Universidade de Brasília -UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Estudo dos fatores associados ao ingresso de estudantes da Universidade de Brasília no curso não desejado: Uma aplicação de regressão logística.

Shayane dos Santos Cordeiro

Orientadora: Prof[a]. Maria Teresa Leão Costa

Brasília
Maio 2021

Shayane dos Santos Cordeiro

Estudo dos fatores associados ao ingresso de estudantes da Universidade de Brasília no curso não desejado: Uma aplicação de regressão logística.

Orientadora: Prof[a]. Maria Teresa Leão Costa

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

Brasília

Maio 2021

AGRADECIMENTOS

Primeiramente agradeço a Deus por me conceder as faculdades necessárias para concretizar mais uma etapa da minha jornada e a mãe de todos os seres, Nossa Senhora, que sempre intercedeu por mim durante as minhas tormentas. Agradeço também:

- A minha mãe, Maria Vanêz dos Santos, por sua sensibilidade e compreensão com a minha constante ausência. Ao meu pai, José Geraldo Campos Cordeiro, pelos bons momentos na minha infância para eu me recordar sempre. Aos meus irmãos em especial Janaina e Samuel pelo amor incondicional.

- Ao meu companheiro, Luiz Gustavo Furlan, pelo apoio nessa segunda jornada, por sua paciência e ombro amigo, por compartilhar aventuras, problemas e momentos felizes comigo.

- A minha orientadora, Maria Teresa Leão Costa, por seu conhecimento compartilhado, humildade ao querer sempre aprender com seus discentes, paciência para ouvi-los e por ter acolhido o meu tema para pesquisa.

- As professoras Ana Maria Nogales Vasconcelos por sua força e motivação com a pesquisa, Juliana Betini Fachini Gomes por me inspirar ao mestrado acadêmico e a Cátia Regina Gonçalves por semear a vontade de estudar estatística em mim.

- Aos colegas de graduação pela generosidade em compartilhar conhecimento, materiais e boas conversas. Todos ficamos mais fortes durante nossa caminhada e aprendemos muito.

- Por fim, a todos os professores e pesquisadores que lutam com paixão por uma educação inclusiva e igualitária, que abdicam de momentos com familiares e amigos para se dedicarem ao conhecimento e a constante busca pela solução de problemas. Meu muito obrigada.

"Você nasceu no lar que precisava nascer,
vestiu o corpo físico que merecia,
mora onde melhor Deus te proporcionou,
de acordo com o teu adiantamento.
Seu ambiente de trabalho é o que você elegeu
espontaneamente para a sua realização.
Teus parentes e amigos são as almas que você mesmo atraiu,
com tua própria afinidade.
Você escolhe, recolhe, elege, atrai, busca, expulsa,
modifica tudo aquilo que te rodeia a existência.
Não reclame, nem se faça de vítima.
Antes de tudo, analisa e observa.
A mudança está em tuas mãos.
Reprograma tua meta, busca o bem e você viverá melhor.
Embora ninguém possa voltar atrás e fazer um novo começo,
qualquer um pode começar agora e fazer um novo fim."
(Chico Xavier)

RESUMO

Conhecer o perfil dos ingressantes do ensino superior é fundamental para subsidiar políticas para acesso e permanência do corpo discente e assegurar sua posterior formação. Nesse sentido, este trabalho teve como objetivo analisar o perfil dos ingressantes da Universidade de Brasília, durante os anos de 2012-2017, utilizando a base de dados da pesquisa "Perfil dos Estudantes da Universidade de Brasília - Etapa Registro" conduzida pelo Observatório da Vida Estudantil. Sendo verificadas mudanças principalmente nas variáveis que compõem aspectos econômicos e vinculados a trajetória de ensino, escolhas e perspectivas do ingressante. Dentre esses aspectos estudados o elevado número de estudantes que não ingressam no curso desejado chamou a atenção. Estudos apontam para um percentual significativo de desistência entre os estudantes que não ingressaram no curso desejado, como consequência torna-se crucial analisar fatores que podem ter peso no processo de escolha do curso pelo estudante. Foi proposto um modelo de regressão logística para estudar quais fatores podem ter relação com esse alto percentual de não ingressantes no curso desejado por sua natureza binária. Após verificar a significância foram definidas as seguintes variáveis para compor o modelo selecionado: ano e modalidade de ingresso, campus, curso por prestígio, sexo, estado civil, renda mensal, nível de escolaridade dos pais, tipo de instituição que fez o ensino médio, se realizou curso preparatório, número de tentativas, trocaria de curso, perspectiva profissional e idade.

Palavras Chaves: Ensino Superior, Universidade de Brasília, Perfil dos estudantes, Ingresso no Curso não Desejado, Regressão logística.

LISTA DE ABREVIATURAS E SIGLAS

CEAM	Centro de Estudos Multidisciplinares
DF	Distrito Federal
OVE	Observatório da Vida Estudantil
UnB	Universidade de Brasília

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Para muitos a oportunidade de ingressar em um curso superior é a garantia de uma melhor remuneração, colocação dentro do mercado de trabalho ou até mesmo seguir uma carreira sonhada. É através dela que ocorre o crescimento pessoal e profissional e a redução nas desigualdades sociais. Isto pode ser verificado através de estudos que mostram que a educação desempenha um papel fundamental para o crescimento econômico e social do país bem como para aqueles que investem na sua formação superior (BARROS et al,2001; WILLMS, 1997). Indo ao encontro desses estudos, SEVERINO et al (2008), destaca o papel do ensino superior não apenas como poderoso mecanismo de ascensão social, mas também sua responsabilidade na construção de uma sociedade marcada pela cidadania, vida coletiva e democracia, cabendo destacar o papel das universidades públicas para a formação de profissionais qualificados. Como consequência, nos últimos anos foi observado um aumento expressivo no número de alunos matriculados na educação superior, além do aumento no número de instituições de ensino superior.

Conhecer o corpo discente das instituições de Ensino Superior é fundamental para a implementação de políticas educacionais voltadas para garantir não apenas o acesso do aluno, mas também sua permanência e conclusão do curso. Controlados os fatores internos que podem provocar a evasão do estudante, principalmente através de políticas educacionais que reduzam as desigualdades e o auxiliem no prosseguimento do curso oferecendo os subsídios necessários para a permanência do ingressante, resta os externos que estão relacionados a escolhas prévias do estudante antes do ingresso no curso.

Os fatores que influenciam a escolha do curso pelo estudante podem estar ligados a diferentes causas como vocação, gosto pessoal, mercado de trabalho, realização pessoal entre outras, porém nem todos os alunos conseguem ingressar no curso desejado e muitas vezes acabam mudando a opção por concorrência/falta de preparo ou tentativas anteriores sem sucesso, optando por cursos de menor prestígio. Outro ponto a ser considerado é a falta de maturidade para a escolha do curso, em muitos casos os estudantes sentem-se pressionados por familiares e terceiros para após o término do ensino médio iniciar imediatamente a graduação. Sem ter uma escolha pessoal definida acabam escolhendo cursos principalmente pelo mercado de trabalho, influência familiar e notoriedade.

