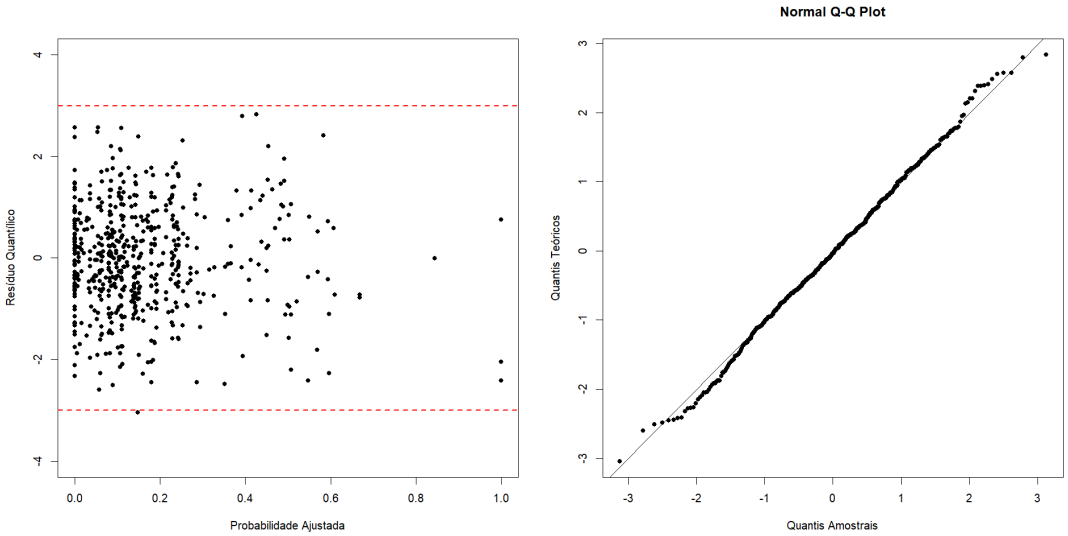


Tabela 2 – Análise gráfica dos resíduos dos modelos ajustados para Ed.Física em 2019.

Para determinar o melhor modelo com base nos critérios AIC e BIC, deve-se observar que valores menores indicam melhor ajuste. O AIC equilibra a qualidade do ajuste com a complexidade do modelo, enquanto o BIC adiciona uma penalização maior para modelos mais complexos. A Tabela 3 apresenta esses valores para os modelos ajustados no curso de Educação Física em 2019.

Tabela 3 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Ed. Física em 2019.

Modelo	**AIC**	**BIC**
m1	525,6	950,3
m2	460,3	577,3
m3	467,0	**557,9**
m4	523,1	917,4
m5	**457,5**	591,8

O modelo m5 apresenta o menor valor de AIC (457,5), sugerindo o melhor ajuste com relação à parcimônia e à qualidade do modelo. E o modelo m2 possui o segundo menor valor de AIC (460,3) e o menor valor de BIC (557,9), o que o torna competitivo, especialmente em contextos onde a penalização por complexidade é mais relevante.

Escolhemos m5 que se adequa melhor aos dados, desta forma temos:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-18,52 + 1,73 \cdot q.32 + 1,22 \cdot q.33 + 0,31 \cdot q.34 \\ &- 16,44 \cdot q.35 - 0,61 \cdot q.72 + 1,60 \cdot q.73 \\ &- 17,68 \cdot q.75 - 31,37 \cdot q.76 + 0,55 \cdot q.77 \\ &- 16,76 \cdot q.78 + 0,23 \cdot q.79\end{aligned} \tag{4.3.1}$$

Legenda dos coeficientes:

- $q.32$: Preta;
- $q.33$: Amarela;
- $q.34$: Parda;
- $q.35$: Indígena;
- $q.72$:Reside em outra cidade do Estado do Paraná situada na região noroeste;
- $q.73$: Reside em uma cidade do Estado do Paraná não situada na região noroeste;
- $q.75$: Reside em cidade do Estado do Rio Grande do Sul;
- $q.76$: Reside em cidade do Estado de São Paulo;
- $q.77$: Reside em cidade do Estado do Mato Grosso;
- $q.78$: Reside em cidade do Estado do Mato Grosso do Sul;
- $q.79$: Reside em cidade situada em Estado não relacionado nos itens anteriores.

A Tabela 4 apresenta os resultados do modelo de Regressão Logística (m5) aplicado aos dados do curso de Educação Física no processo seletivo de 2019. O modelo avalia o impacto das variáveis socioeducacionais (q.32 a q.174) na probabilidade de aprovação dos candidatos. São apresentados os coeficientes estimados (Estimate), erros padrão (Std. Error), valores z, valores-p (Pr(>|z|)), razões de chances (OR) e seus respectivos intervalos de confiança (IC 2,5% - 97,5%). Variáveis com (p-valor < 0,05) indicam uma associação estatisticamente significativa com aprovação, sendo destacadas na tabela. Vale ressaltar que razões de chances superiores a 1 indicam um aumento na probabilidade de ocorrência do desfecho (aprovação do curso no ano em questão), enquanto razões de chances inferiores a 1 sinalizam uma diminuição dessa probabilidade. Os intervalos de confiança representam a precisão das estimativas; intervalos amplos podem refletir uma alta variabilidade nos dados.

Tabela 4 – Modelo de Regressão Logística m5 para Ed. Física em 2019

Variável	Estimate	Std. Error	z value	Pr(>|z|)	OR	IC 95%
q.32	1.725	0.670	2.576	**0.010**	5.615	[1.511, 20.871]
q.33	1.221	0.551	2.218	**0.027**	3.391	[1.152, 9.977]
q.34	0.311	0.300	1.038	0.299	1.365	[0.759, 2.454]
q.72	-0.611	0.317	-1.927	0.054	0.543	[0.292, 1.011]
q.73	1.603	0.393	4.082	**0.0004**	4.970	[2.301, 10.732]
q.79	0.230	1.193	0.193	0.847	1.259	[0.121, 13.055]
q.112	-0.293	1.188	-0.247	0.805	0.746	[0.073, 7.659]
q.113	-2.921	1.567	-1.864	0.062	0.054	[0.003, 1.162]
q.114	0.033	1.191	0.028	0.978	1.034	[0.100, 10.672]
q.115	-0.930	1.161	-0.801	0.423	0.395	[0.041, 3.843]
q.116	-0.298	1.230	-0.242	0.809	0.743	[0.067, 8.278]
q.117	-0.579	1.192	-0.486	0.627	0.560	[0.054, 5.794]
q.172	0.580	0.297	1.952	0.051	1.786	[0.998, 3.198]
q.173	-0.727	0.831	-0.875	0.382	0.483	[0.095, 2.466]
q.174	-0.621	0.779	-0.797	0.426	0.538	[0.117, 2.475]

Entre as variáveis com efeito positivo e significativo na aprovação, destacam-se variável "Raça"categoria "Preta"(q.32), "Raça"categoria "Amarela"(q.33) e "Residência permanente"categoria "cidade do Estado do Paraná não situada na região noroeste"(q.73). A variável Raça Preta apresenta uma razão de chances de 5,61 (IC 95% = [1,51, 20,87]), indicando que candidatos com essa característica têm 5,61 vezes mais chances de serem aprovados nesse curso do que os da categoria "Branca". A variável Raça Amarela também tem um impacto positivo, aumentando em 3,39 vezes as chances de aprovação (IC 95% = [1,15, 9,98]). A variável "Residência permanente em cidade do Estado do Paraná não situada na região noroeste"se destaca com uma razão de chances de 4,97 (IC 95% = [2,30, 10,73]), mostrando um forte efeito positivo e significância estatística robusta.

Algumas variáveis apresentaram efeito negativo ou próximo à neutralidade. A variável "Residência permanente"categoria "outra cidade do Estado do Paraná situada na região noroeste"(q.72) possui uma razão de chances de 0,54 (IC 95% = [0,29, 1,01]), sugerindo uma redução de 46% nas chances de aprovação, embora o intervalo de confiança quase cruze 1, indicando significância marginal. A variável "nível de instrução de sua mãe"com categoria "Ensino Fundamental/1º grau completo"(q.113), com uma OR de 0,05 (IC 95% = [0,002, 1,16]), mostra uma redução drástica nas chances de aprovação, mas o intervalo de confiança inclui o valor 1, o que requer cautela na interpretação. Por outro lado, algumas variáveis não apresentaram significância estatística, como q.34 (OR = 1,36) e q.172 (OR =

1,79), que apesar de possuírem OR superiores a 1, seus intervalos de confiança incluem o valor 1, indicando ausência de efeito significativo. Outras variáveis, como q.114 (OR = 1,03), têm efeitos praticamente nulos.

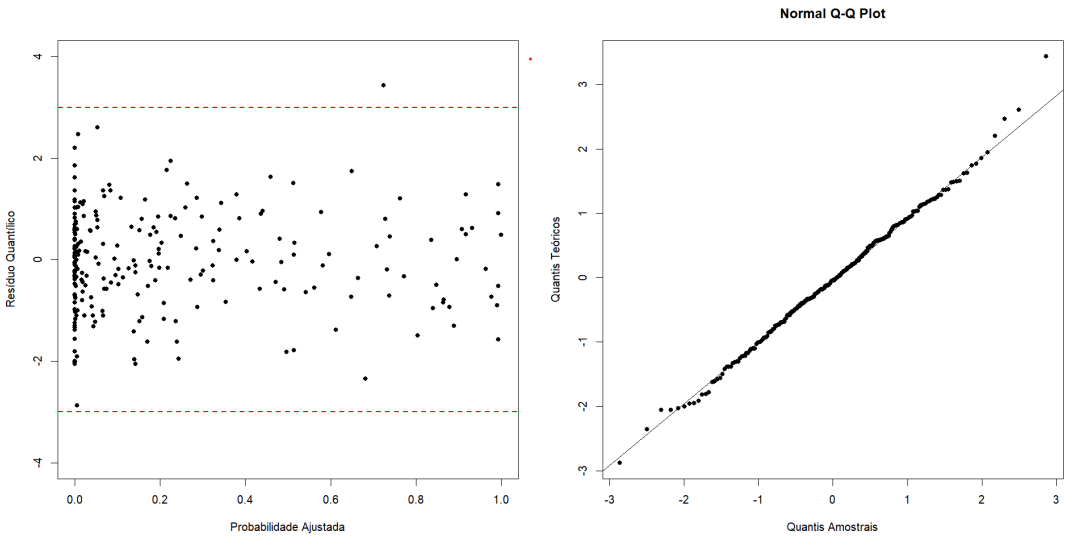
4.3.1.2 2022

Os modelos de regressão para Ed.Física em 2022 são:

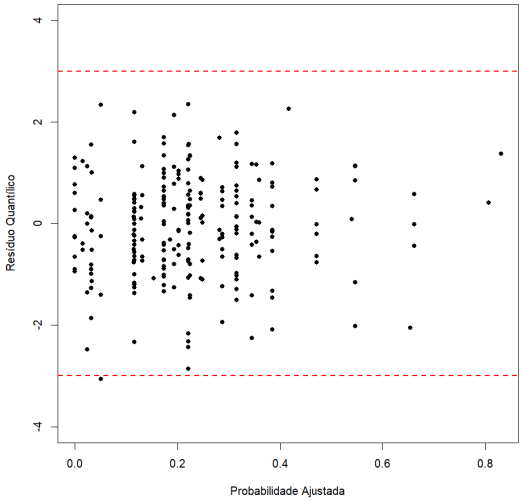
- **m1:** q.1 a q.20;
- **m2:** q.1, q.13 e q.15;
- **m3:** q.7, q.9 e q.12;
- **m4:** do modelo completo foram removidos q.4, q.8, q10 devido a sua associação com outras variáveis como foi mostrado na Figura 4.2.3;
- **m5:** pelo stepAIC, temos q.13, q.15, q.17 e q.18.

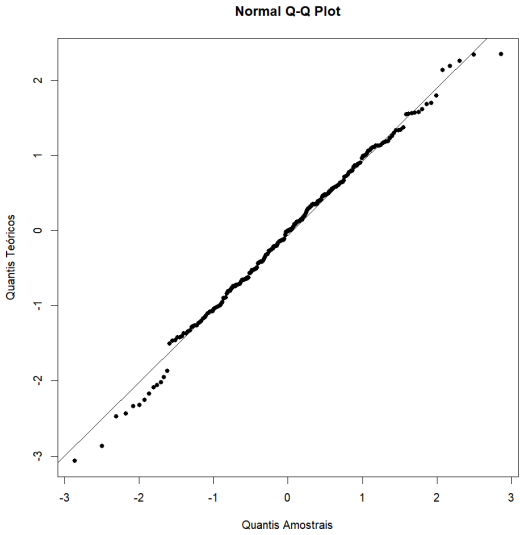
Os gráficos dos resíduos e a tabela dos valores dos critérios de informação, AIC e BIC, desses modelos serão apresentados, respectivamente, a seguir:

m1

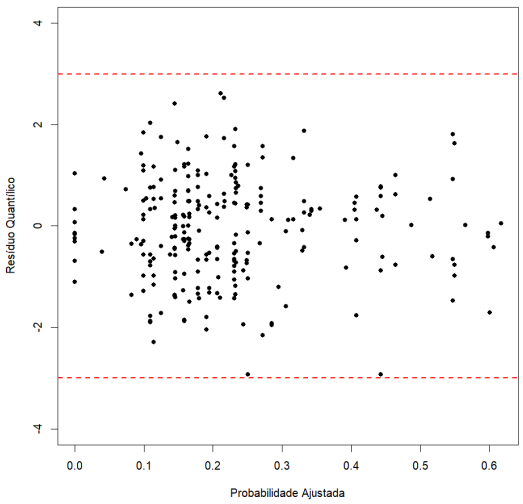


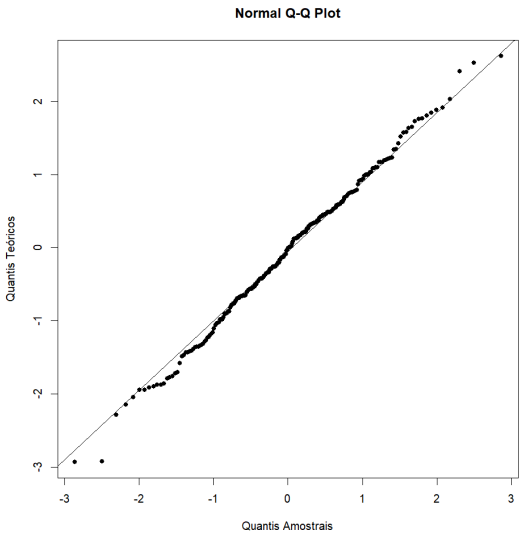
m2


Resíduo Quantílico
Probabilidade Ajustada

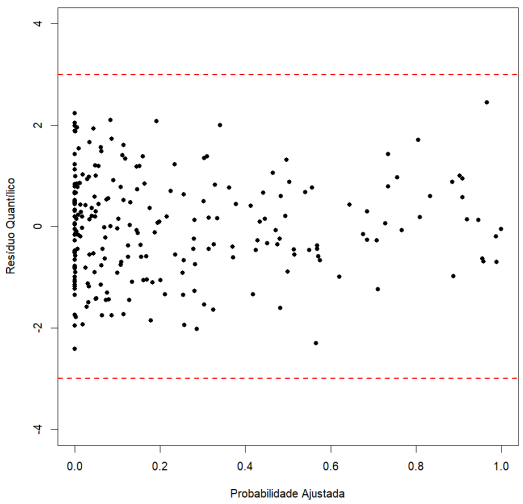

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

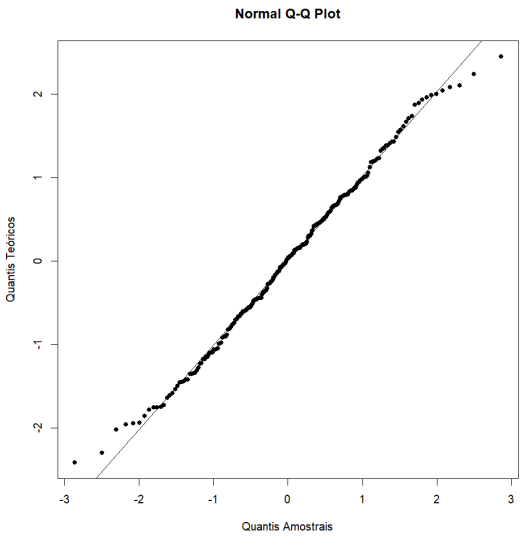
m3


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m4


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m5

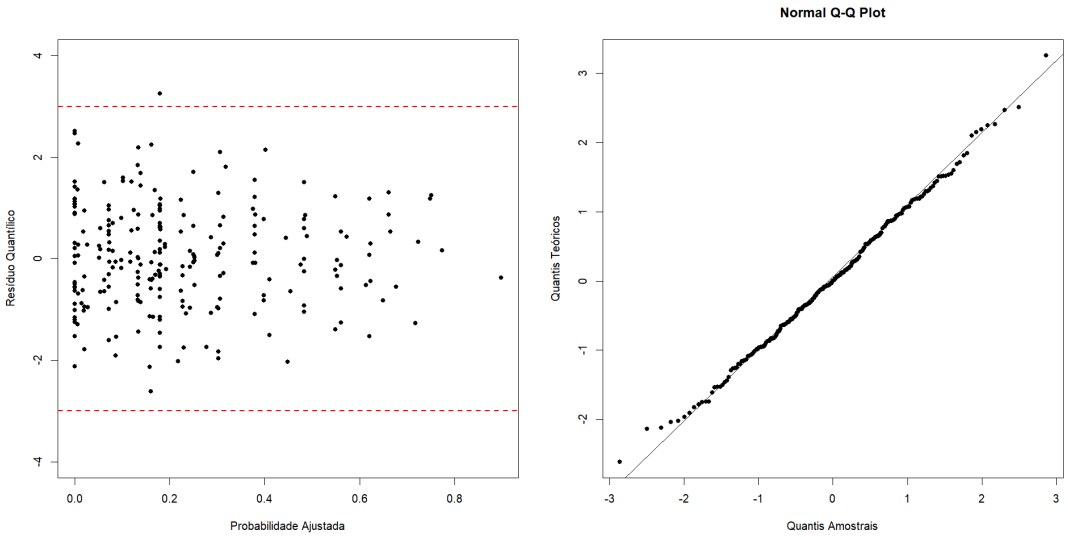


Tabela 5 – Análise gráfica dos resíduos dos modelos ajustados para Ed.Física em 2022.

Tabela 6 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Ed. Física em 2022.

Modelo	**AIC**	**BIC**
m1	323,2	653,8
m2	252,5	**301,2**
m3	271,6	341,3
m4	317,4	609,7
m5	**243,2**	319,8

O m5 apresenta o menor valor de AIC (243,2), e o modelo m2 possui o menor valor de BIC (301,2), conforme mostrado na Tabela 6. Logo, o modelo m5 foi adequado para a análise do curso de Educação Física em 2022, juntamente com a análise do gráfico de resíduos, e pode ser apresentado da seguinte da seguinte forma:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-0,941 + 3,109 \cdot q.132 + 2,372 \cdot q.133 \\ &+ 2,385 \cdot q.134 + 2,714 \cdot q.135 \\ &- 14,557 \cdot q.136 + 4,354 \cdot q.137 \\ &- 15,457 \cdot q.138 - 14,167 \cdot q.139 \\ &- 1,923 \cdot q.152 - 1,499 \cdot q.153 \\ &- 2,964 \cdot q.154 - 2,766 \cdot q.155 \\ &- 1,139 \cdot q.182 - 18,041 \cdot q.183 \\ &- 1,154 \cdot q.184 - 1,379 \cdot q.185 \\ &- 1,029 \cdot q.186 + 1,335 \cdot q.187\end{aligned} \tag{4.3.2}$$

Legenda:

- $q.132$: Não possui casa própria mas possui carro ou moto;
- $q.133$: Possui casa própria e carro ou moto;
- $q.134$: Possui casa própria, carro ou moto e outros imóveis urbanos;
- $q.135$: Possui casa própria, carro ou moto e caminhão;
- $q.136$: Possui casa própria, carro ou moto e propriedade rural;
- $q.137$: Possui casa própria, carro ou moto, caminhão e propriedade rural;
- $q.138$: Possui casa própria, carro ou moto, caminhão, propriedade rural e outros imóveis;
- $q.139$: Possui mais bens além dos relacionados no item anterior;
- $q.152$: Terá que trabalhar desde o primeiro ano, em tempo parcial;
- $q.153$: Terá que trabalhar desde o primeiro ano, em tempo integral;
- $q.154$: Não sabe se terá que trabalhar;
- $q.155$: Não terá que trabalhar;
- $q.182$: Concluiu o Ensino Médio há quatro anos;
- $q.183$: Concluiu o Ensino Médio há três anos;
- $q.184$: Concluiu o Ensino Médio há dois anos;

- $q.185$: Concluiu o Ensino Médio no ano passado;
- $q.186$: Concluiu o Ensino Médio neste ano;
- $q.187$: Concluirá o Ensino Médio no próximo ano.