O ingresso no curso não ambicionado pode ser um dos fatores de evasão desse estudante, gerando muitas vezes um ciclo onde ele começa um curso e não termina. Estudos como os de BARDAGI (2007), o qual evidencia a influência do

desenvolvimento vocacional na evasão do Ensino Superior através de entrevistas com os evadidos e CUNHA et al (2001), que identifica as causas manifestadas por esse grupo para a desistência do curso de Química da Universidade de Brasília, reforçando a falta de orientação ao estudante quanto a escolha: 47,8% dos evadidos tinham Química como o curso de preferência, 39,1% apontaram que Química realmente não era o curso da preferência, 13,0% admitiram ter dúvida quanto à preferência. O estudo de DIAS et al (2010), que através de um levantamento da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros (Unimontes-MG), constata que cerca de 63,2% dos respondentes que abandonaram a graduação indicaram ter cometido falhas no momento de decidir o curso, entre esse percentual os principais motivos foram: escolha do curso como segunda opção (32,4%), por incentivo ou influência da família ou de amigos (21,8%), em vista da baixa concorrência (19%) e a falta de orientação vocacional (16%). Tais estudos apontam para a necessidade de conhecer fatores que podem estar associados ao ingresso do estudante no curso não desejado.

Investir na educação é crucial para um país que busca o desenvolvimento. O problema de não entrar no curso desejado pode ter diversas consequências para a educação superior, principalmente a pública, pois cada aluno que ingressa no ensino superior é aguardado no mercado de trabalho para movimentar a economia e retornar o investimento da sociedade. O outro lado é o do aluno que espera ser um profissional reconhecido, bem remunerado e acima de tudo exercer a profissão desejada.

Importante ressaltar que o processo de seleção das instituições públicas são muitas vezes árduos principalmente para alunos que não tiveram tantas oportunidades na educação básica. Mesmo com a implementação de políticas afirmativas na Educação Superior esse processo ainda não está bem nivelado tendo muitos que abrir mão de um curso sonhado por falta de oportunidades que não tiveram ou pela frustração de tentar uma vaga em um curso concorrido e não conseguir.

Desse modo, surge a necessidade de compreender os fatores que levam o aluno a não entrar no curso desejado. Variáveis referentes ao perfil do estudante podem ajudar a esclarecer quais fatores possivelmente estariam associados a essa mudança na escolha do curso e consequentemente a entrada em um curso não almejado.

Diante do exposto, este estudo tem por objetivo investigar se características sociodemográficas, econômicas e relativas ao ingresso do estudante na Universidade de Brasília estão associadas a não escolha do curso almejado pelo aluno, utilizando um *modelo de regressão logística*.

2 Referencial Teórico

Neste capítulo será desenvolvido a parte teórica da técnica que será utilizada.

2.1 Tabelas de contingência

Dados de variáveis qualitativas são comumente apresentados por tabelas de contingência, cujas células ou caselas, contém o número de observações na categoria analisada. A tabela que representa duas variáveis (X e Y) é chamada de tabela de contingência bivariada $l \times c$, em que l representa as linhas (categorias da variável X) e c representa as colunas (categorias da variável Y). Já a tabela que representa três variáveis é chamada de tabela de contingência trivariada $l \times c \times k$, l são as linhas, c as colunas e k camadas ou estratos. Pode-se expandir esse conceito para tabelas com mais variáveis conhecidas, como tabelas de contingência multivariadas.

Para exemplificar alguns conceitos considere a Tabela de contingência a seguir que apresenta os dados das variáveis categóricas X e Y.

Tabela 1: Tabelas de contingência - variáveis categóricas X e Y.

$X \backslash Y$	Coluna 1	$\cdots$	Coluna j	$\cdots$	Coluna c	Total
Linha 1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Linha i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Linha l	n_{l1}	$\cdots$	n_{lj}	$\cdots$	n_{lc}	n_{l+}
Total	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+c}	n_{++}

O delineamento realizado selecionando uma amostra com n elementos que foram classificados segundo as categorias de X e Y resultando nas frequências apresentadas nas respectivas caselas da Tabela 1. Como o tamanho da amostra foi fixado à priori, mas não os totais das linhas e colunas, esses sendo aleatórios, o modelo amostral é o multinomial.

O interesse é verificar se existe associação entre as variáveis X e Y, ou seja, verificar se a variável X explica o comportamento da variável Y, por exemplo. A abordagem estatística é realizada através do teste de independência.

2.1.1 Teste de Independência

$H_0 : P(X_i \cup Y_j) = P(X_i)P(Y_j)$ com $i = 1, \cdots, l$, e $j = 1, \cdots, c$
H_A : existe pelo menos uma diferença.

Quando os totais marginais para os níveis da variável categórica X são fixos ao invés de aleatórios, a distribuição conjunta de X e Y não é mais significativa, mas sim as distribuições condicionais para cada nível de X, pois as amostras nas linhas são independentes. Quando há duas respostas categóricas o modelo assumido é uma distribuição binomial para a amostra em cada linha.

Assim não faz sentido falar em teste de independência, mas sim em teste de homogeneidade. O objetivo neste teste é verificar se existe diferença no comportamento de uma variável qualitativa ao se comparar r diferentes populações.

2.1.2 Teste de Homogeneidade

- $H_0 : P_1(X_1) = \ldots = P_r(X_1)$; $\quad \ldots \quad ; P_1(X_c) = \ldots = P_r(X_c)$
- H_A : Existe pelo menos uma diferença.

Diferentes medidas estatísticas derivadas da distribuição qui-quadrado χ^2, podem auxiliar a determinar a rejeição de H_0, levando em consideração o p-valor do teste e o nível de significância α estabelecido à priori.

Modelos probabilísticos podem ser utilizados para representar as frequências das tabelas de contingência, sendo que cada sujeito na amostra é aleatoriamente escolhido, de alguma população de interesse e classificado segundo a resposta categórica. Para tabelas 2×2 tem-se as seguintes probabilidades:

Tabela 2: Distribuição de probabilidade conjunta de X e Y (categoricas).

$X \backslash Y$	Y_1	Y_2	Total
X_1	π_{11}	π_{12}	π_1
X_2	π_{21}	π_{22}	π_2

Na Tabela 2 as variáveis X e Y são categóricas e $\pi_{ij} = P(X = i, Y = j)$, representa a probabilidade conjunta de X pertencer a categoria i e Y pertencer a categoria j. Assim, serão definidas algumas medidas importantes, podendo ser estendidas para outras análises.

2.1.3 Razão de Chances

Razão de chances (*Odds Ratio*) é uma medida de associação usada para tabelas de contingência.

A probabilidade de sucesso do nível 1 de uma variável categórica X pode ser representada por π_1, e a chance ($odds_1$) de sucesso é definida como:

$$odds_1 = \frac{\pi_1}{1-\pi_1}. \tag{1}$$

De forma análoga pode-se definir a probabilidade de sucesso do nível 2 e sua chance de sucesso é definida por:

$$odds_2 = \frac{\pi_2}{1-\pi_2}. \tag{2}$$

As chances têm valores não negativos, os quais são maiores que um quando o sucesso é mais provável de ocorrer. Para qualquer nível da variável X , a probabilidade de sucesso é uma função das chances representada por:

$$\pi = \frac{odds}{odds+1}. \tag{3}$$

A razão de chances dos dois níveis de X, mais conhecida como *"odds ratio"*, é expressa por:

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/1-\pi_2}. \tag{4}$$

Quando ambas as variáveis são *variáveis respostas*, a razão de chances pode ser definida através das probabilidades conjuntas:

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{5}$$

A partir da Equação 5 nota-se que θ é a razão do produto cruzado das diagonais da tabela 2×2, motivo pelo qual a razão das chances também é chamada de razão do produto cruzado. A razão de chances para uma amostra tem a seguinte expressão:

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{6}$$

Em pequenas amostras, a distribuição de $\hat{\theta}$ é muito assimétrica. Para contornar tal inconveniente, utiliza-se a transformação logarítmica para obter simetria em torno do zero, aproximando para uma distribuição normal. Devido a esse fato, a estimativa da média é $ln(\theta)$ e os desvio assintótico é representado por:

$$ASE[ln(\hat{\theta})] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{7}$$

A teoria anterior pode ser estendida para tabelas de contingência com dimensões maiores. Para isso, basta escolher uma categoria como referência e comparar as demais com ela, obtendo várias comparações duas a duas em relação ao nível de referência adotado.

2.2 Introdução ao Modelo de Resposta Binária

Como visto anteriormente, tabelas de contingência representam variáveis categóricas, onde suas caselas correspondem as frequências observadas em uma amostra de acordo com determinado estudo, o objetivo principal nessas tabelas é verificar se existe alguma associação entre as variáveis através das frequências observadas.

Variáveis categóricas podem ser interpretadas como variável resposta (dependente) e variável explicativa (independente) sugerindo a presença de um possível modelo que relacione tais variáveis.

Em particular quando a variável resposta em uma tabela de contingência é qualitativa e assume apenas dois valores, será trabalhado com um modelo de regressão baseado na regressão simples, conhecido como modelo regressão logística.

A variável resposta no modelo de regressão logística representa categorias, sucesso ou fracasso, que podem ser transformadas em respostas binárias, 0 ou 1, isso representa uma maior facilidade quando comparado ao modelo de regressão simples, pois os indivíduos são classificados em apenas duas categorias, além disso o modelo fornece resultados em termos de probabilidades o que é bastante útil em muitas áreas que trabalham com previsões e apresentam um pequeno número de suposições, tornando o modelo atrativo para diversos setores.

- Na medicina pode ser usada para prever a mortalidade de pacientes feridos, prever o risco de desenvolver uma dada doença, baseado em características observadas do paciente.

- Na engenharia, especialmente para predizer a probabilidade de falha em um dado processo, sistema ou produto.
- No marketing na previsão da propensão de um cliente para comprar um produto, interromper a assinatura de um serviço.
- Em economia ela pode ser utilizada para prever a probabilidade de uma pessoa estar trabalhando, de um proprietário optar por uma hipoteca.
- Em finanças pode detectar os grupos de risco para a subscrição de um crédito.

A seguir será desenvolvida a teoria sobre o modelo de regressão logística com o objetivo de: estimar seus parâmetros através do método da máxima verossimilhança, calcular intervalos de confiança e realizar testes para verificar a significância dos parâmetros do modelo, em seguida serão propostos alguns testes para investigar a adequabilidade do modelo escolhido.

2.3 Modelo para variável Resposta Binária

De acordo com KUTNER (2005) o modelo de regressão linear simples é representado por:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{8}$$

com as seguintes suposições: $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ e $Y_i \overset{i.i.d}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$.

O problema surge quando a variável resposta Y_i é binária, assumindo apenas valores 0 e 1, simplificando as variáveis encontradas em tabelas de contingência $l \times 2$ com l linhas e 2 colunas. Nesses casos pode-se pensar 1 como sucesso e 0 como fracasso, o que remete $Y_i \overset{i.i.d}{\sim} Bernoulli$.

Y_i	Probabilidade
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Considerando que o valor esperado para modelo de regressão linear simples seja $E(Y_i) = \beta_0 + \beta_1 x_i$ e que $E(Y_i) = \pi_i$, tem-se:

$$E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i. \tag{9}$$

Das suposições do modelo de regressão linear simples, alguns problemas surgem quando Y_i é binário:

1) **Não normalidade dos erros:**

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i).$$

Quando $Y_i = 1 \Rightarrow \epsilon_i = 1 - (\beta_0 + \beta_1 x_i)$.
Quando $Y_i = 0 \Rightarrow \epsilon_i = -(\beta_0 + \beta_1 x_i)$.

Como o erro assume apenas dois valores, logo não são normalmente distribuídos.

2) **Variância dos erros não constante**

Considerando que: $V(Y_i) = V(\epsilon_i) = \pi_i(1 - \pi_i)$ e lembrando que $E(Y_i) = \pi_i = \beta_0 + \beta_1 x_i$

$$V(\epsilon_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 + x_i \beta_1 x_i). \tag{10}$$

Levando a conclusão que $V(\mathcal{E}_i)$ depende de x_i, assim a variância dos erros difere para diferentes níveis de X.

3) **Restrições na função resposta**

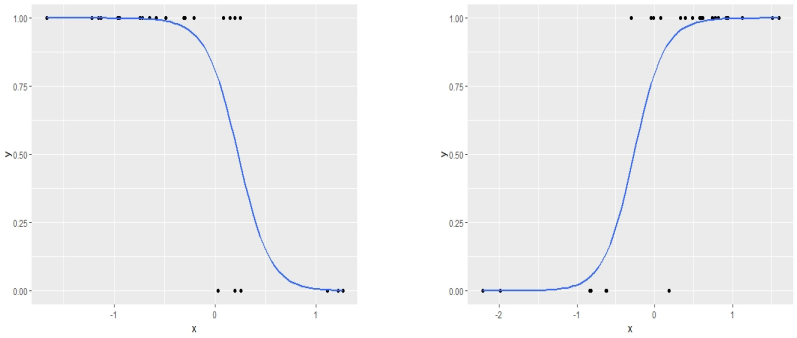
A função resposta representa probabilidades quando a variável indicadora, 0 e 1, é uma variável resposta, assim o valor esperado deve estar restrito a:

$$0 \leq E[Y] = \pi \leq 1. \tag{11}$$

O problema é que muitas funções respostas não possuem automaticamente esta restrição para valores entre 0 e 1. Segundo KUTNER (2005) "considerações empíricas e teóricas sugerem que quando a variável resposta é binária, a forma da função resposta pode frequentemente ser curvilínea ". Os gráficos a seguir, construídos a partir da simulação usando uma amostra de tamanho 25 e valores fixados para β apresentam exemplos da forma dessa função.

Note que elas tem a forma ou inclinação em S e são aproximadamente lineares exceto nos extremos. Estas funções respostas são conhecidas como função resposta logística e são geralmente definidas como sigmoides. Diante dessa análise é possível determinar algumas propriedades.

Figura 1: Funçao resposta (dados simulados).