Tabela 7 – Modelo de Regressão Logística m5 para Ed. Física em 2022

Variável	Estimate	Std. Error	z value	Pr(>|z|)	OR	IC 95%
q.132	3.109	1.180	2.635	**0.008**	**22.40**	[3.223, 486.79]
q.133	2.372	1.136	2.089	**0.037**	**10.72**	[1.745, 222.06]
q.134	2.385	1.213	1.967	**0.049**	**10.86**	[1.424, 244.75]
q.135	2.714	1.911	1.420	0.156	15.09	[0.303, 837.08]
q.137	4.354	1.800	2.418	**0.016**	**77.76**	[2.972, 4904.80]
q.152	-1.923	1.074	-1.790	0.073	0.15	[0.014, 1.134]
q.153	-1.500	1.094	-1.371	0.170	0.22	[0.020, 1.816]
q.154	-2.964	1.139	-2.601	**0.009**	**0.05**	[0.004, 0.445]
q.155	-2.766	1.431	-1.933	0.053	0.06	[0.003, 0.884]
q.172	0.688	0.409	1.683	0.092	1.99	[0.891, 4.468]
q.173	-0.148	0.756	-0.196	0.845	0.86	[0.166, 3.481]
q.182	-1.140	0.975	-1.169	0.242	0.32	[0.036, 1.913]
q.184	-1.155	0.698	-1.655	0.098	0.32	[0.072, 1.164]
q.185	-1.379	0.752	-1.833	0.067	0.25	[0.051, 1.023]
q.186	-1.029	0.466	-2.208	**0.027**	**0.36**	[0.142, 0.890]
q.187	1.335	0.966	1.382	0.167	3.80	[0.595, 28.025]

Conforme a Tabela 7, podemos concluir que candidatos que não possuem casa própria, mas possuem carro ou moto têm 22,4 vezes mais chances de aprovação em relação aos que não possuem casa própria nem carro ou moto; e possuir casa própria e carro ou moto aumenta em 10,72 vezes as chances de aprovação no curso no referente ano. Candidatos que têm casa própria, carro ou moto e outros imóveis urbanos elevam suas chances em 10,86 vezes. Os com casa própria, carro ou moto, caminhão e propriedade rural têm as maiores chances, com um aumento de 77,76 vezes. Temos também os candidatos que não sabem se terão que trabalhar, isso diminui as chances de aprovação em 95% (OR = 0,05) e concluir o Ensino Médio no ano em questão reduz as chances de aprovação em 64% (OR = 0,36) em Ed Física no ano de 2022.

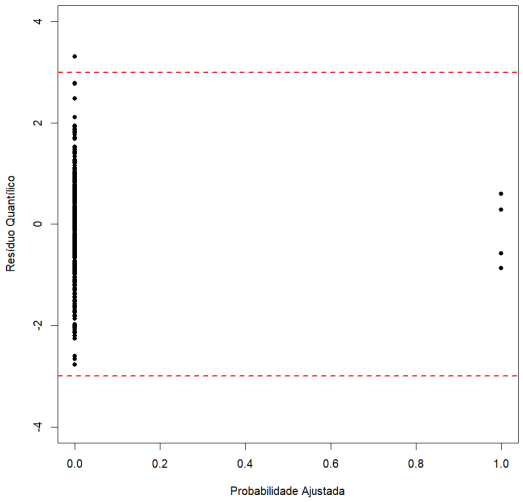
4.3.1.3 2023

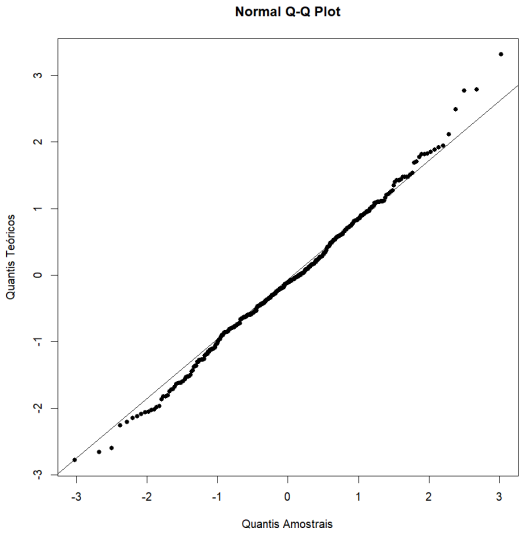
Para Ed.Física em 2023, temos os modelos de regressão :

- **m1:** modelo composto por todas as variáveis preditoras;
- **m2:** somente q.11;
- **m3:** q.7, q.9 e q.12;
- **m4:** q.1, q.2, q.3, q.4, q.5, q.6, q.7, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.17, q.18 e q.19, conforme mostra a Figura 4.2.3;
- **m5:** q.1, q.8, q.15, q.16 e q.17, pelo stepAIC.

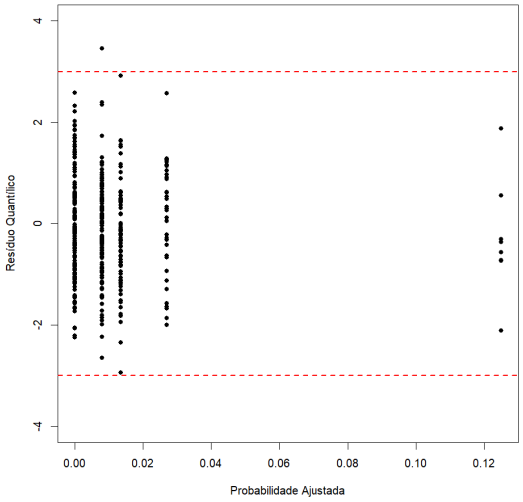
Os gráficos dos resíduos desses modelos serão apresentados, respectivamente:

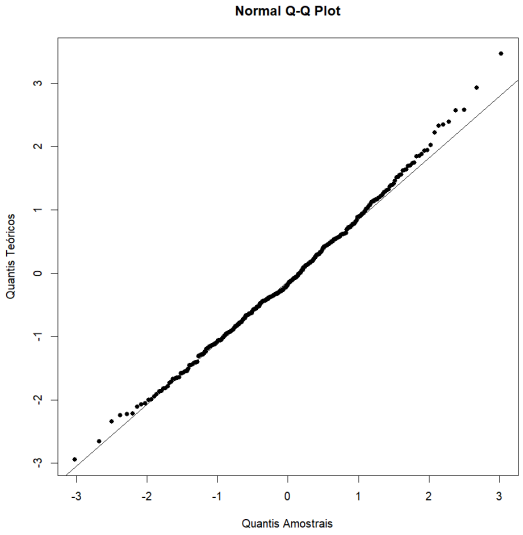
m1



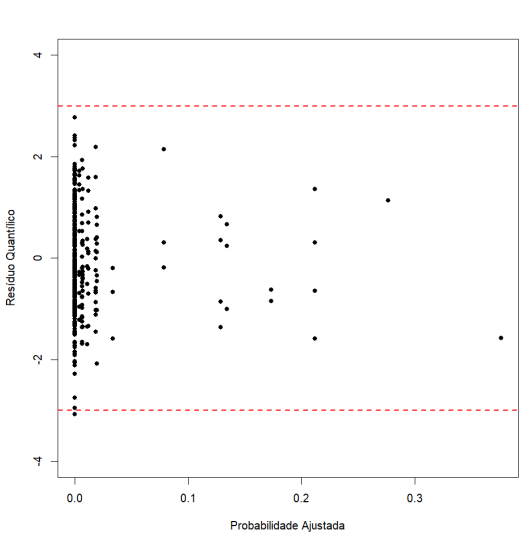


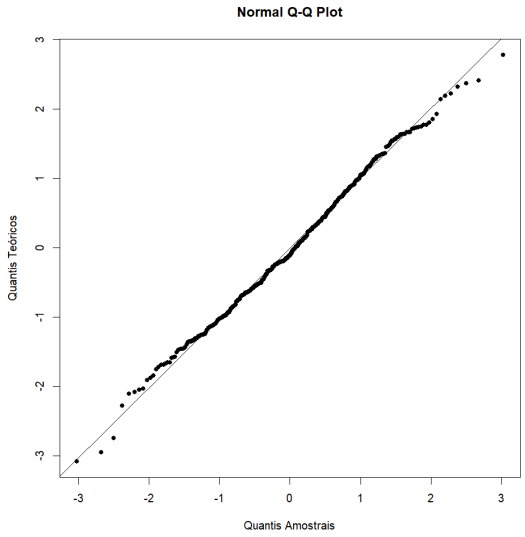
m2



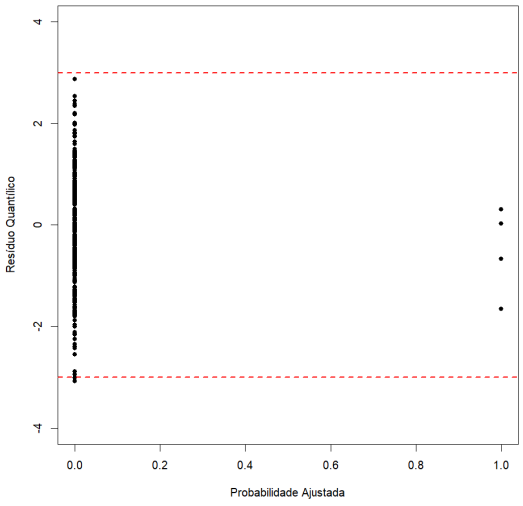


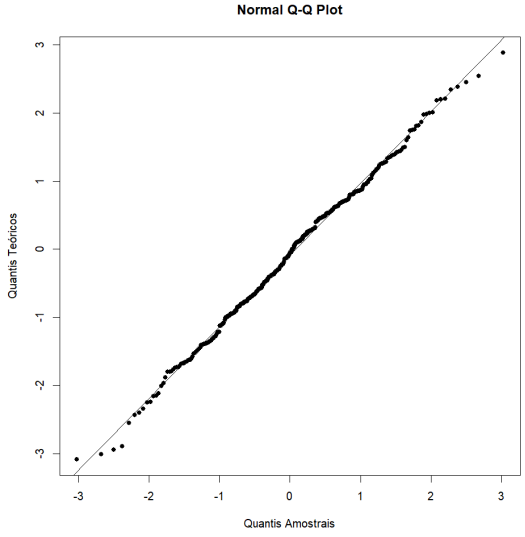
m3



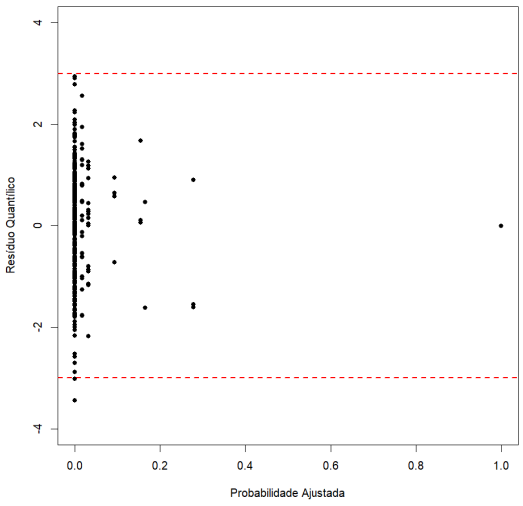


m4





m5



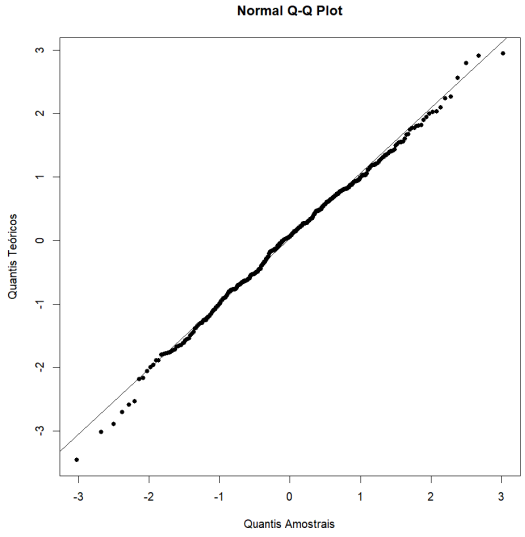


Tabela 8 – Análise gráfica dos resíduos dos modelos ajustados para Ed.Física em 2023.

A tabela abaixo apresenta os valores dos AIC e BIC para os modelos ajustados no curso de Educação Física em 2023.

Tabela 9 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Ed. Física em 2023.

Modelo	**AIC**	**BIC**
m1	194,0	580,4
m2	55,4	**91,3**
m3	66,1	145,8
m4	178,0	532,5
m5	**47,8**	103,5

Os modelos m2 e m5 foram considerados os dois melhores modelos de regressão de acordo com os valores de BIC e AIC, nesta ordem. No entanto, após realizar a analise gráfica da Figura 8, optamos por m5 descrito como:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-22,473 + 39,218 \cdot q.12 + 23,257 \cdot q.82 \\ &- 20,753 \cdot q.152 - 20,170 \cdot q.153 \\ &- 77,239 \cdot q.154 - 40,614 \cdot q.155 \\ &+ 20,855 \cdot q.162 - 16,577 \cdot q.163 \\ &+ 40,944 \cdot q.164 - 38,554 \cdot q.172 \\ &- 58,473 \cdot q.173 - 59,983 \cdot q.174 \\ &+ 17,660 \cdot q.175\end{aligned} \tag{4.3.3}$$

- $q.12$: Sexo Feminino;
- $q.82$: Residência localizada na zona rural;
- $q.152$: Terá que trabalhar desde o primeiro ano, em tempo parcial;
- $q.153$: Terá que trabalhar desde o primeiro ano, em tempo integral;
- $q.154$: Não sabe se terá que trabalhar;
- $q.155$: Não terá que trabalhar;
- $q.162$: Realizou o Ensino Fundamental integralmente em escola particular;
- $q.163$: Realizou o Ensino Fundamental maior parte em escola pública;
- $q.164$: Realizou o Ensino Fundamental maior parte em escola particular;

- $q.172$: Realizou o Ensino Médio integralmente em escola particular;
- $q.173$: Realizou o Ensino Médio maior parte em escola pública;
- $q.174$: Realizou o Ensino Médio maior parte em escola particular;
- $q.175$: Realizou o Ensino Médio maior parte em escolas comunitárias/CNEC;

Tabela 10 – Modelo de Regressão Logística m5 para Ed. Física em 2023

Variável	**Estimate**	**Std. Error**	**Z-Valor**	**Valor-p**
(Intercept)	-22.473	8164.062	-0.003	0.998
q.82	23.257	9609.098	0.002	0.998
q.152	-20.753	7398.496	-0.003	0.998
q.153	-20.170	7398.496	-0.003	0.998
q.162	20.855	8164.062	0.003	0.998
q.174	-59.983	32914.543	-0.002	0.999

O modelo m5 sugere que certas variáveis têm forte influência nas chances de aprovação no curso de Educação Física em 2023. No entanto, a falta de significância estatística e os intervalos de confiança amplos indicam baixa confiabilidade nas estimativas.

Por fim, ao comparar os modelos finais selecionados nos anos 2019, 2022 e 2023, percebe-se uma consistência da variável q.17(Como você realizou ou está realizando o Ensino Médio (2º grau ou equivalente)?), apesar de variações pontuais no conjunto de preditores selecionados. Nos anos pós pandemia (2022 e 2023) temos as variáveis q.15 (Durante o curso superior, você terá que trabalhar?) e q.17 em comum. O modelo m5 demonstrou maior capacidade de generalização e robustez em diferentes anos, confirmando sua adequação como o modelo mais eficaz para prever a aprovação no curso de Educação Física nos anos selecionados.

A trajetória do curso de Educação Física entre os anos analisados reflete a influência crescente das condições socioeconômicas e da estabilidade financeira na aprovação dos candidatos. O ano de 2019 apresentou um cenário mais voltado para a diversidade racial e geográfica, enquanto 2022 e 2023 evidenciaram a importância da posse de bens e da segurança econômica. A ausência de variáveis estatisticamente significativas em 2023 sugere uma possível mudança no perfil dos candidatos ou um desafio na coleta e modelagem dos dados.

4.3.2 Enfermagem

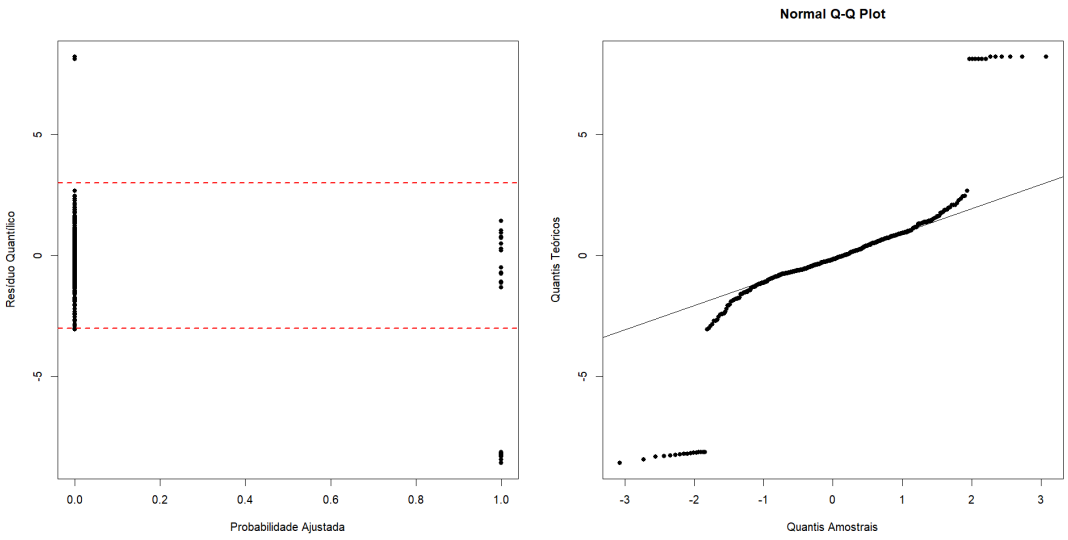
4.3.2.1 2019

Para Enfermagem em 2019, temos os modelos de regressão :

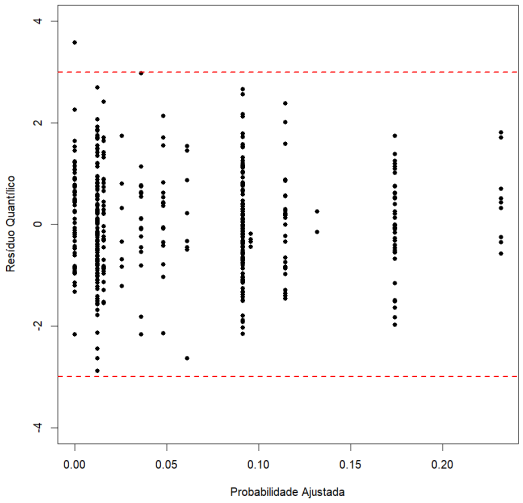
- **m1:** q.1 a q.20;
- **m2:** q.17 e q.18;
- **m3:** q.2, q.7, q.10, q.11, q.12, q.13, q.16 e q.17;
- **m4:** q.1, q.2, q.3, q.4, q.5, q.6, q.7, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.16, q.17, q.18, q.19 e q.20;
- **m5:** q.2, q.5, q.6, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.16, q.17, q.18, q.19 e q.20.