(a) n = 25, $\beta_0 = 2$, $\beta_1 = -3$ (b) n = 25, $\beta_0 = 1$, $\beta_1 = 3$

2.4 Modelo de Regressão Logística Simples

A forma usual para o modelo de regressão logística simples é:

$$Y_i = E[Y_i] + \epsilon_i. \tag{12}$$

A distribuição dos erros depende da distribuição de Bernoulli na variável resposta, já os Y_i são variáveis aleatórias independentes com os valores esperados:

$$E[Y_i] = \pi_i = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}. \tag{13}$$

2.4.1 Propriedades da função resposta logística

a) O sinal de β_1, revela se a função é crescente ou decrescente monotônicas.

b) É quase linear na faixa onde $E[Y]$ está entre 0.2 e 0.8 e gradualmente se aproxima de 0 e 1 nos dois finais da faixa.

c) Facilidade da sua linearização a partir da transformação:

$$logito[\pi_i] = ln\left(\frac{\pi_i}{1-\pi_i}\right) \Rightarrow logito[\pi_i] = \beta_0 + \beta_1 x_i. \tag{14}$$

A Equação 14 é chamada de tranformação logito de probabilidade π_i. A razão $\frac{\pi_i}{1-\pi_i}$, é conhecida por chance ($odds$), β_1 indica a alteração na razão de chances

$$\frac{odds(x_1)}{odds(x_2)} = exp[(\beta_1)(x_1 - x_2)]. \tag{15}$$

Utilizando exponencial e fazendo as devidas modificações na Equação 14, chega-se a forma para a função resposta logística,

$$E[Y_i] = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}, \tag{16}$$

uma forma equivalente é: $E[Y_i] = [1 + exp(-(\beta_0 + \beta_1 x_i))]^{-1}$.

2.4.2 Estimação dos parâmetros

Como $Y_i \overset{i.i.d}{\sim} Bernoulli$, $P(Y_i = 1) = \pi_i$ e $P(Y_i = 0) = 1 - \pi_i$, a distribuição de probabilidade é:

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, \quad Y_i = 0, 1 \quad e \quad i = 1, \cdots, n. \tag{17}$$

Note que: $f_i(1) = \pi_i$ e $f_i(0) = 1 - \pi_i$, consequentemente $f_i(Y_i)$ representa a probabilidade que $Y_i = 1$ ou 0, desde que as Y_i observações sejam independentes, assim a função de probabilidade conjunta, isto é, a função de verossimilhança, é dada por:

$$g(Y_1, \ \ldots \ , Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}. \tag{18}$$

Para encontrar as estimativas de máxima verossimilhança pode-se utilizar o logaritmo da função de probabilidade conjunta, pois as funções são crescentes e o máximo da função log também será da função de distribuição, facilitando a obtenção das estimativas.

$$ln(g(Y_1, \ \ldots \ , Y_n)) = ln\left[\prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}\right] = ln\left[\prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{Y_i}(1 - \pi_i)^1\right] \tag{19}$$

$$= \sum_{i=1}^{n} Y_i ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \sum_{i=1}^{n} ln(1 - \pi_i).$$

Lembrando que $E[Y_i] = \pi_i$,

$$1 - \pi_i = [1 + exp(\beta_0 + \beta_1 x_i)]^{-1} \ \text{ e que } \ ln\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_i.$$

Considerando que $Y_i = y_i$ represente apenas os sucessos para a variável resposta x_i tem-se:

$$lnL(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln[1 + exp(\beta_0 + \beta_1 x_i)]^{-1}. \quad (20)$$

As estimativas para β_0 e β_1, no modelo de regressão logística, são aquelas que maximizam a Equação 19. Os estimadores de Máxima Verossimilhança (MLV), são obtidos pela solução das equações resultantes da derivação da função de verossimilhança em relação a β_0 e β_1.

Como não existe uma solução com forma fechada, os parâmetros são estimados através de procedimentos numéricos, é comum utilizar $\hat{\pi}_i$ como o valor ajustado do i-ésimo caso

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\pi}_i = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}. \quad (21)$$

. A inversa da matriz de informação de Fischer $\mathcal{I}(\beta_0, \beta_1)^{-1}$, constituída com as segundas derivadas de $ln(L)$, equação 20, fornece as variâncias assintóticas e covariâncias para os parâmetros estimados.

$$\mathcal{I}(\beta_0, \beta_1)^{-1} = \begin{pmatrix} V(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_1) \end{pmatrix}$$

A estimação é obtida avaliando a matriz no MLV$(\hat{\beta}_0, \hat{\beta}_1)$.

$$\mathcal{I}(\beta_0, \beta_1)^{-1} = \begin{pmatrix} \widehat{V}(\hat{\beta}_0) & \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{V}(\hat{\beta}_1) \end{pmatrix}$$

2.5 Variável Resposta Binária com Distribuição Binomial

A teoria apresentada considerou $Y_i \overset{i.i.d}{\sim} Bernoulli$, porém, a maioria dos eventos não ocorrem apenas uma vez, são n repetições de *Bernoulli*, também chamada de distribuição *Binomial*.

O estudo tendo como variável resposta uma *Bernoulli* pode ser facilmente extendido para uma variável resposta *Binomial*.

Seja Y_i o número de sucessos em n ensaios de *Bernoulli* com probabilidade π_i, logo $Y_i \overset{i.i.d}{\sim} Bin(n_i, \pi_i)$. A probabilidade de Y assumir um valor inteiro j ($j = 0, 1, \cdots, n_i$) é dada por:

$$P(Y_i = j) = \begin{pmatrix} n_i \\ j \end{pmatrix} \pi_i^j (1-\pi_i)^{n_i - j} = \frac{n_i!}{j!(n_i - j)!} \pi_i (1-\pi_i)^{n-j}.$$

A média de Y_i e a variância são dadas por

$$E(Y_i) = n_i \pi_i \ e \ V(Y_i) = n_i \pi_i (1-\pi_i).$$

Y_i também pode ser descrito como a proporção de sucesso para cada i. Usando as propriedade da esperança e variância obtem-se o seguinte resultado:

$$E\left(\frac{Y_i}{n_i}\right) = \pi_i \ \ e \ \ V\left(\frac{Y_i}{n_i}\right) = \frac{\pi_i(1-\pi_i)}{n_i}.$$

De acordo com SCHEATHER (2009) deve-se considerar a proporção da amostra de "sucessos"(Y_i/n_i) como a resposta desde que:

- Y_i/n_i é um estimador não viesado de π_i.
- Y_i/n_i varia entre 0 e 1.

Para estimação dos parâmetros quando $Y_i \overset{i.i.d}{\sim} Bin(n_i, \pi_i)$, o procedimento realizado é semelhante ao que foi feito quando $Y_i \overset{i.i.d}{\sim} Bernoulli$, salvo pela adição de alguns termos na equação, logo, também são necessários métodos iterativos não númericos como *Newton-Raphson* ou mínimos quadrados ponderados.

A função de log-versossimilhança pode ser declarada como:

$$lnL(\beta_0, \beta_1) = \sum_{i=1}^{c} \left[ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i(\beta_0 + \beta_1 x_i) - n_i ln(1 + exp(\beta_0 + \beta_1 x_i)) \right]. \quad (22)$$

Na Equação 22 os níveis da variável explicativa X são obtidos por $x_1 \cdots x_c$, já o número de observações em cada nível é denotado por n_i e o número de sucessos no nível i é y_i.

2.5.1 Inferências sobre os parâmetros de regressão

As estimativas dos parâmetros e seus respectivos erros padrão (ASE), obtidos através da matriz de informação de Fischer, auxiliam a encontrar intervalos de confiança com um nível de significância α.

$$IC(\beta_1, 1-\alpha) = \hat{\beta}_1 \pm Z_{\left(1-\frac{\alpha}{2}\right)} ASE(\hat{\beta}_1). \tag{23}$$

A interpretação desses intervalos podem ser feitas da seguinte forma, existe $(1-\alpha)\%$ de confiança do verdadeiro valor do parâmetro estar entre esses limites.

Como $exp(\hat{\beta}_1)$ representa a razão de chances, aplicando a função exponencial na Equação 23, chega-se no seu respectivo intervalo de confiança.

$$IC(\hat{\theta}, 1-\alpha) = exp[\hat{\beta}_1 \pm Z_{\left(1-\frac{\alpha}{2}\right)} ASE(\hat{\beta}_1)]. \tag{24}$$

A estimação no ponto x_0 será o $logito[\hat{\pi}(\mathrm{x}_0)] = \hat{\beta}_0 + \hat{\beta}_1 x_i$, em que a variância do estimador é dada por:

$$V(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0) = V(\hat{\beta}_0) + \mathrm{x}_0^2 V(\hat{\beta}_1) + 2\mathrm{x}_0 Cov(\hat{\beta}_0, \hat{\beta}_1). \tag{25}$$

Sendo o erro assintótico estimado, $ASE = \sqrt{\widehat{V}(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0)}$, é possível obter intervalo de confiança para um ponto dado e um nível de significância α.

$$IC(\mathrm{x}_0, 1-\alpha) = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0 \pm Z_{\left(1-\frac{\alpha}{2}\right)} ASE. \tag{26}$$

Da Equação 26, chega-se a um intervalo de confiança para a probabilidade em um ponto x_0, substituindo os limites superiores (L_s) e inferiores (L_i) na Equação 21.

$$\pi[x_0] = \left(\frac{exp(L_i)}{1+exp(L_i)}; \frac{exp(L_s)}{1+exp(L_s)} \right) \tag{27}$$

As estimativas acima podem ser estendidas para outros modelos de regressão, que utilizam mais variáveis explicativas, ou mesmo mais variáveis respostas. Neste caso, os escores e a diagonal da matriz de informação para um dado β_j irá envolver o respectivo x_j. Os elementos fora da diagonal irão envolver produtos de x_i e x_j.

2.6 Modelo de Regressão Logística Múltipla

O modelo de regressão logística simples pode ser extendido para mais de uma variável explicativa, substituindo-se $\beta_0 + \beta_1 x_i$ por $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}$. Para simplificar, será utilizada a notação vetorial a seguir.

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \mathbf{X}_{p\times 1} = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix} \qquad \mathbf{X}_{i,p\times 1} = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

- $\boldsymbol{\beta}'\mathbf{X}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1}$, as observações X são aleatórias.

A função resposta simples é extendida para a múltipla:

$$E[Y_i] = \frac{exp(\boldsymbol{\beta}'\mathbf{X}_i)}{1 + exp(\boldsymbol{\beta}'\mathbf{X}_i)}. \tag{28}$$

De modo análogo a transformação logíto será:

$$\pi' = ln\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{\beta}'\mathbf{X}_i. \tag{29}$$

Similarmente ao modelo de regressão logística simples o modelo de regressão logística múltiplo pode ser declarado como:

$$E[Y_i] = \pi_i = \frac{exp(\boldsymbol{\beta}'\mathbf{X}_i)}{1 + exp(\boldsymbol{\beta}'\mathbf{X}_i)} \quad em\ que\ Y_i \overset{i.i.d}{\sim} Binomial. \tag{30}$$

Os parâmetros do modelo de regressão logística múltipla podem ser obtidos através do método de máxima verossimilhança

$$lnL(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\mathbf{X}_i) - \sum_{i=1}^{n} ln[1 + exp(\boldsymbol{\beta}'\mathbf{X}_i)]. \tag{31}$$

Aqui também serão utilizados procedimentos numéricos para encontrar os valores de $\beta_0, \beta_1, \ldots, \beta_{p-1}$, que maximizam $lnL(\beta)$, as estimativas serão denotadas por $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$, sendo seu vetor:

$$\hat{\boldsymbol{\beta}}_{p\times 1} = \begin{bmatrix} 1 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$$

A função resposta logística e seus valores ajustados podem ser expressos como:

$$\hat{\pi}_i = \frac{exp(\exp \hat{\boldsymbol{\beta}}' \mathbf{X}_i)}{1 + exp(\exp \hat{\boldsymbol{\beta}}' \mathbf{X}_i)} = [1 + exp(-\hat{\boldsymbol{\beta}}' \mathbf{X}_i)]^{-1}. \tag{32}$$

2.6.1 *Odds ratio*

A interpretação da razão de chances (*odds ratio*) para a variável explicativa β_k do modelo de regressão logística múltiplo será a mesma do modelo simples, salvo que todas as outras variáveis são mantidas constantes.

1. **Quando as variáveis explicativas são qualitativas**:
   A chance de sucesso na i-ésima categoria aumenta/diminiu, comparado com a categoria de referência, mantida todas as demais variáveis constantes.

2. **Quando as variáveis explicativas são quantitativas**:
   A chance de sucesso aumenta/diminiu na variável analisada a cada aumento de uma unidade nessa variável, mantida todas as demais constantes.

2.6.2 Inferência sobre a resposta média

Se X_h representa a matriz cuja as variáveis assumem valores interesse para um determinado estudo, a resposta média de será:

$$\pi_h = [1 + exp(-\boldsymbol{\beta}' \mathbf{X}_h)]^{-1}. \tag{33}$$

Em que $\mathbf{X}_h$ possui a seguinte forma matricial:

$$\mathbf{X_h} = \begin{bmatrix} 1 \\ X_{h,1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

O **estimador pontual** de π_h será denotado por $\hat{\pi}_h$

$$\hat{\pi}_h = [1 + exp(-\hat{\boldsymbol{\beta}}'\mathbf{X}_h)]^{-1}. \tag{34}$$

Lembrando que $\hat{\boldsymbol{\beta}}'$ é o vetor transposto de estimadores para o vetor de parâmetros $\boldsymbol{\beta}$.

Para encontrar o **intervalo de confiança** de π_h será necessário calcular os limites de confiança para a resposta média logito $\hat{\pi}'_h$, de forma análoga ao que foi feito para o modelo de regressão quando a resposta é binária, segue que:

- $\hat{\pi}'_h = \hat{\boldsymbol{\beta}}'\mathbf{X}_h$ é o estimador pontual para a resposta média logito.
- $ASE(\hat{\pi}'_h) = ASE(\mathbf{X}'_h\hat{\boldsymbol{\beta}}) = \mathbf{X}'_h ASE(\hat{\boldsymbol{\beta}})\mathbf{X}_h$ o erro assintótico do estimador pontual.