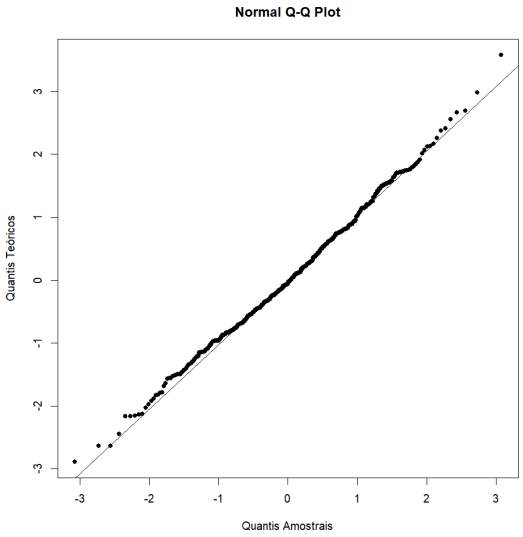
Seguem, respectivamente, os gráficos dos resíduos desses modelos e a tabela com os valores de AIC e BIC:

m1

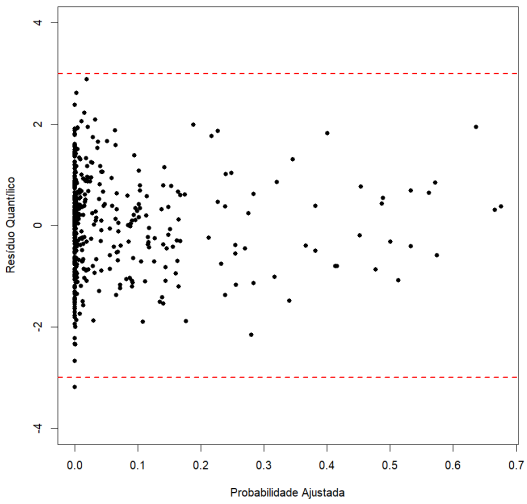


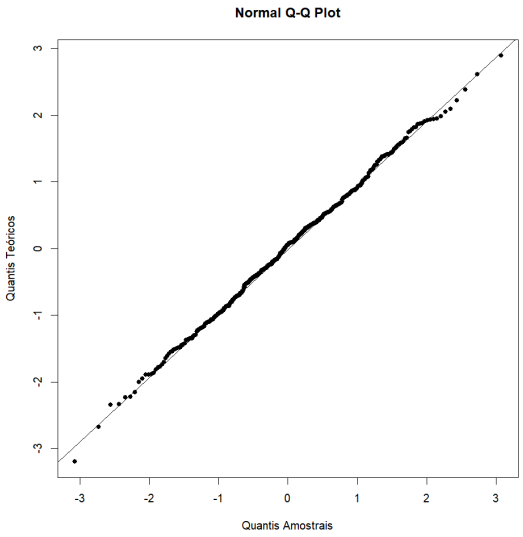
m2


Resíduo Quantílico
Probabilidade Ajustada

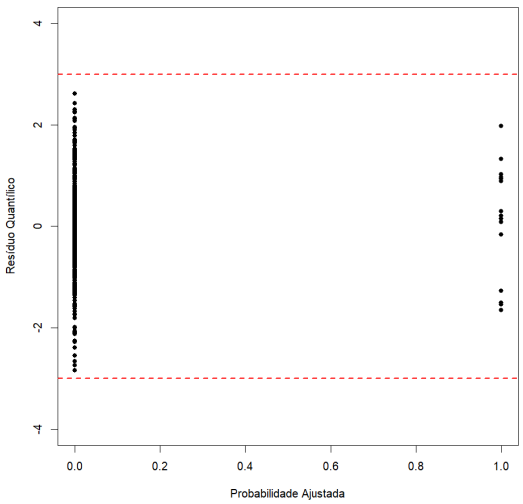

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

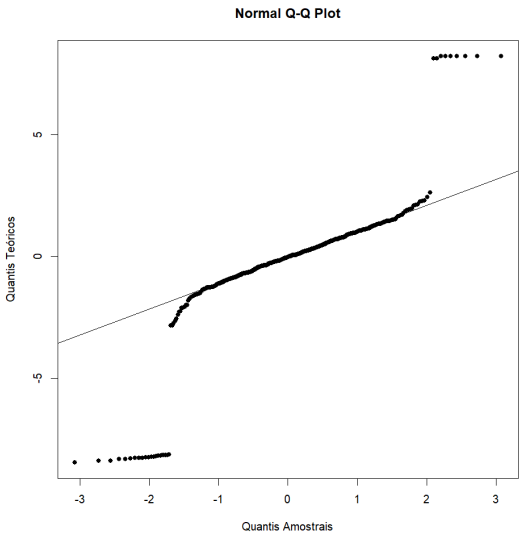
m3


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m4


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m5

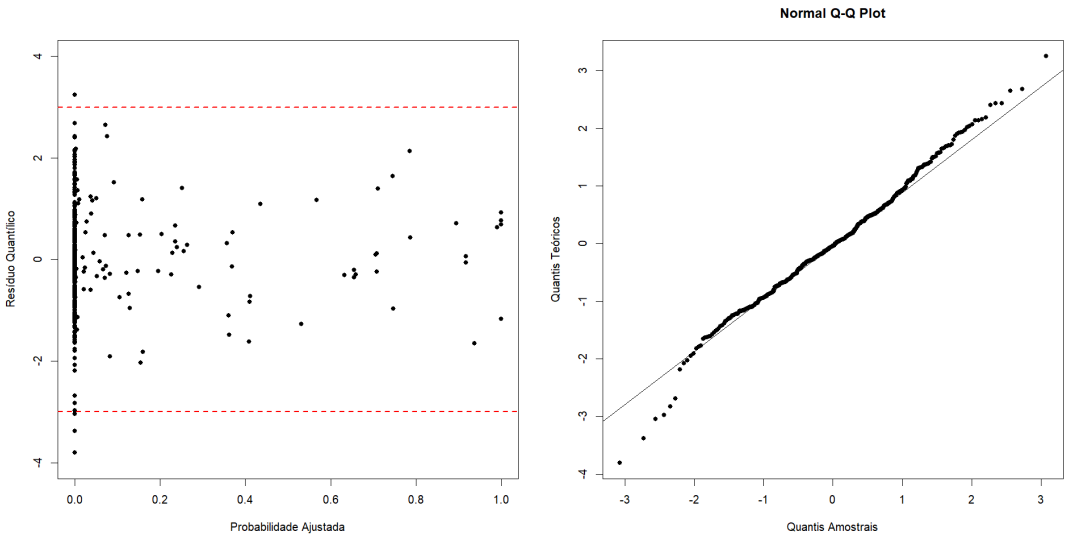


Tabela 11 – Análise gráfica dos resíduos dos modelos ajustados para Enfermagem em 2019.

Tabela 12 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Enfermagem em 2019.

Modelo	**AIC**	**BIC**
m1	2210,4	2609,9
m2	**197,8**	**239,4**
m3	235,6	460,3
m4	2568,8	2964,1
m5	225,5	566,7

Através da análise gráfica dos resíduos na Figura 11 e da tabela de AIC e BIC, temos que m2 é melhor modelo já que possui os menores valores de tais métricas e graficamente, apresenta uma melhor dispersão dos resíduos, sem um padrão muito evidente, menor presença de outliers e estão mais alinhados com a distribuição normal em comparação aos demais modelos. Logo, m2 é dado por:

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -19,700 + 2,083 \cdot q.172 - 15,360 \cdot q.173 + 1,396 \cdot q.174 \tag{4.3.4}$$

Componentes do Modelo:

- $q.172$: Realizou o Ensino Médio integralmente em escola particular;
- $q.173$: Realizou o Ensino Médio maior parte em escola pública;

- $q.174$: Realizou o Ensino Médio maior parte em escola particular.

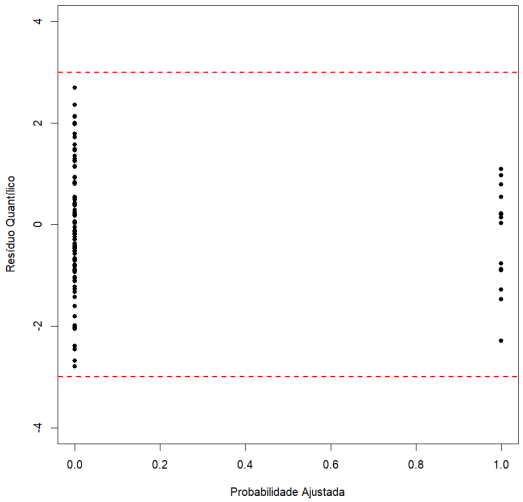
A análise do modelo de regressão logística para o curso de Enfermagem em 2019 revelou que a variável "Realizou o Ensino Médio integralmente em escola particular"(q.172) é a única variável significativa no modelo(valor-p = 0,0011), o que é consideravelmente menor do que o nível de significância padrão de 0,05. Isso indica uma forte evidência estatística de que essa variável influencia a probabilidade de aprovação no processo seletivo do curso. A razão de chances para essa variável é 8,0312, sugerindo que candidatos que cursaram o Ensino Médio em escola particular têm aproximadamente 8 vezes mais chances de serem aprovados em comparação com aqueles que não o fizeram. Além disso, o fato do intervalo de confiança não cruzar o valor 1 reforça a significância e o impacto positivo dessa variável.

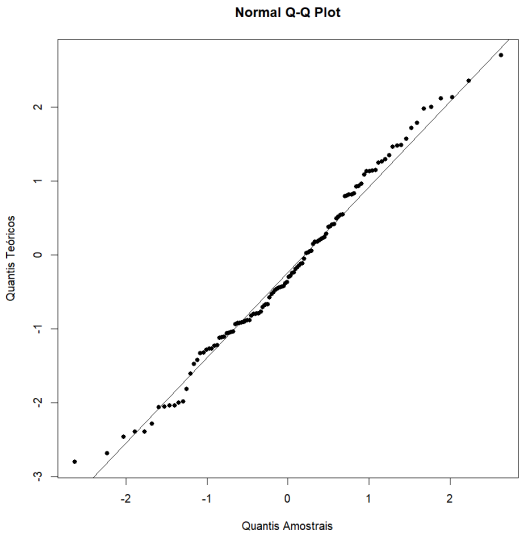
4.3.2.2 2022

Para Enfermagem em 2022, temos os modelos de regressão :

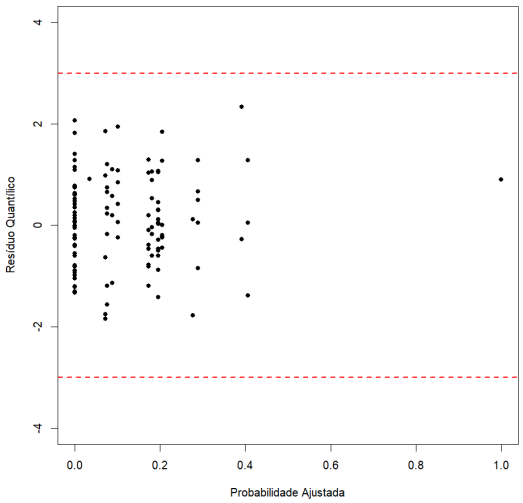
- **m1:** q.1 a q.20.
- **m2:** q.10 e q.12;
- **m3:** q.2, q.7, q.9 e q.14;
- **m4:** q.1, q.2, q.3, q.4, q.5, q.6, q.7, q.9, q.10, q.11, q.12, q.14, q.15, q.16, q.17, q.18, q.19 e q.20;
- **m5:** q.6, q.9, q.12, q.15, q.16, q.17 e q.19.

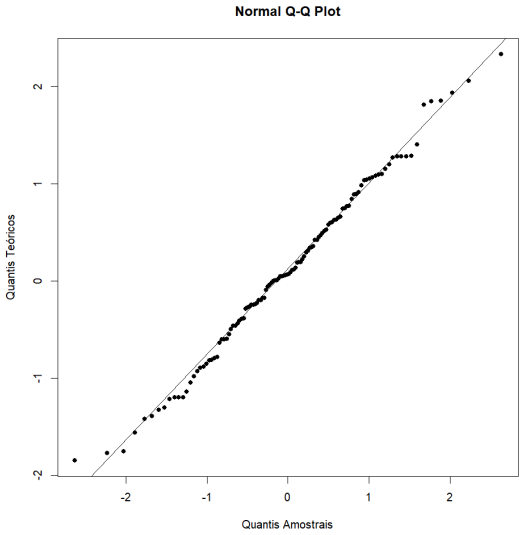
m1


Resíduo Quantílico
Probabilidade Ajustada
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

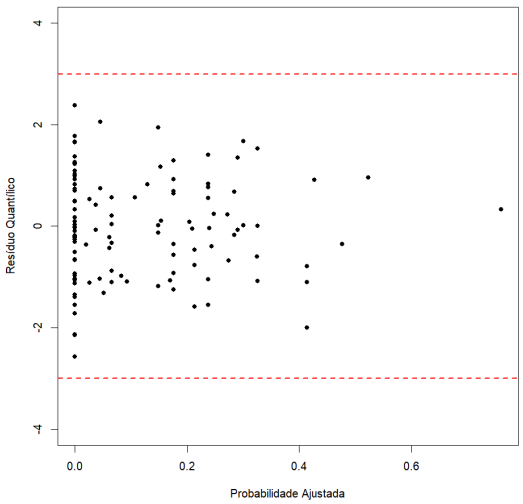

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-3
-2
-1
0
1
2

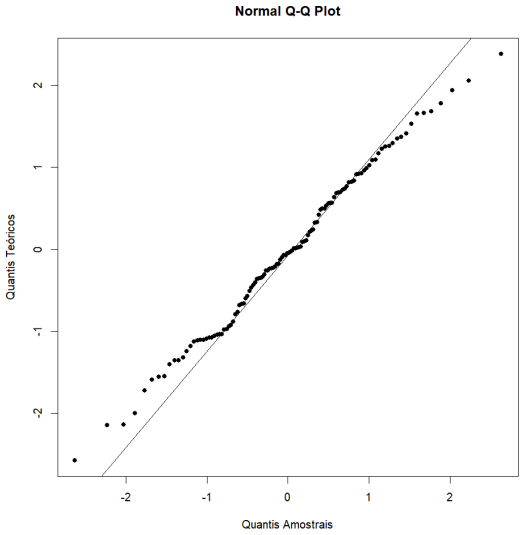
m2


Resíduo Quantílico
Probabilidade Ajustada
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4

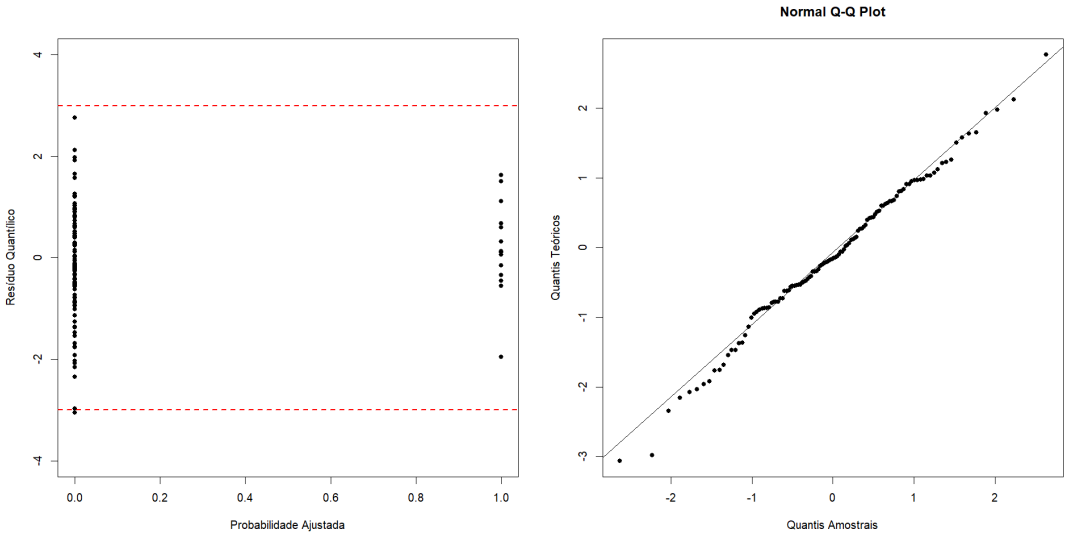

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-2
-1
0
1
2

m3


Resíduo Quantílico
Probabilidade Ajustada
0.0
0.2
0.4
0.6
-4
-2
0
2
4


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-2
-1
0
1
2

m4



m5

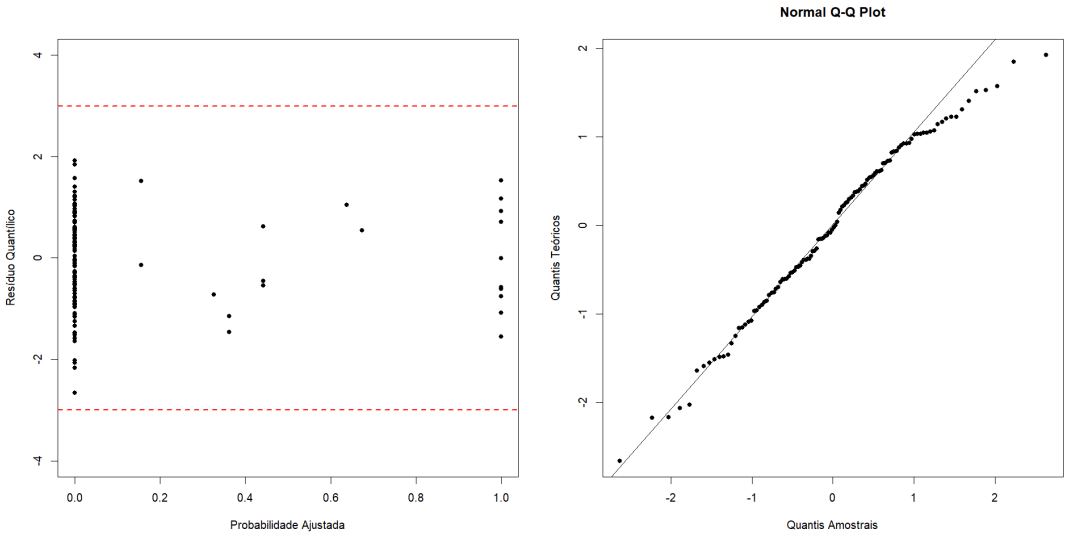


Tabela 13 – Análise gráfica dos resíduos dos modelos ajustados para Enfermagem em 2022.

Tabela 14 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Enfermagem em 2022.

Modelo	**AIC**	**BIC**
m1	170,0	405,5
m2	96,5	**140,9**
m3	107,8	171,5
m4	156,0	372,1
m5	**74,3**	160,1

Ao observarmos a Tabela 14, podemos observar que os menores valores de AIC e BIC pertencem, respectivamente, a m5 (74,3) e m2 (140,9). Por fim, resolvemos optar pelo m2,

devido ao gráfico de resíduos que apresenta uma melhor dispersão dos resíduos indicando um ajuste um pouco mais equilibrado. Resultando no modelo abaixo:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-38,780 + 18,065 \cdot q.102 + 19,666 \cdot q.103 \\ &+ 0,600 \cdot q.104 + 19,213 \cdot q.105 + 20,184 \cdot q.106 \\ &+ 19,059 \cdot q.107 + 0,938 \cdot q.108 + 0,912 \cdot q.109 \\ &+ 18,214 \cdot q.122 - 0,707 \cdot q.123 + 18,156 \cdot q.124 \\ &+ 17,390 \cdot q.125 + 0,067 \cdot q.126 + 17,301 \cdot q.127 \\ &+ 39,162 \cdot q.128\end{aligned} \tag{4.3.5}$$

Definição das Variáveis:

- $q.102$: Pai possui Ensino Fundamental/1º grau incompleto;
- $q.103$: Pai possui Ensino Fundamental/1º grau completo;
- $q.104$: Pai possui Ensino Médio/2º grau incompleto;
- $q.105$: Pai possui Ensino Médio/2º grau completo;
- $q.106$: Pai possui Ensino Superior incompleto;
- $q.107$: Pai possui Ensino Superior completo;
- $q.108$: Pai possui Pós-Graduação;
- $q.109$: Não sabe a escolaridade do pai;
- $q.122$: Renda de mais de 1 e até 2 salários mínimos;
- $q.123$: Renda de mais de 2 e até 3 salários mínimos;
- $q.124$: Renda de mais de 3 e até 5 salários mínimos;
- $q.125$: Renda de mais de 5 e até 10 salários mínimos;
- $q.126$: Renda de mais de 10 e até 15 salários mínimos;
- $q.127$: Renda de mais de 15 e até 20 salários mínimos;
- $q.128$: Renda de mais de 20 salários mínimos.

Tabela 15 – Modelo de Regressão Logística m2 para Enfermagem em 2022

Variável	Estimate	Std. Error	z value	Pr(>|z|)	OR
q.104	0,600	18490,059	0,000	0,999	1,822
q.108	0,938	20083,675	0,000	0,999	2,556
q.109	0,912	18879,775	0,000	0,999	2,490
q.123	-0,707	7066,961	-0,000	0,999	0,493
q.126	0,067	12740,224	0,000	0,999	1,069
q.127	17,301	19812,323	0,001	0,999	$3,27 \times 10^7$

O modelo para o curso de Enfermagem em 2022 estima a probabilidade de aprovação dos candidatos com base na escolaridade do pai e na faixa de renda familiar. Em relação à escolaridade do pai, todas as categorias, desde o Ensino Fundamental incompleto até o Ensino Superior completo, apresentam coeficientes positivos significativos. Ter um pai com Ensino Superior incompleto ou Ensino Fundamental completo mostra forte associação positiva com a aprovação. Por outro lado, as categorias "Pós-Graduação"e "Não sabe a escolaridade do pai"também têm efeito positivo, mas de forma menos expressiva.

Quanto à renda familiar, as faixas de renda até 20 salários mínimos indicam um impacto positivo na aprovação. A faixa de renda mais alta, "mais de 20 salários mínimos", apresenta o coeficiente mais elevado, sugerindo um forte aumento nas chances de aprovação. A única exceção é a faixa de "mais de 2 e até 3 salários mínimos", que tem um coeficiente negativo, indicando uma leve redução nas chances de aprovação.

Os odds ratios (OR) demonstram o quanto cada variável influencia as chances de aprovação. Algumas variáveis, como "Renda de mais de 15 e até 20 salários mínimos"indicam um aumento extremamente alto nas chances de aprovação. No entanto, os intervalos de confiança são extremamente amplos e muitas vezes extremos, como no caso da variável "Escolaridade do pai - Ensino Médio incompleto".

4.3.2.3 2023

A base estava incompleta impossibilitando o ajuste do modelo.

Ao longo dos anos, q.10 e q.12 permaneceram como fatores-chave na aprovação dos candidatos ao curso de Enfermagem, indicando a relevância de aspectos socioeconômicos. No entanto, as mudanças nas variáveis adicionais sugerem que, enquanto em 2019 havia maior peso em histórico acadêmico e participação na economia da família (q.17 e q.18), em 2022 a estrutura do modelo favoreceu uma explicação mais direta baseada nas condições socioeconômicas dos candidatos (q.10 e q.12). Em resumo,enquanto 2019

apresentou uma variável significativa e um modelo mais estável, 2022 destacou variáveis potencialmente influentes, mas a falta de significância estatística e a alta incerteza limitaram a interpretação dos resultados.

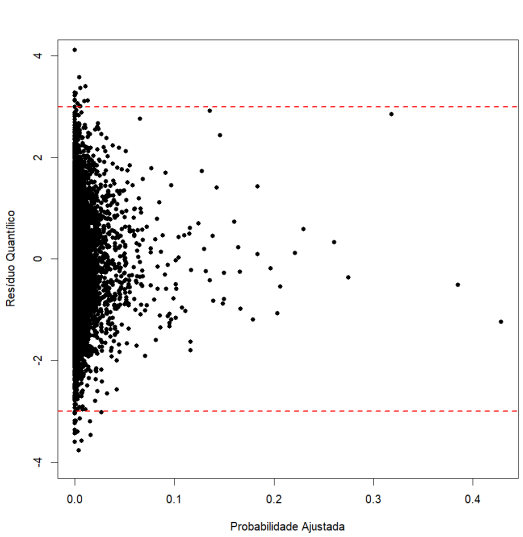
4.3.3 Medicina

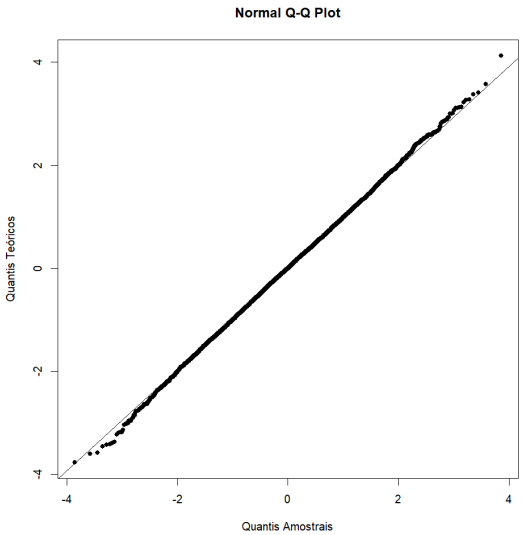
4.3.3.1 2019

Para Medicina em 2019, temos os modelos de regressão abaixo juntamente com a tabela com seus respectivos, AIC e BIC:

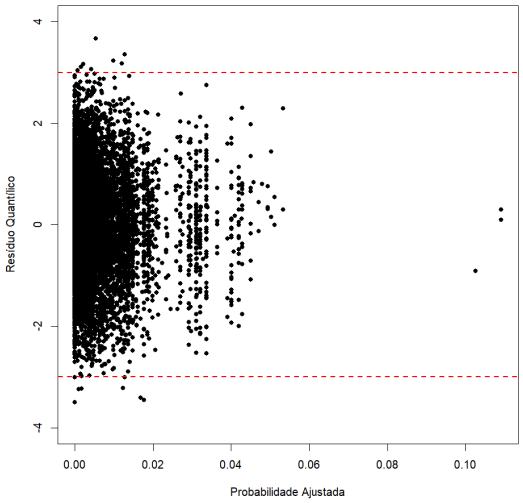
- **m1:** q.1 a q.20;
- **m2:** q.1, q.7, q.11, q.12 e q.17;
- **m3:** q.1, q.2, q.7, q.11, q.18 e q.20;
- **m4:** q.1, q.2, q.3, q.4, q.7, q.8, q.9, q.10, q.11, q.12, q.13, q.15, q.16, q.17, q.18 e q.20.
- **m5:** q.1, q.7, q.15 e q.17.

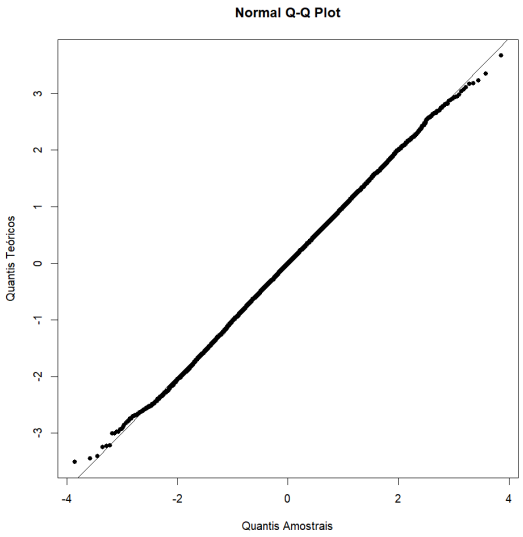
m1



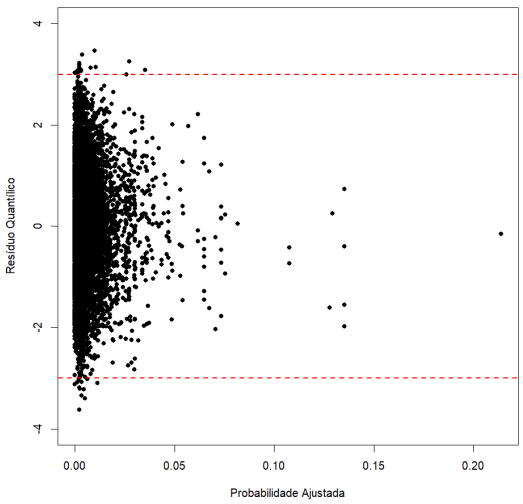


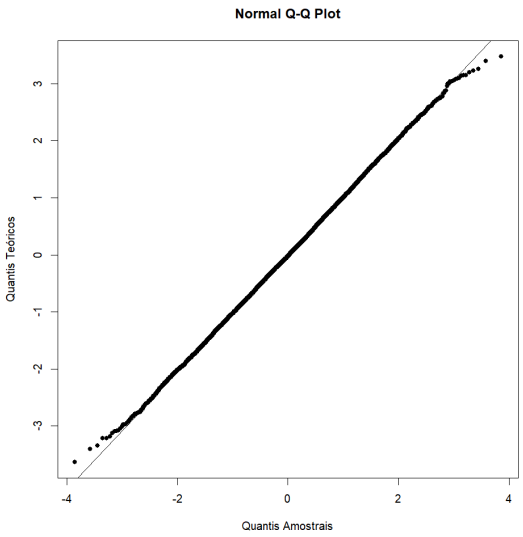
m2


Resíduo Quantílico
Probabilidade Ajustada
0.00
0.02
0.04
0.06
0.08
0.10
4
2
0
-2
-4

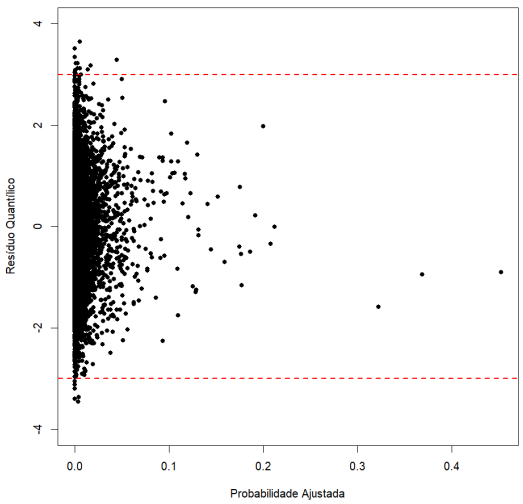

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-4
-2
0
2
4
3
2
1
0
-1
-2
-3

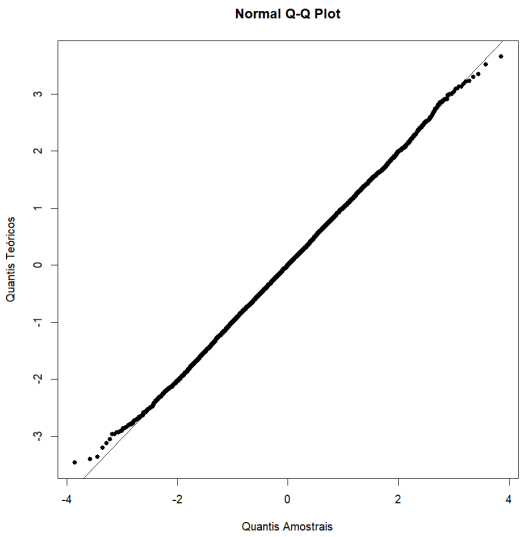
m3


Resíduo Quantílico
Probabilidade Ajustada
0.00
0.05
0.10
0.15
0.20
4
2
0
-2
-4


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-4
-2
0
2
4
3
2
1
0
-1
-2
-3

m4


Resíduo Quantílico
Probabilidade Ajustada
0.0
0.1
0.2
0.3
0.4
4
2
0
-2
-4


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais
-4
-2
0
2
4
3
2
1
0
-1
-2
-3

m5

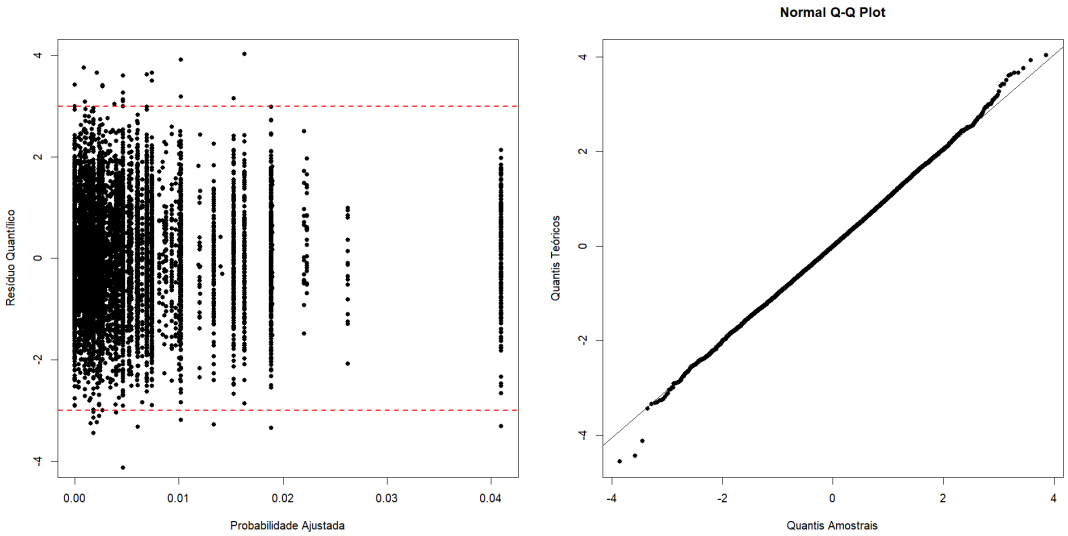


Tabela 16 – Análise gráfica dos resíduos dos modelos ajustados para Medicina em 2019.

Tabela 17 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Medicina em 2019.

Modelo	**AIC**	**BIC**
m1	706,5	1442,6
m2	632,2	837,5
m3	649,8	911,7
m4	681,8	1283,5
m5	**616,1**	**743,5**

Analisando o gráfico de resíduos dos modelos, observa-se que embora no modelo m2 há poucos valores de resíduos abaixo de -3 e acima de 3, há indício de heterocedasticidade, diferente do comportamento no modelo 5. Alem disso, m5 é o melhor modelo com base nos critérios quantitativos (AIC e BIC), pois apresenta uma boa simplicidade, considerando que utiliza apenas 4 variáveis (q.1, q.7, q.15 e q.17), enquanto mantém um ajuste adequado e o alinhamento no Q-Q Plot, indicando um ajuste mais consistente e confiável.Desta forma, dentre os modelos analisados, optamos pelo modelo m5.

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-20,356 - 0,798 \cdot q.12 - 1,013 \cdot q.72 \\ &- 1,421 \cdot q.73 - 0,627 \cdot q.74 - 16,397 \cdot q.75 \\ &- 1,148 \cdot q.76 - 16,143 \cdot q.77 - 16,191 \cdot q.78 \\ &- 2,052 \cdot q.79 + 1,509 \cdot q.172 + 0,716 \cdot q.173 \\ &+ 0,869 \cdot q.174 + 14,542 \cdot q.175\end{aligned} \tag{4.3.6}$$

- $q.12$: Sexo Feminino;
- $q.72$: Reside em outra cidade do Estado do Paraná situada na região noroeste;
- $q.73$: Reside em uma cidade do Estado do Paraná não situada na região noroeste;
- $q.74$: Reside em cidade do Estado do Santa Catarina;
- $q.75$: Reside em cidade do Estado do Rio Grande do Sul;
- $q.76$: Reside em cidade do Estado de São Paulo;
- $q.77$: Reside em cidade do Estado do Mato Grosso;
- $q.78$: Reside em cidade do Estado do Mato Grosso do Sul;
- $q.79$: Reside em cidade situada em Estado não relacionado nos itens anteriores;
- $q.172$: Realizou o Ensino Médio integralmente em escola particular;
- $q.173$: Realizou o Ensino Médio maior parte em escola pública;
- $q.174$: Realizou o Ensino Médio maior parte em escola particular;
- $q.175$: Realizou o Ensino Médio maior parte em escolas comunitárias/CNEC.

Tabela 18 – Modelo de Regressão Logística m5 para Medicina em 2019

Variável	Estimate	Std. Error	z value	Pr(>|z|)	OR	IC 95%
q.12	-0,798	0,280	-2,849	**0,004**	0,450	[0,259, 0,782]
q.72	-1,013	0,380	-2,668	**0,008**	0,363	[0,165, 0,582]
q.73	-1,421	0,393	-3,613	**0,000**	0,241	[0,106, 0,504]
q.74	-0,627	0,741	-0,846	0,398	0,534	[0,085, 1,824]
q.76	-1,148	0,459	-2,499	**0,012**	0,317	[0,117, 0,732]
q.79	-2,052	1,023	-2,006	**0,045**	0,129	[0,007, 0,378]
q.172	1,509	0,600	2,515	**0,012**	4,523	[1,639, 18,739]
q.173	0,716	1,160	0,617	0,537	2,046	[0,561, 16,160]
q.174	0,869	0,918	0,947	0,344	2,386	[0,312, 14,534]

A equação do modelo m5 indicou que ser do sexo feminino (-0,798) e residir no Estado do Rio Grande do Sul (-16,397) e Mato Grosso (-16,143), diminuem significativamente as chances de aprovação. Por outro lado, realizar o Ensino Médio integralmente em escola particular (+1,509) ou em escolas comunitárias/CNEC (+14,542) aumentou as chances de aprovação.

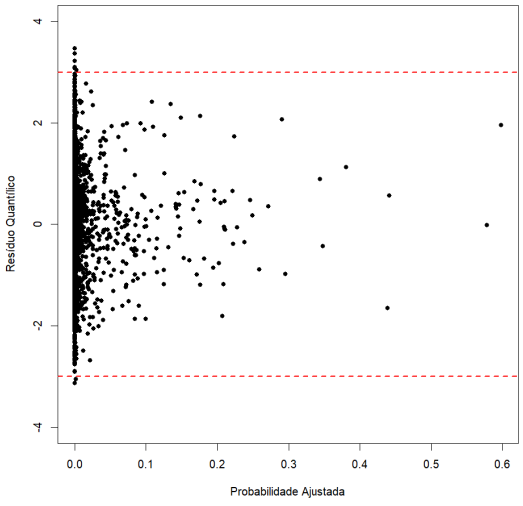
A tabela 18 mostrou que variáveis e suas categorias como "Sexo Feminino"(OR = 0,450), "Residência na região noroeste do Paraná"(OR = 0,363) e "Realizou o Ensino Médio integralmente em escola particular"(OR = 4,523) foram estatisticamente significativas. O odds ratio (OR) de 4,523 para a variável "Ensino Médio em escola particular"indica que os candidatos com essa característica tinham aproximadamente 4,5 vezes mais chances de serem aprovados em relação ao grupo de referência.

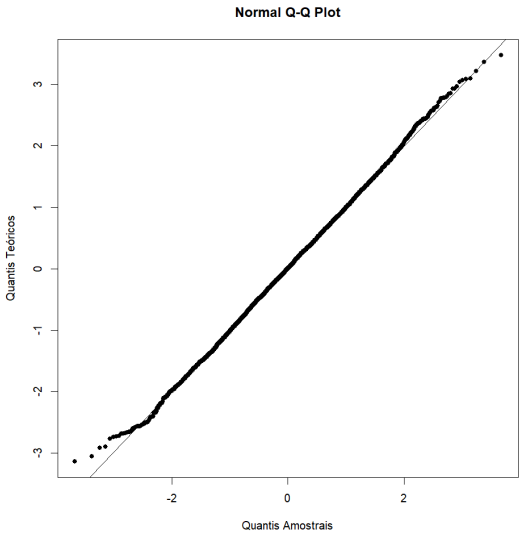
4.3.3.2 2022

Para Medicina em 2022, temos os modelos de regressão :

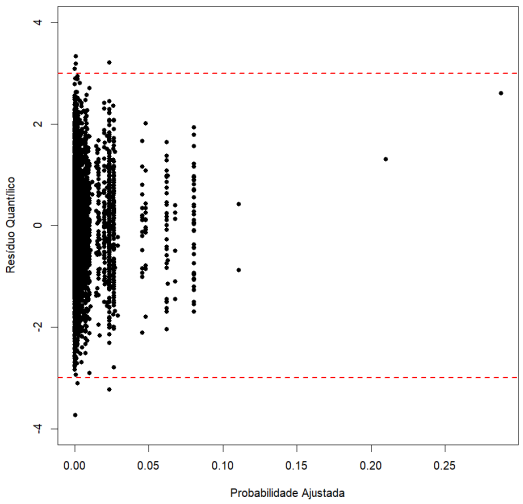
- **m1:** q.1 a q.20;
- **m2:** q.2, q.7 e q.20;
- **m3:** q.1, q.2, q.7, q.9, q.10 q.11, q.12, q.13, q.14, q.15, q.18 e q.20;
- **m4:** q.1, q.2, q.3, q.4, q.6, q.7, q.9, q.10, q.12, q.13, q.14, q.15, q.16, q.18 e q.20;
- **m5:** q.7, q.18 e q.20.