Assim os limites do intervalo de confiança Inferior(I) e Superior(S) para a resposta média logito são respectivamente:

$$I = \hat{\pi}'_h - Z_{\left(1-\frac{\alpha}{2}\right)} ASE(\hat{\pi}'_h) \;\; e \;\; S = \hat{\pi}'_h + Z_{\left(1-\frac{\alpha}{2}\right)} ASE(\hat{\pi}'_h). \tag{35}$$

Usando a relação monotômica entre π_h e $\hat{\pi_h}'$ para converter os limites de confiança entre I e S para π_h tem-se:

$$I^* = [1 + exp(-I)]^{-1} \;\; e \;\; S^* = [1 + exp(-S)]^{-1}. \tag{36}$$

chegando assim ao intervalo de confiança para as respostas médias.

2.7 Testes para os parâmetros do modelo

Para determinar se um parâmetro estimado é signficativo para o modelo de regressão logística alguns testes podem ser realizados como, *Wald*, razão de verossimilhança e *Score*.

2.7.1 Teste sobre um único parâmetro do modelo

O teste de *Wald* para os parâmetros do modelo $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$, pode ser realizado sobre um único parâmetro $\hat{\beta}_k$. A hipótese estatística e a regra de decisão do teste são respectivamente:

H_0: $\hat{\beta}_k = 0$
H_A: $\hat{\beta}_K \neq 0$.

- A estatística do teste é:

$$z^* = \frac{\hat{\beta}_k}{ASE(\hat{\beta_k})} \qquad (z^*) \sim N(0,1) \;\; e \;\; (z^*)^2 \sim \chi^2_1. \tag{37}$$

- A regra de decisão sob H_0 é:

$|Z^*| \leq Z_{\left(1-\frac{\alpha}{2}\right)}$ conclui-se H_0.
$|Z^*| > Z_{\left(1-\frac{\alpha}{2}\right)}$ conclui-se H_A.

2.7.2 Teste sobre vários parâmetros do modelo

Também é possível realizar o teste para vários $\hat{\beta}_k = 0$, utilizando o teste da razão de verossimilhança, para isso será necessário definir o modelo completo e o modelo reduzido.

$L(F) = [1+exp(-\mathbf{X}\hat{\boldsymbol{\beta}}_F)]^{-1}$ *sendo* : $\mathbf{X}\hat{\boldsymbol{\beta}}_F = \hat{\beta}_0+\hat{\beta}_1 x+\ldots+\hat{\beta}_{p-1}x_{p-1}$ **Modelo Completo**.

$L(R) = [1+exp(-\mathbf{X}\hat{\boldsymbol{\beta}}_R)]^{-1}$ *sendo* : $\mathbf{X}\hat{\boldsymbol{\beta}}_R = \hat{\beta}_0+\hat{\beta}_1 x+\ldots+\hat{\beta}_{q-1}x_{q-1}$ **Modelo Reduzido**.

A hipótese, a estatística e a regra de decisão do teste são respectivamente:

H_0: $\hat{\beta}_q = \hat{\beta}_{q+1} = \ldots = \hat{\beta}_{p-1} = 0$
H_A: Pelo menos $\hat{\beta}_k$ em H_0 é diferente de zero.

$$G^2 = 2ln\left[\frac{L(R)}{L(F)}\right] = 2[lnL(R) - lnL(F)] = G^2_{H_o} - G^2_{H_A}. \tag{38}$$

- $G^2 \leq \chi^2_{(1-\alpha;p-q)}$ conclui-se H_0.
- $G^2 > \chi^2_{(1-\alpha;p-q)}$ conclui-se H_A.

Observe que é utilzada a diferença das *deviances* (G^2) nos modelos, $G^2_{H_o}$ do modelo reduzido e $G^2_{H_A}$ do modelo completo. A forma para a *deviance* será apresentada Subseção 2.10.

2.8 Métodos para a seleção do modelo

Para a seleção do modelo de regressão logística, é necessário escolher as variáveis que irão compor esse modelo, tornando-o o mais parcimonioso possível, ou seja, um modelo que irá explicar bem o comportamento da variável resposta envolvendo o mínimo de parâmetros possíveis na estimação.

Existem alguns procedimentos que auxiliam a escolha das variáveis que farão parte do modelo entre eles:

1) Observar cuidadosamente cada potencial preditor. Para variáveis nominais e ordinais, verificar através de tabelas de contingência categorias com zero, unindo ou eliminando-as. Já para variáveis contínuas é necessário examinar o modelo.
2) Selecionar variáveis no modelo univariado utilizando um p-valor mais abrangente do que seria utilizado em modelo com múltiplas variáveis.
3) Utilizar os procedimentos *Stepwise/Forward/Backward* para seleção das variáveis preditoras. Esses métodos são algoritmos que utilizam algum critério definido à priori pelo analista, para escolha das variáveis que permanecerão no modelo.

- No método *Forward* pode-se incluir variáveis no modelo mais simples, chegando a um modelo mais complexo, até que a adição de uma nova variável não melhore o modelo, de acordo com um critério indicado pelo analista.
- Ao contrário do método *Forward*, o método *Backward* parte de um modelo mais complexo eliminando as variáveis explicativas. O processo "para" quando a retirada de uma variável produz um ajuste inadequado, de acordo com um critério indicado pelo analista.
- Também é possível fazer adição ou remoção atendendo a regra de decisão em cada etapa, nesse caso é utilizado o método *Stepwize*.

No caso em que variáveis explicativas de um modelo maior estão contidas em um modelo menor pode-se utilizar o Critério de Informação de Akaike AIC_p ou o Critério Bayesiano de Schwarz SBC_p, também chamado de Critério de Informação Bayesiano de Schwarz BIC, ambos baseados no máximo da função de verossimilhança (MFV), para escolher entre os modelos, em que:

$$AIC_p = -2ln(L(\hat{\boldsymbol{\beta}}')) + 2p. \tag{39}$$

$$BIC = -2ln(L(\hat{\boldsymbol{\beta}}')) + p\ ln(n). \tag{40}$$

Modelos promissores com p parâmetros apresentam pequenos valores AIC_p e SBC_p.

2.9 Análise dos Resíduos

Os resíduos são simplesmente $y_i - \hat{\pi}_i$ e $\hat{\pi}_i$ é o termo previsto no modelo. De acordo com a literatura Vários tipos de resíduos são aplicáveis à regressão logística. SHEATHER (2009) cita pelo menos três tipos de resíduos na regressão logística:

2.9.1 Resíduo da resposta

$$r_i = y_i/n_i - \hat{\pi}_i.$$

$\hat{\pi}_i$ é o i-ésimo valor ajustado do modelo de regressão logística.

Problemas surgem quando a variância de y_i/n_i é não constante o que dificulta a interpretação. Esse problema pode ser contornado com os resíduos de Pearson.

2.9.2 Resíduo de Pearson e resíduo de Pearson padronizados

$$r_{Pearson,i} = \frac{y_i - \hat{\pi}_i}{\sqrt{V(y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} = \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Já os resíduos padronizados de Pearson são dados por

$$sr_{Pearson,i} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{V(y_i - n_i\hat{\pi}_i)}}.$$

De acordo AGRESTI (1996) quando n_i é grande, o resíduo de Pearson $r_{Pearson,i}$ tem aproximadamente uma distribuição normal. Se o número de parâmetros do modelo for pequeno em comparação com o número de $logito[\pi_i]$ da amostra, os resíduos de Pearson são tratados como desvios, com valores absolutos maiores que 2 indicando possível falta de ajuste.