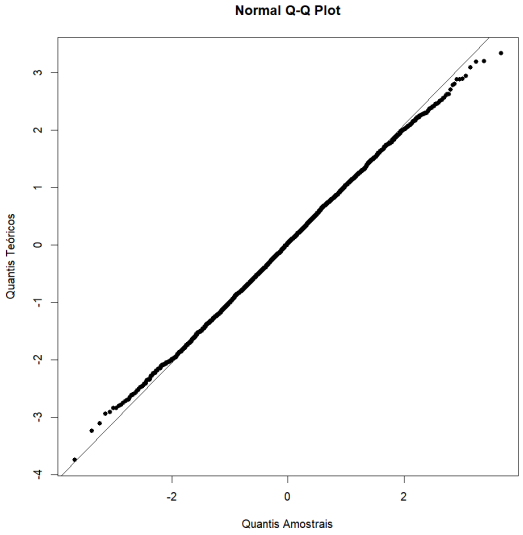
m1



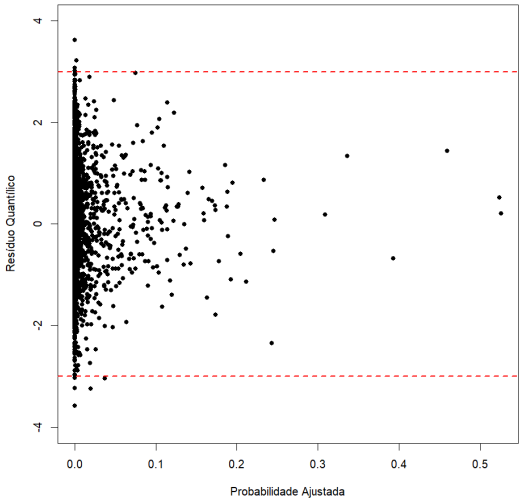


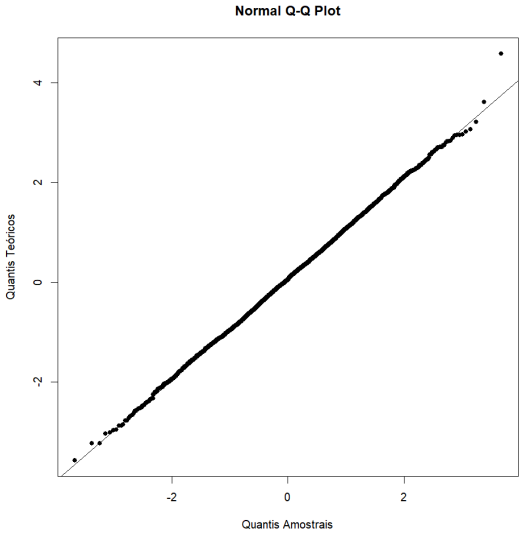
m2


Resíduo Quantílico
Probabilidade Ajustada

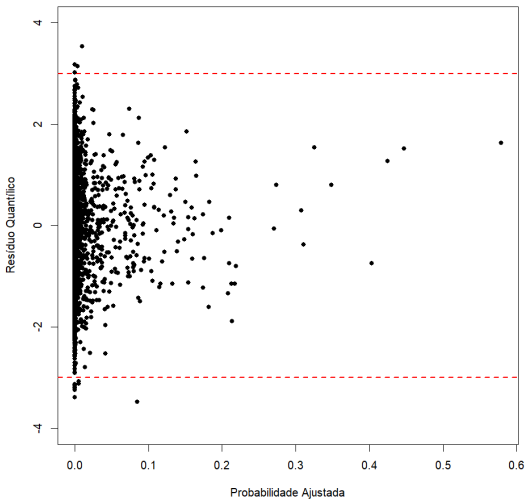

Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

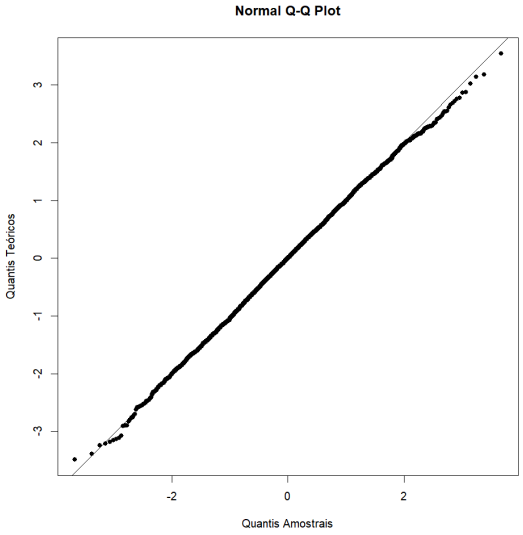
m3


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m4


Resíduo Quantílico
Probabilidade Ajustada


Normal Q-Q Plot
Quantis Teóricos
Quantis Amostrais

m5

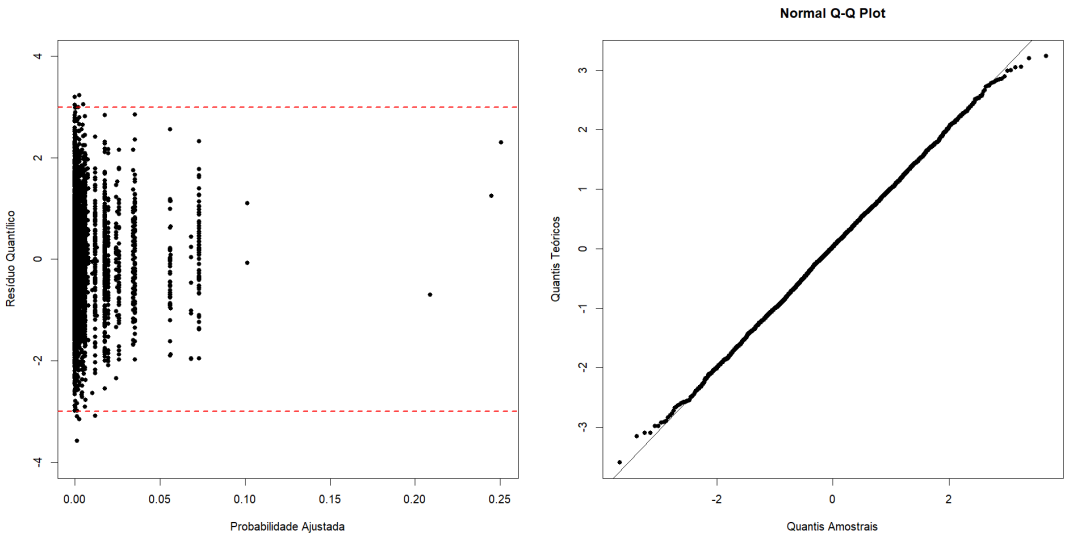


Tabela 19 – Análise gráfica dos resíduos dos modelos ajustados para Medicina em 2022.

Tabela 20 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Medicina em 2022.

Modelo	AIC	BIC
m1	382,3	1034,4
m2	298,2	437,4
m3	341,7	810,1
m4	351,6	870,7
m5	**288,7**	**409,0**

O m5 é o melhor modelo,pois apresenta os menores AIC e BIC, e possui boa simplicidade já que apenas 3 variáveis significativas, como mostra abaixo:

$$\begin{aligned}\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = &-21,552 - 1,086 \cdot q.72 - 2,595 \cdot q.73 - 17,487 \cdot q.74 \\ &- 17,409 \cdot q.75 - 2,566 \cdot q.76 - 17,727 \cdot q.77 \\ &- 17,861 \cdot q.78 - 17,576 \cdot q.79 + 0,486 \cdot q.182 \\ &- 16,387 \cdot q.183 - 0,027 \cdot q.184 - 0,448 \cdot q.185 \\ &+ 2,213 \cdot q.186 + 3,697 \cdot q.187\end{aligned} \tag{4.3.7}$$

- $q.72$: Reside em outra cidade do Estado do Paraná situada na região noroeste;
- $q.73$: Reside em uma cidade do Estado do Paraná não situada na região noroeste;
- $q.74$: Reside em cidade do Estado do Santa Catarina;

- $q.75$: Reside em cidade do Estado do Rio Grande do Sul;
- $q.76$: Reside em cidade do Estado de São Paulo;
- $q.77$: Reside em cidade do Estado do Mato Grosso;
- $q.78$: Reside em cidade do Estado do Mato Grosso do Sul;
- $q.79$: Reside em cidade situada em Estado não relacionado nos itens anteriores;
- $q.182$: Concluiu o Ensino Médio há quatro anos;
- $q.183$: Concluiu o Ensino Médio há três anos;
- $q.184$: Concluiu o Ensino Médio há dois anos;
- $q.185$: Concluiu o Ensino Médio no ano passado;
- $q.186$: Concluiu o Ensino Médio neste ano;
- $q.187$: Concluirá o Ensino Médio no próximo ano;

Tabela 21 – Modelo de Regressão Logística m5 para Medicina em 2022

Variável	Estimate	Std. Error	z value	Pr(>|z|)	OR	IC 95%
q.72	**-1,086**	**0,486**	**-2,236**	**0,025**	**0,337**	**[0,120, 0,831]**
q.73	**-2,595**	**0,769**	**-3,375**	**0,001**	**0,075**	**[0,012, 0,274]**
q.76	**-2,566**	**1,047**	**-2,450**	**0,014**	**0,077**	**[0,004, 0,394]**
q.182	0,486	0,779	0,623	0,533	1,625	[0,312, 7,582]
q.184	-0,027	0,776	-0,035	0,972	0,973	[0,188, 4,519]
q.185	-0,448	1,348	-0,332	0,740	0,639	[0,025, 7,280]
q.186	**2,213**	**0,991**	**2,232**	**0,026**	**9,141**	**[1,437, 66,755]**
q.187	**3,697**	**1,093**	**3,381**	**0,001**	**40,320**	**[5,216, 362,869]**

A equação desse modelo indicou que residir em certas regiões do Brasil, como no Estado de São Paulo (-2,566), Rio Grande do Sul (-17,409) e Santa Catarina (-17,487), diminuem consideravelmente as chances de aprovação. E candidatos que concluíram o Ensino Médio há três anos (-16,387) também tiveram menor probabilidade de aprovação. Por outro lado, os candidatos que concluíram o Ensino Médio no ano corrente (+2,213) ou que ainda concluirão no próximo ano (+3,697) tiveram chances significativamente aumentadas.

As variáveis como "Reside em cidade do Paraná fora da região noroeste"(OR = 0,075) e "Concluirá o Ensino Médio no próximo ano"(OR = 40,320) foram estatisticamente significativas ($p < 0,05$). Indicando que candidatos que ainda vão concluir o Ensino Médio têm

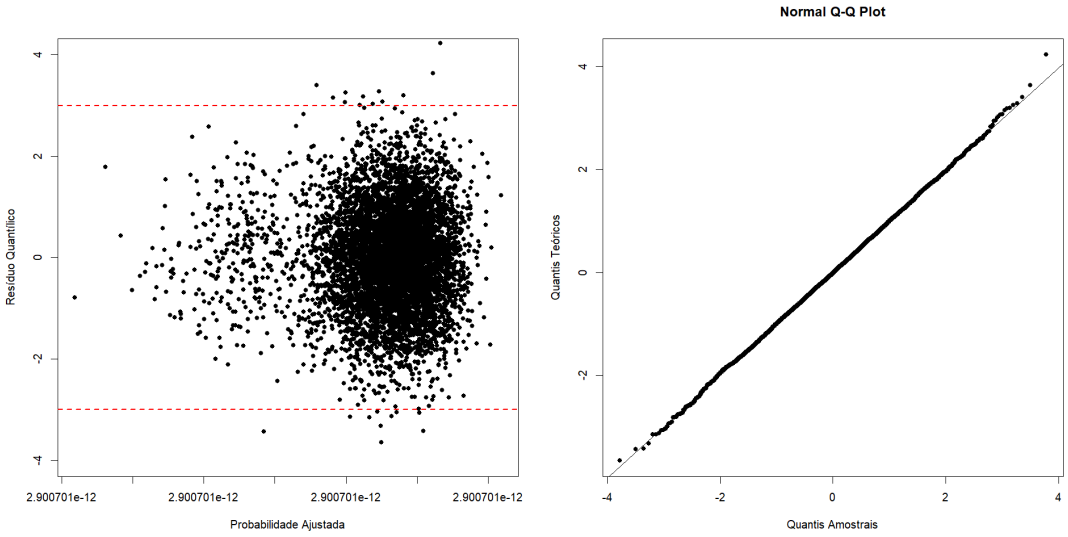
aproximadamente 40 vezes mais chances de serem aprovados em relação ao grupo de referência.

4.3.3.3 2023

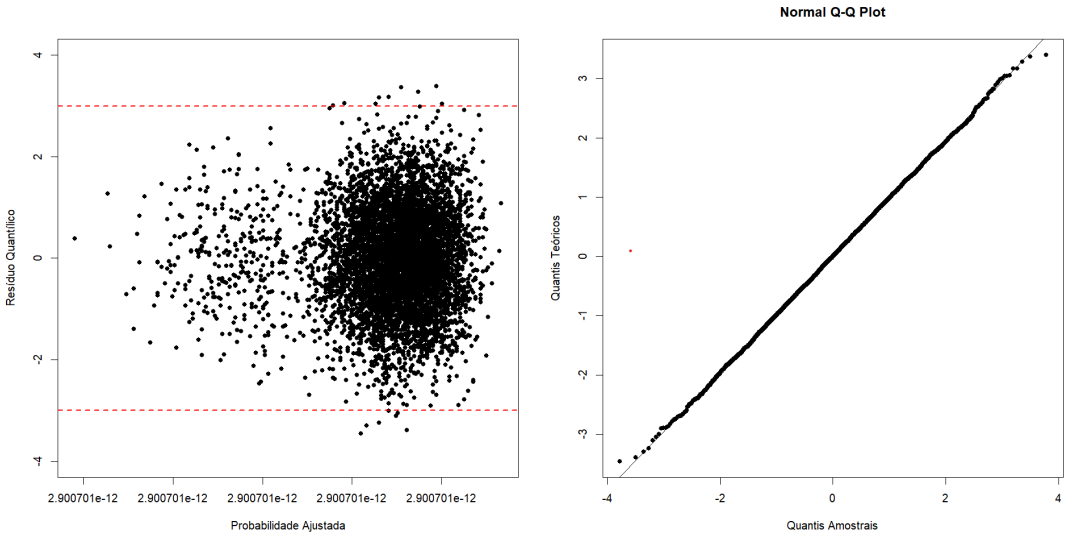
Para Medicina em 2023, temos os modelos de regressão :

- **m1:** q.1 a q.20;
- **m2:** q.1, q.2, q.3, q.4, q.5, q.6, q.8, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.16, q.17, q.19 e q.20;
- **m3:** q.1, q.2, q.7, q.9, q.10, q.11, q.12, q.13, q.14, q.15, q.16, q.17 e q.20;
- **m4:** q.1, q.2, q.3, q.4, q.7, q.8, q.9, q.11, q.12, q.13, q.14, q.15, q.16, q.17, q.18 e q.20;
- **m5:** 1.

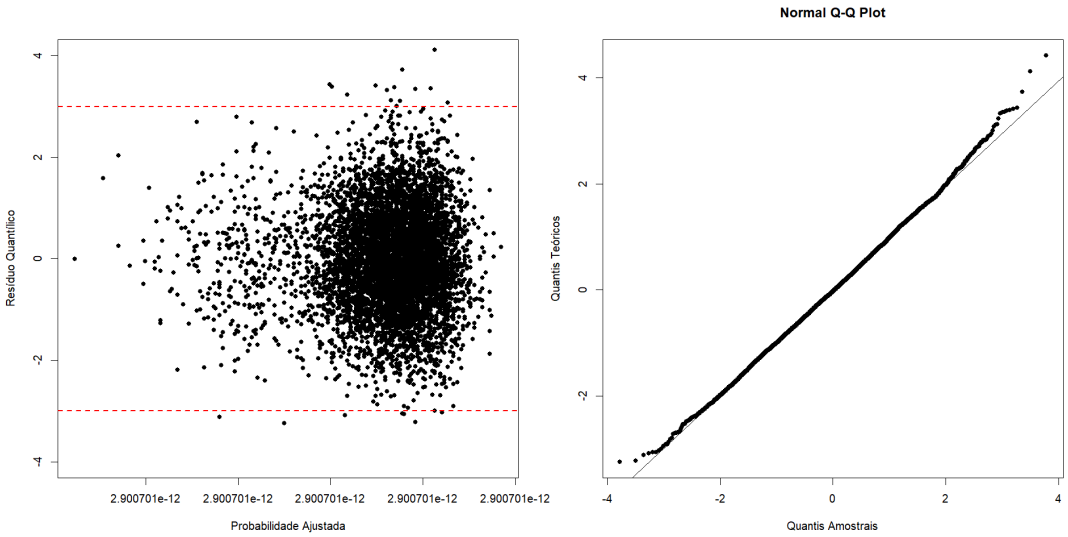
m1



m2



m3



m4

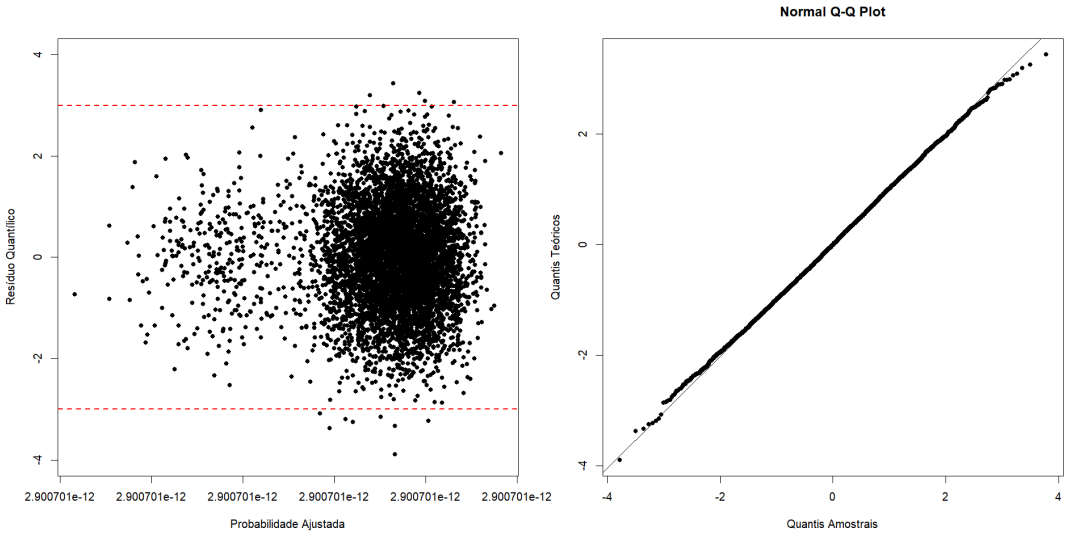


Tabela 22 – Análise gráfica dos resíduos dos modelos ajustados para Medicina em 2023.

Tabela 23 – Valores de AIC e BIC para os modelos m1, m2, m3, m4 e m5 para o curso de Medicina em 2023.

Modelo	**AIC**	**BIC**
m1	206,0	900,7
m2	178,0	778,3
m3	150,0	655,9
m4	160,0	699,6
m5	**2,0**	**8,7**

Apesar de m5 ser o melhor modelo, por apresentar menores AIC (2) e BIC (8,7), ele é um modelo nulo (apenas com o intercepto). Logo, este modelo não será considerado. Pela Tabela 23, foi selecionado m3 que apresenta os segundos menores AIC e BIC, e nos gráficos expressa um ajuste satisfatório.

O modelo m3 apresentou coeficientes estimados próximos de zero, altos erros padrão, valores de z praticamente nulos e valores p iguais a 1, indicando que nenhuma variável independente tem efeito significativo no modelo. Os odds ratios iguais a 1, com intervalos de confiança amplos, reforçam a falta de influência das variáveis na aprovação para Medicina em 2023.

Pode-se observar que presença da variável q.7 em todos os modelos finais escolhidos de 2019, 2022 e 2023 isso sugere que ela pode ser um fator determinante para a aprovação no curso de Medicina. Além disso, os modelos de 2019 e 2022 priorizaram modelos mais simples e eficientes. Em 2023, houve um aumento na complexidade, o que pode indicar mudanças no perfil dos candidatos ou na relação das variáveis com a aprovação.

De forma geral, as variáveis socioeconômicas (como renda familiar, posse de bens e necessidade de trabalhar durante o curso), acadêmicas (como o histórico de ensino médio) e geográficas (local de residência) foram fatores determinantes na aprovação dos candidatos nos três cursos. Antes da pandemia, a diversidade e o histórico escolar tinham maior influência, enquanto nos anos pós-pandemia, a estabilidade financeira ganhou destaque, reforçando o impacto das dificuldades econômicas no desempenho acadêmico. A trajetória do curso de Medicina ao longo dos anos demonstrou uma complexidade crescente nos modelos, sugerindo que as relações entre as variáveis e a aprovação dos candidatos podem ter se tornado mais complexas ao longo do tempo.

Neste capítulo, foram apresentadas a análise descritiva sobre os processos seletivos da Universidade Estadual de Maringá destacando os anos de 2019, 2020 e 2023, com enfoque tanto na descrição do perfil dos candidatos quanto na identificação das variáveis socioeducacionais determinantes para aprovação. Também foi realizada a análise de correspondência múltipla, que permitiu identificar similaridades e relações entre as variá-

veis independentes. No entanto, a análise de regressão logística mostrou a presença de heterocedasticidade, indicando inconsistências que podem comprometer a precisão das estimativas dos parâmetros do modelo, o que, por consequência, torna necessária uma nova abordagem para melhor adequação do modelo aos dados.

Capítulo 5

Conclusão

Este estudo analisou o perfil dos candidatos ao vestibular da Universidade Estadual de Maringá (UEM) no período de 2015 a 2024, com o objetivo de identificar variáveis socioeducacionais que influenciam a aprovação nos cursos de Educação Física, Enfermagem e Medicina (que ganharam destaque no cenário pós-pandêmico) com foco nos anos de 2019, 2022 e 2023, permitindo uma análise comparativa dos períodos pré-pandemia, transição e consolidação do pós-pandemia. Para isso, foram empregadas análises descritivas, Análise de Correspondência Múltipla (MCA) e Modelos de Regressão Logística.

Os resultados indicaram que variáveis como necessidade de trabalhar (q.15) e renda familiar (q.12) foram determinantes para a aprovação nos cursos de Educação Física e Enfermagem. Já no curso de Medicina, além da renda, a residência permanente do candidato (q.7) teve um peso significativo na seleção. Além disso, a formação acadêmica anterior (q.17) foi uma variável comum nos modelos de Educação Física e Medicina, sugerindo que a escolaridade prévia dos candidatos impacta diretamente na aprovação.

Os achados também revelaram diferenças no perfil dos candidatos aprovados e reprovados. A origem escolar foi um fator relevante, uma vez que candidatos que estudaram em escolas particulares tiveram maior taxa de aprovação em comparação aos egressos da rede pública. Além disso, foi identificado um aumento da complexidade dos modelos preditivos ao longo dos anos, o que pode refletir mudanças no perfil dos candidatos e na relação entre as variáveis socioeducacionais e a aprovação.

No contexto pós-pandemia, a necessidade de trabalhar durante o curso (q.15) apareceu com maior frequência, evidenciando o impacto das dificuldades financeiras no desempenho acadêmico dos candidatos. Este fator sugere que políticas de assistência estudantil podem desempenhar um papel fundamental na permanência e no sucesso dos alunos

aprovados.

Entretanto, a análise de regressão logística revelou a presença de heterocedasticidade, o que pode comprometer a precisão das estimativas dos parâmetros do modelo. Assim, torna-se necessária uma abordagem alternativa para garantir uma melhor adequação dos modelos preditivos aos dados.

Dessa forma, este estudo contribui para a compreensão do impacto das variáveis socioeducacionais na aprovação dos candidatos da UEM, fornecendo subsídios para o desenvolvimento de políticas educacionais mais equitativas e inclusivas. Os achados também fornecem subsídios para que a UEM e outras instituições possam formular ações que tornem seus processos seletivos mais inclusivos, beneficiando não apenas a comunidade acadêmica, mas a sociedade como um todo. A partir desses resultados, recomenda-se que futuras pesquisas aprofundem a análise do impacto de fatores socioeconômicos na trajetória acadêmica dos alunos, explorando novas metodologias estatísticas para aprimorar os modelos de previsão de aprovação.

Referências

ABDI, H.; VALENTIN, D. Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, p. 651–657, 2007. 31

Agência Estadual de Notícias do Paraná. *Universidade Estadual de Maringá implanta cotas sociais no vestibular*. 2009. Acesso em: 11 set. 2024. Disponível em: <https://arquivo2003.aen.pr.gov.br/Noticia/Universidade-Estadual-de-Maringa-implanta-cotas-sociais-no-vestibular>. 57

AVERSA, V. d. O.; FLORENTINO, R. *O Perfil dos Candidatos da UNESP: Uma Análise a Partir do Desempenho*. Tese (Doutorado) — Universidade Estadual Paulista, 2022.

CALIL, P. R. M. *Aplicação de técnicas multivariadas para análise dos escores de classificação dos candidatos ao concurso vestibular 2007 da UFSM*. Tese (Doutorado) — Universidade Federal de Santa Maria, Santa Maria, RS, Brasil, 2007.

CARVALHO, P. da S.; OLIVEIRA, M. S.; FIGUEIROA, M. L.; JúNIOR, L. A. de J. Probabilidade para aprovação no vestibular do curso de estatística da ufs: Uma aplicação logística binária. In: ASSOCIAçãO BRASILEIRA DE ESTATíSTICA – ABE. *XX SINAPE – Simpósio Nacional de Probabilidade e Estatística*. João Pessoa, PB, 2012.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Piracicaba: USP*, p. 31, 2008.

CORRêA, R. P.; CASTRO, H. C.; FERREIRA, R. R.; ARAúJO-JORGE, T.; STEPHENS, P. R. S. The perceptions of brazilian postgraduate students about the impact of covid-19 on their well-being and academic performance. *International Journal of Educational Research Open*, Elsevier, v. 3, p. 100185, 2022. Disponível em: <https://doi.org/10.1016/j.ijedro.2022.100185>.

DIAS, T. F.; LAGE, L. V.; RIBEIRO, R. L.; ROCHA, G. H.; RODRIGUES, J. G.; SANTOS, T. R.; FRANCO, G. C.; LOSCHI, R. H.; BRAGA, M. M. Cursos diurnos e noturnos: fatores de aprovação no vestibular da ufmg. *Cadernos de Pesquisa*, SciELO Brasil, v. 38, p. 127–146, 2008.

DOBSON, A. J. *An Introduction to Generalized Linear Models*. 2. ed. Boca Raton: Chapman & Hall/CRC, 2002. 225 p. (Texts in Statistical Science Series).

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

Fundação Oswaldo Cruz (Fiocruz). *Impactos sociais, econômicos, culturais e políticos da pandemia*. 2024. Acessado em: 17 fev. 2025. Disponível em: <https://portal.fiocruz.br/impactos-sociais-economicos-culturais-e-politicos-da-pandemia>. 19

GARCIA, F. T. Identificação de variáveis determinantes na seleção de candidatos, para os cursos de engenharia, no processo seletivo da universidade federal de santa maria, rs. Universidade Federal de Santa Maria, 2010.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: Wiley New York, 2000. 33

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3. ed. Hoboken, NJ: John Wiley & Sons, 2013. 528 p.

JOHINSON, R. A. *Applied Mutivarity Statisical Analysis*. 2007. 30

LIBERA, M. L. M. D. Avaliação do desempenho no vestibular/2003 projeção da classificação. Universidade Federal de Santa Maria, 2005.

LIMA, A. F. R.; DÍAZ, M. E. P.; JÚNIOR, S. B. F. As condições socioeconômicas e sua relação com o sucesso no vestibular: evidências a partir do processo seletivo da universidade federal de goiás. *Revista de Economia do Centro-Oeste*, v. 3, n. 1, p. 36–50, 2017.

LOPES, C. B.; RIBEIRO, R. L.; CARVALHO, M. G.; FRANCO, G. C.; LOSCHI, R. H.; BRAGA, M. M. Identificação das características associadas com a aprovação de candidatos de escolas públicas e privadas, vestibular-2004, ufmg. *Educação em Revista*, SciELO Brasil, p. 167–194, 2007.

LUCCA, R. M. *Análise da influência de fatores socioeducacionais na aprovação de candidatos nos vestibulares da UEM*. Maringá, PR, Brasil: [s.n.], 2024. Relatório Final de Atividades de Estágio Supervisionado (Bacharelado em Estatística).

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2023. Accessed: 2024-09-10. Disponível em: <https://posit.co/download/rstudio-desktop/>. 37

Secretaria da Ciência, Tecnologia e Ensino Superior - Paraná. *UEM divulga lista dos 1760 aprovados no vestibular de verão e no PAS*. 2024. Accessed: 2024-09-09. Disponível em: <https://www.seti.pr.gov.br/Noticia/UEM-divulga-lista-dos-1760-aprovados-no-vestibular-de-verao-e-no-PAS>.

SILVA, T.; PERIÇARO, G. A. Classificação dos candidatos ao vestibular da fecilcam via técnicas estatísticas multivariadas. In: *Congresso Nacional de Matemática Pura e Aplicada, Cuiabá. Disponível em: Anais do XXXII CNMAC*. [S.l.: s.n.], 2009. v. 2.

Universidade Estadual de Maringá. *UEM celebra três anos de cota racial no Dia da Consciência Negra*. 2017. Acesso em: 11 set. 2024. Disponível em: <https://noticias.uem.br/index.php?option=com_content&view=article&id=27132:

uem-celebra-tres-anos-de-cota-racial-no-dia-da-consciencia-negra&catid=986:pgina-central&Itemid=211>. 57

Universidade Estadual de Maringá. *Rankings*. 2024. Accessed: 2024-09-09. Disponível em: <https://www.uem.br/a-uem/ranking/rankings>.

Universidade Estadual de Maringá. *UEM completa 54 anos com mais de 14 mil alunos matriculados na instituição*. 2024. Accessed: 2024-09-09. Disponível em: <https://noticias.uem.br/uemnamidia/index.php/13063-uem-completa-54-anos-com-mais-de-14-mil-alunos-matriculados-na-instituicao>.

Universidade Estadual de Maringá. *Universidade Estadual de Maringá*. 2024. Acesso em: 16 ago. 2024. Disponível em: <https://www.uem.br/a-uem>.

Universidade Estadual de Maringá. *Vestibular UEM*. 2024. Accessed: 2024-09-09. Disponível em: <https://www.vestibular.uem.br/>.

ANEXO A

Script do R- pacotes utilizados

A.0.1 Pacotes utilizados

```
library(MASS)
#source("FGerais2.R")
library(caret)
library(recipes)
library(data.table)
library(descr)
library(DT)
library(mfx)
library(dplyr)
library(ROCR)
library(kableExtra)
library(Factoshiny)
library(ggplot2)
library(FactoMineR)
library(factoextra)
library(graphics)
library(ggmosaic)
library(vcd)
library(ggfortify)
library(reshape2)
library(openxlsx)
library(report)
```

```
library(stringr)
library(hnp)
library(statmod)
library(stats)
library(MASS)
library(car)
```