2.9.3 Resíduo *deviance* e resíduo *deviance* padronizado

São definidos de maneira análoga aos resíduos de Pearson com a estatística de adequação de Pearson substituída pela *deviance* G^2

$$r_{Deviance} = sign(y_i - n_i\hat{\pi}_i)g_i,$$

em que $G^2 = \sum_i^n g_i^2$. Os resíduos padronizados são definidos por:

$$sr_{Deviance,i} = \frac{r_{Deviance,i}}{\sqrt{V(y_i - n_i\hat{\pi}_i)}}.$$

Embora os resíduos de Pearson sejam os mais utilizados, os resíduos *deviance* (ou resíduos de *deviance* padronizados) são realmente os escolhidos, uma vez que sua distribuição está mais próxima dos resíduos de quadrados mínimos de acordo com SIMONOFF (2003).

KUTNER (2005) recorda que como Y_i assume apenas valores 0 e 1 consequentemente os resíduos não são normalmente distribuídos, ao contrário do

modelo de regressão linear gráficos dos resíduos versus valores ajustados ou variáveis preditoras serão pouco informativos, porém, essas medidas serão de grande utilidade para verificar o ajuste do modelo a seguir.

2.10 Testes para a qualidade do ajuste do modelo (diagnóstico)

O diagnóstico para verificar se o modelo se ajusta aos dados é extremamente importante, uma forma de realizá-lo é investigar as suposições do modelo através de testes e análise gráfica. Os testes para ratificar a qualidade do ajuste estão apresentados a seguir.

H_0: $E[Y_i] = [1 + exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i)]^{-1}$

H_A: $E[Y_i] \neq [1 + exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_i)]^{-1}$.

As hipóteses acima equivalem a seguinte afirmação:

H_0 *O modelo de regressão logística é apropriado*

H_A: *O modelo de regressão logística é inapropriado.*

2.10.1 Teste X^2

$$X^2 = \sum_{i=0}^{s} r_i^2 = \sum_{i=1}^{s} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \sim \chi^2_{s-2}. \tag{41}$$

- Se $X^2 \leq \chi^2_{(1-\alpha,m-(p+1))}$ conclui-se H_0.
- Se $X^2 > \chi^2_{(1-\alpha,m-(p+1))}$ conclui-se H_A.

A estatística X^2 é baseada nos resíduos de Pearson (HOSMER e LEMESHOW, 2000).

2.10.2 Teste da *Deviance*

$$G^2 = 2\sum_{i=1}^{s}\left\{y_i ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)ln\left(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right)\right\}. \tag{42}$$

– Caso $G^2 \leq \chi^2_{(1-\alpha;m-(p+1))}$ conclui-se H_0.
– Caso $G^2 > \chi^2_{(1-\alpha;m-(p+1))}$ conclui-se H_A.

A estatística *deviance* é baseada nos resíduos da *deviance* (HOSMER e LEMESHOW, 2000). Aqui é utilizada apenas a *deviance* residual.

Os graus de liberdade *(gl)* da estatística X^2 e *Deviance* é dado por (m - (p+1)) onde m são as amostras $logito[\pi_i]$ (amostras da binomial) e p é o número de parâmetros do modelo. Quando $m_i = 1$, dados binários, a qualidade do ajuste X^2 e G^2 são mais discutíveis e os gráficos dos resíduos podem ser mais difíceis de interpretar (SHEATHER, 2009).

2.10.3 Teste de Hosmer-Lemeshow para preditores contínuos

Consiste em agrupar em classes casos com valores ajustados $\hat{\pi}$ similares com aproximadamente o mesmo número de casos. Assim n/10 observações com maiores probabilidades estimadas são colocadas na primeira classe e assim por diante.

Considerando as classes formadas por $\hat{\pi}$ é possível calcular a estatística χ^2 de Pearson, cuja regra de decisão é dada por:

– Caso $X^2 \leq \chi^2_{(1-\alpha;c-2)}$ conclui-se H_0.
– Caso $X^2 > \chi^2_{(1-\alpha;c-2)}$ conclui-se H_A.

Em que c-2 é o número de classes.

2.10.4 Pseudo R^2

Semelhante ao coeficiente de determinação R^2 da regressão múltipla, a medida de pseudo R^2 representa o ajuste geral do modelo proposto. Sua interpretação, portanto, é semelhante à regressão múltipla.

$$R^2_{dev} = 1 - \frac{G^2_F}{G^2_R}.$$

- Lembrando que F representa o modelo completo (todos os parâmetros), já R representa o modelo reduzido (sem algum parâmetro).
- É possível obter diferentes medidas para o pseudo R^2 dentre elas constam as estipuladas por *Cox e Snell*, *Nagelkerke*, *McFadden*, *Tjur* e *Pearson*.

2.11 Predição para uma nova observação

A predição para observações futuras no modelo de regressão logístico difere do modelo de regressão simples, aqui será necessário a escolha de um ponto de corte, assim as observações que estejam abaixo desse valor representam fracasso ($Y_i = 0$) e acima desse valor representem sucesso ($Y_i = 1$), uma vez que a função de regressão estima a probabilidade de sucesso $\hat{\pi}_i$.

De forma simplificada, em geral o resultado 1 será previsto se o valor estimado $\hat{\pi}_h$, para níveis dados de X_h, for alto e o resultado 0 será previsto se o valor estimado para π_h for pequeno para dados níveis de X_h, a dificuldade consiste em encontrar o ponto de corte. Segundo KUTNER (2005), três abordagens podem ser consideradas.

1) Essa é a forma mais simples consiste em usar 0.5 como ponto de corte.
2) Encontrar o melhor ponto de corte para o conjunto de dados no qual o modelo de regressão logística é baseado. Utilizar diferentes pontos de corte para avaliar os casos no modelo de construído, o ponto de corte escolhido será aquele para o qual a proporção de predições for o menor.
3) Usar probabilidades *à priori* e custos incorretos de predição na determinação do ponto de corte.

2.12 Tabela de Classificação

Uma maneira prática de qualificar o ajuste do modelo logístico é através dos acertos e erros das suas predições definindo uma regra que classifica as probabilidades estimadas pertencendo a 0 ou 1.

Intuitivamente pode-se supor que para probabilidades altas, próximas de 1, o indivíduo é classificado como $\hat{Y}_i = 1$, de forma contrária para probabilidades baixas, próximas de 0, o indivíduo é classificado como $\hat{Y}_i = 0$, mas qual ponto de corte utilizar.

O problema de escolher o ponto de corte já foi discutido na section 2.11. Usualmente na literatura se utiliza o valor 0.5, mas dependendo do estudo outro nível pode ser proposto (HOSMER e LEMESHOW, 2000).

Determinado o ponto de corte será avaliado o poder de discriminação do modelo baseado nos acertos ou erros ao classificar os eventos ($\hat{Y}_i = 1$) dos não eventos ($\hat{Y}_i = 0$). Essa projeção do modelo na tabela de classificação é conhecida como matriz de classificação.

A tabela de classificação, Tabela 3, é formada pelas observações de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN).

Tabela 3: Tabela de classificação do modelo.