A.0.2 Análise Descritiva

```
setwd("C:/Users/LIAO do MU/Documents/BRUNA - MESTRADO")
dados <- read.csv("arquivo_unico.csv")

# Número total de inscritos e variáveis
inscritos <- nrow(dados)
variaveis <- ncol(dados)

names(dados)

dados$st_final <- factor(dados$st_final, levels = c("Aprovado",
↪  "Aprovado negro", "Aprovado PcD",
"Aprovado sociais", "Aprovado sociais negro",
"Classificado", "Não homologado", "Reprovado"))


dados$q.1 <- factor(dados$q.1, levels = c(1, 2))
dados$q.2 <- factor(dados$q.2, levels = 1:10)
dados$q.3 <- factor(dados$q.3, levels = 1:5)
dados$q.4 <- factor(dados$q.4, levels = 1:3)
dados$q.5 <- factor(dados$q.5, levels = 1:8)
dados$q.6 <- factor(dados$q.6, levels = 1:7)
dados$q.7 <- factor(dados$q.7, levels = 1:9)
dados$q.8 <- factor(dados$q.8, levels = 1:2)
dados$q.9 <- factor(dados$q.9, levels = 1:7)
dados$q.10 <- factor(dados$q.10, levels = 1:9)
dados$q.11 <- factor(dados$q.11, levels = 1:9)
dados$q.12 <- factor(dados$q.12, levels = 1:8)
```

```
dados$q.13 <- factor(dados$q.13, levels = 1:9)
dados$q.14 <- factor(dados$q.14, levels = 1:5)
dados$q.15 <- factor(dados$q.15, levels = 1:5)
dados$q.16 <- factor(dados$q.16, levels = 1:5)
dados$q.17 <- factor(dados$q.17, levels = 1:5)
dados$q.18 <- factor(dados$q.18, levels = 1:7)
dados$q.19 <- factor(dados$q.19, levels = 1:5)
dados$q.20 <- factor(dados$q.20, levels = 1:5)
dados$q.21 <- factor(dados$q.21, levels = 1:7)
dados$q.22 <- factor(dados$q.22, levels = 1:6)
dados$q.23 <- factor(dados$q.23, levels = 1:4)
dados$q.24 <- factor(dados$q.24, levels = 1:8)
dados$q.25 <- factor(dados$q.25, levels = 1:5)
dados$q.26 <- factor(dados$q.26, levels = 1:13)
dados$q.27 <- factor(dados$q.27, levels = 1:5)
dados$q.28 <- factor(dados$q.28, levels = 1:4)
dados$q.29 <- factor(dados$q.29, levels = 1:2)
dados$q.30 <- factor(dados$q.30, levels = 1:2)

summary(dados)

#Distribuição do número total de inscritos por Centro Acadêmico

inscritos_por_centro <- as.data.frame(table(dados$lt_centro))

colnames(inscritos_por_centro) <- c("Centro_Academico",
↪  "Inscritos")

inscritos_por_centro <-
↪  inscritos_por_centro[order(inscritos_por_centro$Inscritos,
↪  decreasing = TRUE), ]

ggplot(inscritos_por_centro, aes(x = reorder(Centro_Academico,
↪  -Inscritos), y = Inscritos)) +
geom_bar(stat = "identity", fill = "#3498db", width = 0.7) +  #
↪  Cor azul e barras finas
geom_text(aes(label = Inscritos), vjust = -0.3, color =
↪  "#2c3e50", size = 3.5) +  # Números no topo das barras
```

```
theme_minimal(base_size = 15) +  # Estilo minimalista com fontes
↪  maiores
labs(title = "Número de Inscritos por Centro Acadêmico",
subtitle = "entre os anos de 2015 e 2024",
x = "Centro Acadêmico",
y = "Número de Inscritos") +
theme(
plot.title = element_text(face = "bold", hjust = 0.5, color =
↪  "#34495e"),  # Título em negrito e centralizado
plot.subtitle = element_text(hjust = 0.5, color = "#34495e"),
axis.title.x = element_text(face = "bold", color = "#2c3e50"),  #
↪  Estilizar título do eixo X
axis.title.y = element_text(face = "bold", color = "#2c3e50"),  #
↪  Estilizar título do eixo Y
axis.text.x = element_text(angle = 45, hjust = 1),  # Rotacionar
↪  os rótulos do eixo X
panel.grid.major.x = element_blank(),  # Remover as linhas de
↪  grade verticais principais
panel.grid.minor.x = element_blank()
)


#Distribuição do número total de inscritos por Centro Acadêmico a
↪  cada ano


inscritos_por_ano_centro <- as.data.frame(table(dados$nu_ano,
↪  dados$lt_centro))

colnames(inscritos_por_ano_centro) <- c("Ano",
↪  "Centro_Academico", "Inscritos")

inscritos_por_ano_centro$Ano <-
↪  as.numeric(as.character(inscritos_por_ano_centro$Ano))

ggplot(inscritos_por_ano_centro, aes(x = Ano, y = Inscritos,
↪  color = Centro_Academico, group = Centro_Academico)) +
geom_line(size = 1) +  # Adicionar as linhas
```

```
geom_point(size = 2) +  # Adicionar os pontos nas linhas
theme_minimal(base_size = 15) +  # Estilo minimalista com fontes
↪  maiores
labs(title = "Número de Inscritos por Centro Acadêmico",
subtitle = "De 2015 a 2024",
x = "Ano",
y = "Número de Inscritos",
color = "Centro Acadêmico") +
scale_x_continuous(breaks =
↪  seq(min(inscritos_por_ano_centro$Ano),
↪  max(inscritos_por_ano_centro$Ano), by = 1)) +  # Mostrar
↪  todos os anos
theme(
plot.title = element_text(face = "bold", hjust = 0.5, color =
↪  "#34495e"),  # Título em negrito e centralizado
plot.subtitle = element_text(hjust = 0.5, color = "#34495e"),  #
↪  Centralizar o subtítulo
axis.title.x = element_text(face = "bold", color = "#2c3e50"),  #
↪  Estilizar título do eixo X
axis.title.y = element_text(face = "bold", color = "#2c3e50"),  #
↪  Estilizar título do eixo Y
axis.text.x = element_text(angle = 45, hjust = 1),  # Rotacionar
↪  os rótulos do eixo X
panel.grid.major.x = element_blank(),  # Remover as linhas de
↪  grade verticais principais
panel.grid.minor.x = element_blank()
)

inscritos_por_ano_centro <- dados %>%
group_by(nu_ano, lt_centro) %>%  # Agrupar por ano e centro
↪  acadêmico
summarise(Inscritos = n()) %>%    # Contar o número de inscritos
ungroup()

total_inscritos_por_ano <- inscritos_por_ano_centro %>%
group_by(nu_ano) %>%
summarise(Total_Inscritos = sum(Inscritos)) %>%
ungroup()
```

```
inscritos_por_ano_centro <- inscritos_por_ano_centro %>%
left_join(total_inscritos_por_ano, by = "nu_ano") %>%
mutate(Percentual = (Inscritos / Total_Inscritos) * 100)  #
↪  Calcular a porcentagem

ggplot(inscritos_por_ano_centro, aes(x = nu_ano, y = Percentual,
↪  color = lt_centro, group = lt_centro)) +
geom_line(size = 1) +  # Adicionar as linhas
geom_point(size = 2) +  # Adicionar pontos para as observações
theme_minimal(base_size = 15) +  # Usar um tema minimalista
labs(title = "Porcentagem de Inscritos por Centro Acadêmico ao
↪  Longo dos Anos",
x = "Ano",
y = "Porcentagem de Inscritos (%)",
color = "Centro Acadêmico") +
scale_y_continuous(labels = scales::percent_format(scale = 1),
↪  limits = c(0, 60)) +
scale_x_continuous(breaks =
↪  seq(min(inscritos_por_ano_centro$nu_ano),
↪  max(inscritos_por_ano_centro$nu_ano), by = 1)) +  # Mostrar
↪  todos os anos no eixo X
theme(
plot.title = element_text(face = "bold", hjust = 0.5),  #
↪  Centralizar e deixar em negrito o título
axis.title.x = element_text(face = "bold"),  # Negrito para o
↪  título do eixo X
axis.title.y = element_text(face = "bold"),  # Negrito para o
↪  título do eixo Y
axis.text.x = element_text(angle = 45, hjust = 1),  # Rotacionar
↪  os rótulos do eixo X
panel.grid.major.x = element_blank(),  # Remover as linhas de
↪  grade verticais principais
panel.grid.minor.x = element_blank(),
panel.grid.major = element_line(color = "#bdc3c7"),  # Cor das
↪  grades principais
panel.grid.minor = element_blank(),
legend.title = element_blank()
```

```
)

# Análise de perfil dos inscritos

for (i in 1:30) {
        var_name <- paste0("q.", i)
        if (var_name %in% colnames(dados)) {
                freq_table <- table(dados[[var_name]])
                max_value <- which.max(freq_table)
                cat("Variável:", var_name, "\n")
                print(freq_table)
                cat("Valor mais frequente:", names(max_value),
                ↪  "com frequência", freq_table[max_value],
                ↪  "\n\n")
        } else {
                cat("A variável", var_name, "não existe no
                ↪  dataframe.\n\n")
        }
}

for (i in 1:30) {
        var_name <- paste0("q.", i)
        if (var_name %in% colnames(dados)) {
                freq_table <- table(dados[[var_name]])
                max_value <- which.max(freq_table)
                max_freq <- freq_table[max_value]
                total_obs <- sum(freq_table)  # Total de
                ↪  observações
                percentage <- (max_freq / total_obs) * 100  #
                ↪  Cálculo da porcentagem

                cat("Variável:", var_name, "\n")
                print(freq_table)
                cat("Valor mais frequente:", names(max_value),
                ↪  "com frequência", max_freq,
                "(", round(percentage, 2), "%)", "\n\n")
        } else {
```

```
                cat("A variável", var_name, "não existe no
                ↪  dataframe.\n\n")
        }
}


#Após realizar uma triagem de dados, reduzimos o banco de dados
↪  as categorias de aprovação(Aprovado, Aprovado negro, Aprovado
↪  PcD, Aprovado sociais, Aprovado sociais negro) para observar
↪  quais categorias são mais frequentes entre as variáveis no
↪  banco de dados "Aprovação" e através dessa frequência
↪  estabelecer um perfil para os candidatos aprovados:

dados_aprovacao <- dados %>%
filter(!(st_final %in% c("Não homologado", "Reprovado",
↪  "Classificado")))

dados_aprovacao <- dados_aprovacao %>%
group_by(st_final) %>%
filter(n() > 0)


for (i in 1:30) {
        var_name <- paste0("q.", i)
        if (var_name %in% colnames(dados_aprovacao)) {
                freq_table <- table(dados_aprovacao[[var_name]])
                max_value <- which.max(freq_table)
                cat("Variável:", var_name, "\n")
                print(freq_table)
                cat("Valor mais frequente:", names(max_value),
                ↪  "com frequência", freq_table[max_value],
                ↪  "\n\n")
        } else {
                cat("A variável", var_name, "não existe no
                ↪  dataframe.\n\n")
        }
}
```

```
# Total de inscritos (supondo que seja o número de linhas do
↪ dataframe)
total_inscritos <- nrow(dados)

calcular_probabilidade_frequencia <- function(dados_completos,
↪ dados_aprovacao, variavel) {
        tab_completa <-
        ↪ as.data.frame(table(dados_completos[[variavel]]))
        colnames(tab_completa) <- c("Categoria", "Total")
        tab_aprovacao <-
        ↪ as.data.frame(table(dados_aprovacao[[variavel]]))
        colnames(tab_aprovacao) <- c("Categoria", "Aprovados")
        tab_resultado <- merge(tab_completa, tab_aprovacao, by =
        ↪ "Categoria", all.x = TRUE)
        tab_resultado$Aprovados[is.na(tab_resultado$Aprovados)]
        ↪ <- 0
        tab_resultado$Probabilidade_Aprovacao <-
        ↪ tab_resultado$Aprovados / tab_resultado$Total

        return(tab_resultado)
}

# Laço de repetição de q.1 a q.30 para calcular e imprimir as
↪ probabilidades
for (i in 1:30) {
        variavel <- paste0("q.", i)  # Criar o nome da variável
        ↪ dinamicamente (q.1, q.2, ..., q.30)

        # Verificar se a variável existe no dataframe
        if (variavel %in% colnames(dados)) {
                # Calcular a probabilidade usando tabelas de
                ↪ frequência para a variável atual
                probabilidade <-
                ↪ calcular_probabilidade_frequencia(dados,
                ↪ dados_aprovacao, variavel)

                # Imprimir os resultados diretamente
                cat("\nProbabilidades para", variavel, ":\n")
```

```
                print(probabilidade)
        } else {
                cat("A variável", variavel, "não existe no
                ↪  dataframe.\n")
        }
}

## Correlação entre as variáveis

#Criando uma nova coluna chamada aprovação

dados$aprovacao <- ifelse(dados$st_final %in% c("Aprovado",
↪  "Aprovado negro", "Aprovado PcD",
"Aprovado sociais", "Aprovado sociais negro"),
1, 0)

print(head(dados))

## Teste qui-quadrado

O teste qui-quadrado (chisq.test) foi utilizado para testar a
↪  existência de uma associação significativa entre as variáveis
↪  do questionário socioeducacional e a coluna nova criada
↪  aprovação.

# Transformar a coluna "aprovação" em fator, se ainda não for
dados$aprovacao <- factor(dados$aprovacao)

# Transformar as colunas Q1 a Q30 em fatores, se necessário
dados[, paste0("q.", 1:30)] <- lapply(dados[, paste0("q.",
↪  1:30)], factor)


# Loop para realizar o chisq.test para Q1 a Q30

resultados <- list()

for (i in 1:30) {
```

```
        # Nome da coluna
        coluna <- paste0("q.", i)

        # Remover NAs das duas colunas antes de criar a tabela de
        ↪ contingência
        dados_sem_na <- na.omit(dados[, c("aprovacao", coluna)])

        # Criar a tabela de contingência
        tabela <- table(dados_sem_na$aprovacao,
        ↪ dados_sem_na[[coluna]])

        # Realizar o teste qui-quadrado
        resultado <- chisq.test(tabela)

        # Armazenar o resultado do teste
        resultados[[coluna]] <- resultado
}

# Verificar os resultados
resultados


## Cálculo do V de Cramer


# Criar uma lista para armazenar os resultados de V de Cramer
resultados_v_cramer <- list()

for (i in 1:30) {
        # Nome da coluna (Q1, Q2, ..., Q30)
        coluna <- paste0("q.", i)

        # Criar a tabela de contingência, excluindo NAs
        tabela <- table(dados$aprovacao, dados[[coluna]], useNA =
        ↪ "no")

        # Verificar se há valores suficientes para o cálculo
        if (all(tabela > 0)) {
```

```
                # Calcular o V de Cramer
                v_cramer <- assocstats(tabela)$cramer

                # Armazenar o resultado do V de Cramer na lista
                resultados_v_cramer[[coluna]] <- v_cramer
        } else {
                # Caso haja problemas com a tabela (ex.:
                ↪ frequências insuficientes)
                resultados_v_cramer[[coluna]] <- "Valores
                ↪ insuficientes"
        }
}

# Exibir os resultados de V de Cramer para cada variável Q
resultados_v_cramer

# Inicializar uma lista para armazenar as tabelas de dupla
↪ entrada
tabelas_dupla_entrada <- list()

# Loop para calcular as tabelas de dupla entrada para Q1 a Q30 em
↪ relação à "aprovação"
for (i in 1:30) {
        # Nome da variável (Q1, Q2, ..., Q30)
        coluna <- paste0("q.", i)

        # Criar a tabela de dupla entrada entre "aprovação" e a
        ↪ variável Q
        tabela <- table(dados[[coluna]], dados$aprovacao)

        # Armazenar a tabela de dupla entrada na lista
        tabelas_dupla_entrada[[coluna]] <- tabela

        # Exibir a tabela de dupla entrada
        print(paste("Tabela de dupla entrada para", coluna))
        print(tabela)

        # Calcular a frequência relativa global
```

```
        frequencia_relativa_global <- prop.table(tabela)
        print(paste("Frequência relativa global para", coluna))
        print(frequencia_relativa_global)

        # Calcular a frequência relativa por linha (dentro de
        ↪ cada grupo de q.1 a q.30)
        frequencia_relativa_linha <- prop.table(tabela, margin =
        ↪ 1)
        print(paste("Frequência relativa por linha para",
        ↪ coluna))
        print(frequencia_relativa_linha)

        # Calcular a frequência relativa por coluna (dentro de
        ↪ cada grupo de aprovação)
        frequencia_relativa_coluna <- prop.table(tabela, margin =
        ↪ 2)
        print(paste("Frequência relativa por coluna para",
        ↪ coluna))
        print(frequencia_relativa_coluna)
}
```

A.0.3 Análise Multivariada

```
        \begin{Verbatim}

                dados <- read.csv("arquivo_unico.csv");dados

                str(dados)

                summary(dados)

                inscritos <- nrow(dados)
                inscritos

                variaveis <- ncol(dados)
```

```
variaveis

names(dados)

dados$st_final <- factor(dados$st_final, levels =
↪  c("Aprovado", "Aprovado negro", "Aprovado
↪  PcD",
"Aprovado sociais", "Aprovado sociais negro",
"Classificado", "Não homologado", "Reprovado"))


dados$aprovacao <- ifelse(dados$st_final %in%
↪  c("Aprovado", "Aprovado negro", "Aprovado
↪  PcD",
"Aprovado sociais", "Aprovado sociais negro"),
1, 0)

dados$q.1 <- factor(dados$q.1, levels = c(1, 2))
dados$q.2 <- factor(dados$q.2, levels = 1:10)
dados$q.3 <- factor(dados$q.3, levels = 1:5)
dados$q.4 <- factor(dados$q.4, levels = 1:3)
dados$q.5 <- factor(dados$q.5, levels = 1:8)
dados$q.6 <- factor(dados$q.6, levels = 1:7)
dados$q.7 <- factor(dados$q.7, levels = 1:9)
dados$q.8 <- factor(dados$q.8, levels = 1:2)
dados$q.9 <- factor(dados$q.9, levels = 1:7)
dados$q.10 <- factor(dados$q.10, levels = 1:9)
dados$q.11 <- factor(dados$q.11, levels = 1:9)
dados$q.12 <- factor(dados$q.12, levels = 1:8)
dados$q.13 <- factor(dados$q.13, levels = 1:9)
dados$q.14 <- factor(dados$q.14, levels = 1:5)
dados$q.15 <- factor(dados$q.15, levels = 1:5)
dados$q.16 <- factor(dados$q.16, levels = 1:5)
dados$q.17 <- factor(dados$q.17, levels = 1:5)
dados$q.18 <- factor(dados$q.18, levels = 1:7)
dados$q.19 <- factor(dados$q.19, levels = 1:5)
dados$q.20 <- factor(dados$q.20, levels = 1:5)
dados$q.21 <- factor(dados$q.21, levels = 1:7)
```

```
dados$q.22 <- factor(dados$q.22, levels = 1:6)
dados$q.23 <- factor(dados$q.23, levels = 1:4)
dados$q.24 <- factor(dados$q.24, levels = 1:8)
dados$q.25 <- factor(dados$q.25, levels = 1:5)
dados$q.26 <- factor(dados$q.26, levels = 1:13)
dados$q.27 <- factor(dados$q.27, levels = 1:5)
dados$q.28 <- factor(dados$q.28, levels = 1:4)
dados$q.29 <- factor(dados$q.29, levels = 1:2)
dados$q.30 <- factor(dados$q.30, levels = 1:2)
dados$aprovacao <- factor(dados$aprovacao, levels
↪  = 0:1)


dados <- dados %>%
mutate(tipo_evento = case_when(
grepl("PAS-UEM",  nm_evento) ~ "PAS-UEM",
grepl("Vestibular de Verão", nm_evento) ~
↪  "Vestibular de Verão",
grepl("Vestibular de Inverno", nm_evento) ~
↪  "Vestibular de Inverno",
grepl("Vestibular", nm_evento) ~ "Vestibular",
))

summary(dados)

inscritos <- nrow(dados)
inscritos

variaveis <- ncol(dados)
variaveis

str(dados)

names(dados)

reglog <- dados %>%
filter(nu_ano %in% 2015:2023, nu_opcao_cotas ==
↪  "Não cotista")
```

```
        reg <- reglog[, c(1:26,37)]; reg

        table(reg$nu_ano)

        reg <- na.omit(reg); reg

        summary(reg)

        nrow(reg)

        ncol(reg)

        str(reg)

        names(reg)

# MCA - Multiple Correspondence Analysis in R: Essentials

res.mca <- MCA(reg, graph = FALSE)
print(res.mca)

eig.val <- get_eigenvalue(res.mca)
print(eig.val)

# The percentages of inertia explained by each MCA
↪  dimensions

fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0,
↪  15))


# The biplot of individuals and variable categories

#fviz_mca_biplot(res.mca, repel = TRUE, ggtheme =
↪  theme_minimal())
```

```
# Graph of variables
# Results
var <- get_mca_var(res.mca); var

# Coordinates
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)

# Correlation between variables and principal dimensions

fviz_mca_var(res.mca, choice = "mca.cor", repel = TRUE,
↪  ggtheme = theme_minimal())

#fviz_mca_var(res.mca, axes = c(1, 3), choice =
↪  "mca.cor", repel = TRUE, ggtheme = theme_minimal())

#fviz_mca_var(res.mca, axes = c(1,2), choice = "mca.cor",
↪  repel = TRUE, ggtheme = theme_minimal())


#fviz_mca_var(res.mca, axes = c(2, 5), choice =
↪  "mca.cor", repel = TRUE, ggtheme = theme_minimal())


# Coordinates of variable categories

fviz_mca_var(res.mca,repel = TRUE,ggtheme =
↪  theme_minimal())


# Quality of representation of variable categories

fviz_mca_var(res.mca, col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())
```

```
fviz_mca_var(res.mca, axes = c(1,3), col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())



fviz_mca_var(res.mca, axes = c(2,3), col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())


fviz_mca_var(res.mca, alpha.var="cos2", repel = TRUE,
↪  ggtheme = theme_minimal())

fviz_mca_var(res.mca, axes = c(1,3), alpha.var="cos2",
↪  repel = TRUE, ggtheme = theme_minimal())

fviz_mca_var(res.mca, axes = c(2,3), alpha.var="cos2",
↪  repel = TRUE, ggtheme = theme_minimal())

# Cos2 of variable categories on Dim.1 and Dim.2

fviz_cos2(res.mca, choice = "var", axes = 1:2)



head(round(var$contrib,2), 4)

# Contributions of rows to dimension 1


fviz_contrib(res.mca, choice = "var", axes = 1, top = 15)


# Contributions of rows to dimension 2
```

```
fviz_contrib(res.mca, choice = "var", axes = 2, top = 15)


# Contributions of rows to dimension 3


fviz_contrib(res.mca, choice = "var", axes = 3, top = 15)


# Total contribution to dimension 1 and 2

fviz_contrib(res.mca, choice = "var", axes = 1:2, top =
↪  15)


# Total contribution to dimension 1 and 3

fviz_contrib(res.mca, choice = "var", axes = 1:3, top =
↪  15)

# Total contribution to dimension 2 and 3

fviz_contrib(res.mca, choice = "var", axes = 2:3, top =
↪  15)


# The most important variable categories highlighted on
↪  the scatter plot'

fviz_mca_var(res.mca, col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())



fviz_mca_var(res.mca,axes = c(1,3), col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())
```

```
fviz_mca_var(res.mca, axes = c(2,3), col.var = "contrib",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())


# Graph of individuals

ind <- get_mca_ind(res.mca)
ind

# Coordinates of column points
head(ind$coord)
# Quality of representation
head(ind$cos2)
# Contributions
head(ind$contrib)

# Plots: quality and contribution

fviz_mca_ind(res.mca, col.ind = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE, ggtheme = theme_minimal())

# Cos2 of individuals
fviz_cos2(res.mca, choice = "ind", axes = 1:2, top = 20)

# Contribution of individuals to the dimensions
fviz_contrib(res.mca, choice = "ind", axes = 1:2, top =
↪  20)

# Color individuals by groups

fviz_mca_ind(res.mca,
label = "none", habillage = "Q1",
palette = c("#00AFBB", "#E7B800"),
addEllipses = TRUE, ellipse.type = "confidence",
```

```
ggtheme = theme_minimal())

fviz_mca_ind(res.mca,
label = "none", habillage = "Q2",
palette = c("#00A", "#E7B"),
addEllipses = TRUE, ellipse.type = "confidence",
ggtheme = theme_minimal())


fviz_mca_ind(res.mca,
label = "none", habillage = "Q3",
palette = c("#FF3800", "#FFFF00"),
addEllipses = TRUE, ellipse.type = "confidence",
ggtheme = theme_minimal())

# habillage = index of the column to be used as grouping
↪  variable
fviz_mca_ind(res.mca, habillage = 2, addEllipses = TRUE)

# habillage = external grouping variable
fviz_mca_ind(res.mca, habillage = X$Q1, addEllipses =
↪  TRUE)

fviz_ellipses(res.mca, c("Q2", "Q3"),
geom = "point")

fviz_ellipses(res.mca, 1:4, geom = "point")

fviz_ellipses(res.mca, 5:8, geom = "point")

fviz_ellipses(res.mca, 9:12, geom = "point")

fviz_ellipses(res.mca, 13:16, geom = "point")

fviz_ellipses(res.mca, 17:20, geom = "point")

# Biplot of individuals and variable categories
fviz_mca_biplot(res.mca, repel = TRUE,
```

```
ggtheme = theme_minimal())

fviz_mca_var(res.mca, choice = "mca.cor",
repel = TRUE)


enf2019mca <- enf2019[ , !(names(enf2019) %in%
↪  c("aprovacao"))]

res.mca <- MCA(enf2019mca, graph = FALSE)
print(res.mca)

eig.val <- get_eigenvalue(res.mca)
print(eig.val)

# The percentages of inertia explainenf by each MCA
↪  dimensions

fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0,
↪  15))


# The biplot of individuals and variable categories

#fviz_mca_biplot(res.mca, repel = TRUE, ggtheme =
↪  theme_minimal())


# Graph of variables
# Results
var <- get_mca_var(res.mca); var

# Coordinates
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)
```

```
# Correlation between variables and principal dimensions

fviz_mca_var(res.mca, choice = "mca.cor", repel = TRUE,
↪  ggtheme = theme_minimal())



#### ENFERMAGEM 2022 ####


ed2022mca <- enf2022[ , !(names(enf2022) %in%
↪  c("aprovacao"))]

res.mca <- MCA(ed2022mca, graph = FALSE)
print(res.mca)

eig.val <- get_eigenvalue(res.mca)
print(eig.val)

# The percentages of inertia explained by each MCA
↪  dimensions

fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0,
↪  15))


# The biplot of individuals and variable categories

#fviz_mca_biplot(res.mca, repel = TRUE, ggtheme =
↪  theme_minimal())


# Graph of variables
# Results
var <- get_mca_var(res.mca); var

# Coordinates
```

```
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)

# Correlation between variables and principal dimensions

fviz_mca_var(res.mca, choice = "mca.cor", repel = TRUE,
↪  ggtheme = theme_minimal())


#### ENFERMAGEM 2023 ####

ed2023mca <- enf2023[ , !(names(enf2023) %in%
↪  c("aprovacao"))]

res.mca <- MCA(ed2023mca, graph = FALSE)
print(res.mca)

eig.val <- get_eigenvalue(res.mca)
print(eig.val)

# The percentages of inertia explained by each MCA
↪  dimensions

fviz_screeplot(res.mca, addlabels = TRUE, ylim = c(0,
↪  15))


# The biplot of individuals and variable categories

#fviz_mca_biplot(res.mca, repel = TRUE, ggtheme =
↪  theme_minimal())


# Graph of variables
# Results
```

```
var <- get_mca_var(res.mca); var

# Coordinates
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)

# Correlation between variables and principal dimensions

fviz_mca_var(res.mca, choice = "mca.cor", repel = TRUE,
↪  ggtheme = theme_minimal())
```

A.0.4 Modelos de Regressão Logística

```
med2019 <-  reg %>%
filter( nu_ano == 2019, grepl("(Medicina)", nm_curso),
↪  nu_opcao_cotas == "Não cotista")

med2019 <- med2019[ , !(names(med2019) %in%
↪  c("nu_ano","nm_evento","st_final","lt_centro","nm_cu⌋
↪  rso","nu_opcao_cotas"))]


##### Ajuste de modelo de MLG com todas as variaveis
↪  #####

m1 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.5+ q.6+ q.7+
↪  q.8+ q.9+ q.10 +q.11 +q.12 +q.13 +q.14 +q.15 +q.16
↪  +q.17 +q.18+q.19+q.20, data = med2019, family =
↪  "binomial"(link = "logit"))
summary(m1)

### RESIDUOS
```

```
par(mfrow = c(1,2))

resid(m1)

resid(m1, type = "pearson")

qqnorm(resid(m1), pch = 20, cex = 1.5)
qqline(resid(m1))


#Gráficos de Resíduos vs. Valores Ajustados

residuos = resid(m1)
ajustados <- predict(m1)
plot(ajustados,residuos, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.

hnp(m1, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos <- qresid(m1)
ajustados <- predict(m1, type="response")

plot(residuos ~ ajustados, pch = 20, cex = 1.4, xlab =
↪  "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.
```

```
qqnorm(residuos, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos)


anova(m1,test = "Chisq")


#####Ajuste de modelo de MLG com as variaveis
↪  significativas ####

m2 <- glm(aprovacao ~ q.1 + q.7 + q.11 + q.12 + q.17,
↪  data = med2019, family = "binomial"(link = "logit"))
summary(m2)

### RESIDUOS

resid(m2)

resid(m2, type = "pearson")

qqnorm(resid(m2), pch = 20, cex = 1.5)
qqline(resid(m2))


#Gráficos de Resíduos vs. Valores Ajustados

residuos2 = resid(m2)
ajustados2 <- predict(m2)
plot(ajustados2,residuos2, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.

hnp(m2, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")
```

```
### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos21 <- qresid(m2)
ajustados21 <- predict(m2, type = "response")


plot(residuos21 ~ ajustados21, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))

abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)

### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos21, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos21)


##### Ajuste de modelo de MLG com as variaveis
↪  explicativas entre aprovados e não aprovados #####

m3 <- glm(aprovacao ~ q.1 +  q.2 + q.7 + q.11 + q.18 +
↪  q.20, data = med2019, family = "binomial"(link =
↪  "logit"))
summary(m3)

### RESIDUOS

resid(m3)
resid(m3, type = "pearson")

qqnorm(resid(m3), pch = 20, cex = 1.5)
qqline(resid(m3))

#Gráficos de Resíduos vs. Valores Ajustados
```

```
residuos3 = resid(m3)
ajustados3 <- predict(m3)
plot(ajustados3,residuos3, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m3, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos31 <- qresid(m3)
ajustados31 <- predict(m3,type="response")

plot(residuos31 ~ ajustados31, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos31, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos31)


#####Ajuste de modelo de MLG com as variaveis
↪  explicativas com MCA ####

# As variáveis similares são: q1 e q5, q6 e q8, q10 e
↪  q14, e q15 e q19
```

```
m4 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.7+ q.8+ q.9+
↪  q.10 +q.11 +q.12 +q.13 + q.15 +q.16 +q.17 +q.18+q.20,
↪  data = med2019, family = "binomial"(link = "logit"))
summary(m4)

### RESIDUOS

resid(m4)

resid(m4, type = "pearson")

qqnorm(resid(m4), pch = 20, cex = 1.5)
qqline(resid(m4))


#Gráficos de Resíduos vs. Valores Ajustados

residuos4 = resid(m4)
ajustados4 <- predict(m4)
plot(ajustados4,residuos4, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m4, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos41 <- qresid(m4)
ajustados41 <- predict(m4, type = "response")


plot(residuos41 ~ ajustados41, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
```

```
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos41, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos41)


#####Ajuste de M1 de MLG através do stepAIC ####

m5 <- stepAIC(m1)
summary(m5)

### RESIDUOS

resid(m5)
resid(m5, type = "pearson")

qqnorm(resid(m5), pch = 20, cex = 1.5)
qqline(resid(m5))

#Gráficos de Resíduos vs. Valores Ajustados

residuos5 = resid(m5)
ajustados5 <- predict(m5)
plot(ajustados5,residuos5, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m5, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
```

```
residuos51 <- qresid(m5)
ajustados51 <- predict(m5, type = "response")

plot(residuos51 ~ ajustados51, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos51, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos51)


edAIC <- AIC(m1,m2,m3,m4,m5)

edBIC <- BIC(m1,m2,m3,m4,m5)

table_df <- data.frame(
Model = rownames(edAIC),
AIC = edAIC$AIC,
BIC = edBIC$BIC
)

table_df

# O Modelo 5 (m5) é o melhor modelo com base nos
↪  critérios quantitativos (AIC e BIC), pois apresenta:
# Uma boa simplicidade, considerando que utiliza apenas 4
↪  variáveis (q.1, q.7, q.15, e q.17), enquanto mantém
↪  um ajuste adequado.
# Se a simplicidade e interpretabilidade são prioridades,
↪  o Modelo 5 (m5) é a melhor escolha. Se você deseja um
↪  modelo mais abrangente,
```

```
# o Modelo 4 (m4) poderia ser considerado, mas com
↪  penalização por complexidade.


##### ODDS RATIO #####

summary(m5)$coefficients

odds_ratios <- exp(coef(m5))
print(odds_ratios)

# Intervalo de confiança para o OR
odds_ratios_ci <- exp(cbind(OR = coef(m5), confint(m5)))
print(odds_ratios_ci)


#### MEDICINA 2022 ####

med2022 <-  reg %>%
filter( nu_ano == 2022, grepl("(Medicina)", nm_curso),
↪  nu_opcao_cotas == "Não cotista")

med2022 <- med2022[ , !(names(med2022) %in%
↪  c("nu_ano","nm_evento","st_final","lt_centro","nm_cu
↪  rso","nu_opcao_cotas"))]


##### Ajuste de modelo de MLG com todas as variaveis
↪  #####

m1 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.5+ q.6+ q.7+
↪  q.8+ q.9+ q.10 +q.11 +q.12 +q.13 +q.14 +q.15 +q.16
↪  +q.17 +q.18+q.19+q.20, data = med2022, family =
↪  "binomial"(link = "logit"))
summary(m1)

### RESIDUOS
```

```
resid(m1)
resid(m1, type = "pearson")

qqnorm(resid(m1), pch = 20, cex = 1.5)
qqline(resid(m1))


#Gráficos de Resíduos vs. Valores Ajustados

residuos = resid(m1)
ajustados <- predict(m1)
plot(ajustados,residuos, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.

hnp(m1, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos <- qresid(m1)
ajustados <- predict(m1, type="response")

plot(residuos ~ ajustados, pch = 20, cex = 1.4, xlab =
↪  "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos)
```

```
anova(m1,test = "Chisq")


#####Ajuste de modelo de MLG com as variaveis
↪  significativas ####

m2 <- glm(aprovacao ~ q.2 + q.7 + q.20 , data = med2022,
↪  family = "binomial"(link = "logit"))
summary(m2)

### RESIDUOS

resid(m2)
resid(m2, type = "pearson")

qqnorm(resid(m2), pch = 20, cex = 1.5)
qqline(resid(m2))


#Gráficos de Resíduos vs. Valores Ajustados

residuos2 = resid(m2)
ajustados2 <- predict(m2)
plot(ajustados2,residuos2, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.

hnp(m2, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")


### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos21 <- qresid(m2)
```

```
ajustados21 <- predict(m2, type = "response")


plot(residuos21 ~ ajustados21, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))

abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)

### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos21, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos21)


##### Ajuste de modelo de MLG com as variaveis
↪  explicativas entre aprovados e não aprovados #####

m3 <- glm(aprovacao ~ q.1 +  q.2 + q.7 + q.9 + q.10 +
↪  q.11 + q.12 + q.13 + q.14 + q.15 + q.18 + q.20, data
↪  = med2022, family = "binomial"(link = "logit"))
summary(m3)

### RESIDUOS

resid(m3)
resid(m3, type = "pearson")

qqnorm(resid(m3), pch = 20, cex = 1.5)
qqline(resid(m3))

#Gráficos de Resíduos vs. Valores Ajustados

residuos3 = resid(m3)
ajustados3 <- predict(m3)
plot(ajustados3,residuos3, pch = 20, cex = 1.4)
```

```
### Gráfico de resíduos versus valores ajustados.

hnp(m3, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos31 <- qresid(m3)
ajustados31 <- predict(m3,type="response")

plot(residuos31 ~ ajustados31, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos31, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos31)


#####Ajuste de modelo de MLG com as variaveis
↪  explicativas com MCA ####

# As variáveis similares são: q1, q5 e q8, q10, q11 e
↪  q17, q13 e q19

m4 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.6+ q.7+ q.9+
↪  q.10 +q.12 +q.13 +q.14 +q.15 +q.16 + q.18 + q.20,
↪  data = med2022, family = "binomial"(link = "logit"))
summary(m4)

### RESIDUOS
```

```
resid(m4)

resid(m4, type = "pearson")

qqnorm(resid(m4), pch = 20, cex = 1.5)
qqline(resid(m4))


#Gráficos de Resíduos vs. Valores Ajustados

residuos4 = resid(m4)
ajustados4 <- predict(m4)
plot(ajustados4,residuos4, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m4, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos41 <- qresid(m4)
ajustados41 <- predict(m4, type = "response")


plot(residuos41 ~ ajustados41, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.
```

```
qqnorm(residuos41, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos41)


#####Ajuste de M1 de MLG através do stepAIC ####

m5 <- stepAIC(m1)
summary(m5)

### RESIDUOS

resid(m5)
resid(m5, type = "pearson")

qqnorm(resid(m5), pch = 20, cex = 1.5)
qqline(resid(m5))

#Gráficos de Resíduos vs. Valores Ajustados

residuos5 = resid(m5)
ajustados5 <- predict(m5)
plot(ajustados5,residuos5, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m5, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos51 <- qresid(m5)
ajustados51 <- predict(m5, type = "response")
```

```
plot(residuos51 ~ ajustados51, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos51, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos51)

##### AIC, BIC e ANOVA #####

edAIC <- AIC(m1,m2,m3,m4,m5)

edBIC <- BIC(m1,m2,m3,m4,m5)

table_df <- data.frame(
Model = rownames(edAIC),
AIC = edAIC$AIC,
BIC = edBIC$BIC
)

table_df

anova(m1,m2,m3,m4,m5)

# O Modelo 5 (m5) é o melhor modelo porque apresenta: O
↪  menor AIC (288.75) e BIC (409.03).
# Boa simplicidade, com apenas 3 variáveis
↪  significativas.

##### ODDS RATIO #####

summary(m5)$coefficients
```

```
# Intervalo de confiança para o OR
odds_ratios_ci <- exp(cbind(OR = coef(m5), confint(m5)))
print(odds_ratios_ci)



#### MEDICINA 2023 ####

med2023 <-  reg %>%
filter( nu_ano == 2023, grepl("(Medicina)", nm_curso),
↪  nu_opcao_cotas == "Não cotista")

med2023 <- med2023[ , !(names(med2023) %in%
↪  c("nu_ano","nm_evento","st_final","lt_centro","nm_cu
↪  rso","nu_opcao_cotas"))]


##### Ajuste de modelo de MLG com todas as variaveis
↪  #####

m1 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.5+ q.6+ q.7+
↪  q.8+ q.9+ q.10 +q.11 +q.12 +q.13 +q.14 +q.15 +q.16
↪  +q.17 +q.18+q.19+q.20, data = med2023, family =
↪  "binomial"(link = "logit"))
summary(m1)

### RESIDUOS

resid(m1)
resid(m1, type = "pearson")

qqnorm(resid(m1), pch = 20, cex = 1.5)
qqline(resid(m1))

# Gráficos de Resíduos vs. Valores Ajustados

residuos = resid(m1)
ajustados <- predict(m1)
```

```
plot(ajustados,residuos, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.

hnp(m1, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos <- qresid(m1)
ajustados <- predict(m1, type="response")

plot(residuos ~ ajustados, pch = 20, cex = 1.4, xlab =
↪  "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos)


anova(m1,test = "Chisq")

# Comentário: nenhuma das variáveis no modelo é
↪  estatisticamente significativa
# para prever a variável resposta aprovacao, dado o
↪  conjunto de dados.
```

```
##########################################################
↪  ####################

# Inicializar uma matriz para armazenar os coeficientes
↪  de Cramér

num_variaveis <- 20
matriz_cramer <- matrix(NA, nrow = num_variaveis, ncol =
↪  num_variaveis)
colnames(matriz_cramer) <- paste0("q.", 1:num_variaveis)
rownames(matriz_cramer) <- paste0("q.", 1:num_variaveis)

# Loop para calcular o coeficiente de Cramér entre todas
↪  as combinações de variáveis
for (i in 1:num_variaveis) {
        for (j in i:num_variaveis) {
                # Criar a tabela de contingência para as
                ↪  variáveis q.i e q.j
                tabela_contingencia <-
                ↪  table(med2023[[paste0("q.", i)]],
                ↪  med2023[[paste0("q.", j)]])

                # Calcular o coeficiente de Cramér
                resultado <-
                ↪  assocstats(tabela_contingencia)

                # Armazenar o valor na matriz
                matriz_cramer[i, j] <- resultado$cramer
                matriz_cramer[j, i] <- resultado$cramer
                ↪  # A matriz é simétrica
        }
}


# Substituir NaN por zero (ou outro valor, se desejar)
matriz_cramer[is.nan(matriz_cramer)] <- 0

# Visualizar a matriz de coeficientes de Cramér
```

```
matriz_cramer

# Selecionar os pares de variáveis com maiores
↪ associações
limite_associacao <- 0.5  # Defina um limite para
↪ considerar alta associação
pares_altas_associacoes <- which(matriz_cramer >
↪ limite_associacao, arr.ind = TRUE)

if (nrow(pares_altas_associacoes) > 0) {
        # Filtrar para manter apenas as associações não
        ↪ redundantes (acima da diagonal da matriz)
        pares_altas_associacoes <- pares_altas_associaco
        ↪ es[pares_altas_associacoes[, 1] <
        ↪ pares_altas_associacoes[, 2], , drop = FALSE]

        # Criar um data frame dos pares com alta
        ↪ associação
        resultado_pares <- data.frame(
        Variavel1 = rownames(matriz_cramer)[pares_altas_
        ↪ associacoes[, 1]],
        Variavel2 = colnames(matriz_cramer)[pares_altas_
        ↪ associacoes[, 2]],
        Associacao =
        ↪ matriz_cramer[pares_altas_associacoes]
        )

        # Exibir os resultados
        print(resultado_pares)
} else {
        # Caso não existam pares acima do limite
        cat("Nenhuma associação foi encontrada acima do
        ↪ limite definido.\n")
}

# q.6 e q.7  0.6042309
# q.2 e q.18  0.7591645
```

```
##########################################################
↪  ##################
vif(m1)

# Variáveis que apresentam multicolinearidade alta: não
↪  há


#####Ajuste de modelo de MLG com as variaveis
↪  significativas ####

#obs: como não há variáveis significativas somente foram
↪  excuidas as variaveis coom correlação e VIF altos.

m2 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.5+ q.6+ q.8+
↪  q.9+ q.10 +q.11 +q.12 +q.13 +q.14 +q.15 +q.16 +q.17
↪  +q.19+q.20, data = med2023, family = "binomial"(link
↪  = "logit"))
summary(m2)

### RESIDUOS

resid(m2)
resid(m2, type = "pearson")

qqnorm(resid(m2), pch = 20, cex = 1.5)
qqline(resid(m2))


#Gráficos de Resíduos vs. Valores Ajustados

residuos2 = resid(m2)
ajustados2 <- predict(m2)
plot(ajustados2,residuos2, pch = 20, cex = 1.4)


### Gráfico de resíduos versus valores ajustados.
```

```
hnp(m2, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")


### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos21 <- qresid(m2)
ajustados21 <- predict(m2, type = "response")


plot(residuos21 ~ ajustados21, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))

abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)

### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos21, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos21)


##### Ajuste de modelo de MLG com as variaveis
↪  explicativas entre aprovados e não aprovados #####

m3 <- glm(aprovacao ~ q.1 + q.2 + q.7 + q.9 + q.10 + q.11
↪  + q.12 + q.13 + q.14 + q.15 + q.16 + q.17 + q.20,
↪  data = med2023, family = "binomial"(link = "logit"))
summary(m3)

### RESIDUOS

resid(m3)
resid(m3, type = "pearson")
```

```
qqnorm(resid(m3), pch = 20, cex = 1.5)
qqline(resid(m3))

#Gráficos de Resíduos vs. Valores Ajustados

residuos3 = resid(m3)
ajustados3 <- predict(m3)
plot(ajustados3,residuos3, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m3, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))
residuos31 <- qresid(m3)
ajustados31 <- predict(m3,type="response")

plot(residuos31 ~ ajustados31, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos31, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos31)


#####Ajuste de modelo de MLG com as variaveis
↪  explicativas com MCA ####
```

```
# As variáveis similares são: q1 e q5, q6 e q8, q10 e
↪  q12, q15 e q19

m4 <- glm(aprovacao ~ q.1+ q.2+ q.3+ q.4+ q.7+ q.8+
↪  q.9+q.11 +q.12 +q.13 +q.14 +q.15+q.16 +q.17
↪  +q.18+q.20, data = med2023, family = "binomial"(link
↪  = "logit"))
summary(m4)

### RESIDUOS

resid(m4)

resid(m4, type = "pearson")

qqnorm(resid(m4), pch = 20, cex = 1.5)
qqline(resid(m4))


#Gráficos de Resíduos vs. Valores Ajustados

residuos4 = resid(m4)
ajustados4 <- predict(m4)
plot(ajustados4,residuos4, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m4, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")

### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos41 <- qresid(m4)
ajustados41 <- predict(m4, type = "response")
```

```
plot(residuos41 ~ ajustados41, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos41, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos41)


#####Ajuste de M1 de MLG através do stepAIC ####

m5 <- stepAIC(m1)
summary(m5)

### RESIDUOS

resid(m5)
resid(m5, type = "pearson")

qqnorm(resid(m5), pch = 20, cex = 1.5)
qqline(resid(m5))

#Gráficos de Resíduos vs. Valores Ajustados

residuos5 = resid(m5)
ajustados5 <- predict(m5)
plot(ajustados5,residuos5, pch = 20, cex = 1.4)

### Gráfico de resíduos versus valores ajustados.

hnp(m5, pch = 20, cex = 1.2, halfnormal = FALSE, xlab =
↪  "Resíduos", ylab = "Quantis Teóricos")
```

```
### Para os resíduos quantílicos aleatorizados, temos:

par(mfrow = c(1,2))

residuos51 <- qresid(m5)
ajustados51 <- predict(m5, type = "response")

plot(residuos51 ~ ajustados51, pch = 20, cex = 1.4, xlab
↪  = "Probabilidade Ajustada", ylab = "Resíduo
↪  Quantílico", ylim = c(-4,4))
abline(h = 3, col = "red", lty = 2, lwd = 2)
abline(h = -3, col = "red", lty = 2, lwd = 2)


### Gráfico de resíduos versus valores ajustados.

qqnorm(residuos51, pch = 20, cex = 1.4, xlab = "Quantis
↪  Amostrais", ylab = "Quantis Teóricos")
qqline(residuos51)

##### AIC, BIC e ANOVA #####

edAIC <- AIC(m1,m2,m3,m4,m5)

edBIC <- BIC(m1,m2,m3,m4,m5)

table_df <- data.frame(
Model = rownames(edAIC),
AIC = edAIC$AIC,
BIC = edBIC$BIC
)

table_df

# O Modelo 5 (m5) é o melhor modelo. Apesar de ser um
↪  modelo nulo (apenas com o intercepto),
# ele apresenta: O menor AIC (2) e BIC (8.75).
```

```
##### ODDS RATIO #####

summary(m3)$coefficients

# Intervalo de confiança para o OR
odds_ratios_ci <- exp(cbind(OR = coef(m3), confint(m3)))
print(odds_ratios_ci)
```