Valor estimado	**Valor Observado**	
	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	VP	FP
$\hat{Y} = 0$	FN	VN

Com base nas frequências da matriz de classificação alguns parâmetros podem ser estimados. Com P representando o total de eventos positivos ($Y = 1$) e N o total de não eventos ($Y = 0$), tem-se:

Precisão: $ACC = \frac{VP+VN}{P+N}$.

Sensibilidade: $SEN = \frac{VP}{FN+VP}$.

Especificidade: $ESP = \frac{VN}{VN+FP}$.

Verdadeiro Preditivo Positivo: $VPP = \frac{VP}{VN+FP}$.

Verdadeiro Preditivo Negativo: $VPN = \frac{VN}{VP+FN}$.

2.13 Curva ROC

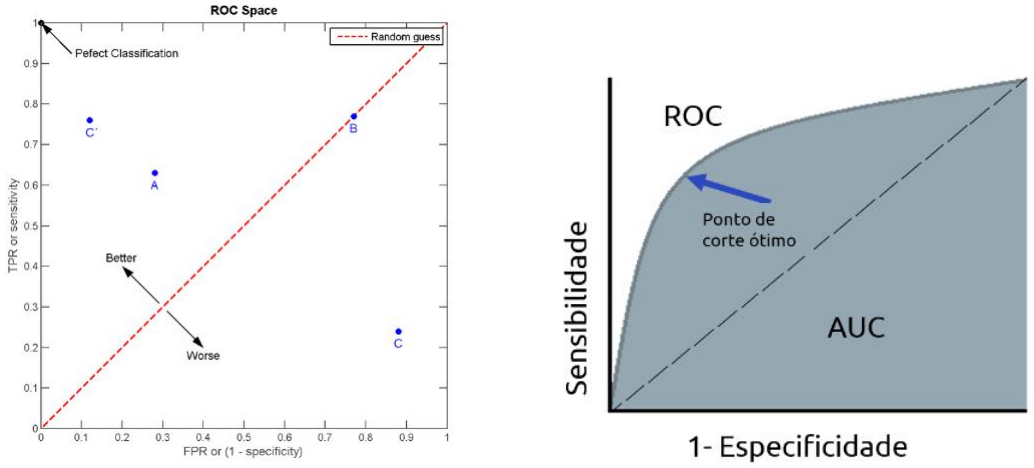
O gráfico da sensibilidade (Taxa dos Verdadeiros Positivos) versus 1 - Especificidade (Taxa dos Falso Positivos) para todos os possíveis pontos de corte é denominado de curva ROC (*Receiver Operating Characteristic Curve*) sendo que a área sob a curva fornece uma medida de discriminação que é a probabilidade de $Y_i = 1$ (HOSMER e LEMESHOW, 2000).

Como regra geral:

$\mathbf{ROC} = 0{,}5$	Sugere que não há discriminação.
$0,5 < \mathbf{ROC} < 0,7$	Discriminação pobre
$0,7 < \mathbf{ROC} < 0,8$	Discriminação aceitável.
$0,8 \leq \mathbf{ROC} < 0,9$	Discriminação excelente.
$\mathbf{ROC} \geq 0,9$	Discriminação excepcional.

A área sob a curva **ROC** fornece uma descrição mais completa da acurácia da classificação, ou seja, a capacidade do modelo discriminar entre os sujeitos que vivenciam o resultado e aqueles que não.

Figura 2: Classificador ROC.



(a) Espaço ROC - 4 classificadores. (b) Gráfico da Área Sobre a Curva (AUC).

Fontes: wikipedia(a) e Juliana Scudilio(b)

O ponto com coordenadas (0,1) do espaço ROC apresentado na Figura 2 (a) representa uma excelente classificação. Nesse ponto a sensibilidade é 100% (isto é, não há falsos negativos) e 100% de especificidade (isto é, não há falsos positivos). A Figura 2 (b) mostra o ponto de corte ótimo da curva.

3 Metodologia

A base de dados utilizada nesse estudo foi fornecida pelo Observatório da Vida Estudantil (OVE) tendo apoio do Laboratório de População e Desenvolvimento (LPD), ambos vinculados ao Centro de Estudos Avançados Multidisciplinares (CEAM), que realizou um levantamento das características dos ingressantes da UnB através da aplicação de um questionário que era disponibilizado via internet para os estudantes preencherem e obrigatório para completar o processo de registro na Universidade.

As variáveis que foram utilizadas para coletar informações, são em sua maioria de natureza qualitativa, apresentando um grande número de categorias para abranger melhor possível a heterogeneidade dos ingressantes. Ao todo são 60 perguntas divididas em 10 blocos temáticos que englobam variáveis como características sociodemográficas, formação escolar anterior, acesso ao ensino superior, escolhas relacionadas ao curso e expectativas de inserção profissional.

O estudo se refere ao período de 1/2012 a 2/2017, totalizando 12 semestres com 45.043 observações. Para facilitar a análise dos dados será utilizado os blocos temáticos abordados anteriormente, formando três grandes grupos: identificação social/demográfica, formação escolar anterior e acesso ao ensino superior. Cada grupo será constituído por variáveis apontadas como relevantes com base em estudos anteriores.

Através de uma análise descritiva das variáveis categóricas será feito um levantamento do perfil dos alunos que ingressaram na instituição buscando identificar possíveis fatores que atuam no fato do estudante não escolher o curso desejado ao participar do processo seletivo para ingresso na universidade. Tendo em vista que a variável de interesse é categórica assumindo dois valores: 0 - *o estudante ingressou no curso desejado* ou 1 - *o estudante não ingressou no curso desejado*, será adotado o **modelo de regressão logística**.

Para auxiliar nas análises descritivas será utilizada uma ferramenta de *Bussiness Intelligence*. O *Microsoft Power BI* auxilia na tomada de decisões a partir de dados históricos, através da criação de *Dashboards* iterativos que são em síntese painéis com diferentes modelos de gráficos e tabelas interligados, favorecendo a análise de múltiplas variáveis simultaneamente. O Quadro 4 apresenta as variáveis selecionadas para o presente estudo.

Quadro 4: Variáveis selecionadas para o estudo.

Variáveis Relacionadas ao Ingresso na Universidade	**Variáveis Educacionais**
Semestre	Ensino fundamental
Sistema de Ingresso	Ensino médio
Modalidade de Ingresso	Tipo de ensino médio
Campus	Atividades extracurriculares:
Turno	Realizou curso preparatório
Curso	Tipo de curso preparatório
Necessidade especial	Número de tentativas de ingresso na UnB
Variáveis Sociodemográficas	É o curso desejado?
Sexo	Motivos para não ingressar no curso desejado
Nacionalidade	Trocaria de curso?
Região de nascimento	Motivos da escolha do curso
Estado civil	**Variáveis Socioeconômicas**
Raça/Cor/Etnia	Posse de bens
Região de residência	Renda mensal familiar
Com quem reside	Número de pessoas que vivem da renda
Variáveis Relacionadas a Perspectiva Profissional	Escolaridade do pai
Grau máximo de estudo pretendido	Escolaridade da mãe
Atividade econômica	Convenio ou plano de saúde
Horas de trabalho	Assistência médica
Perspectiva profisional	

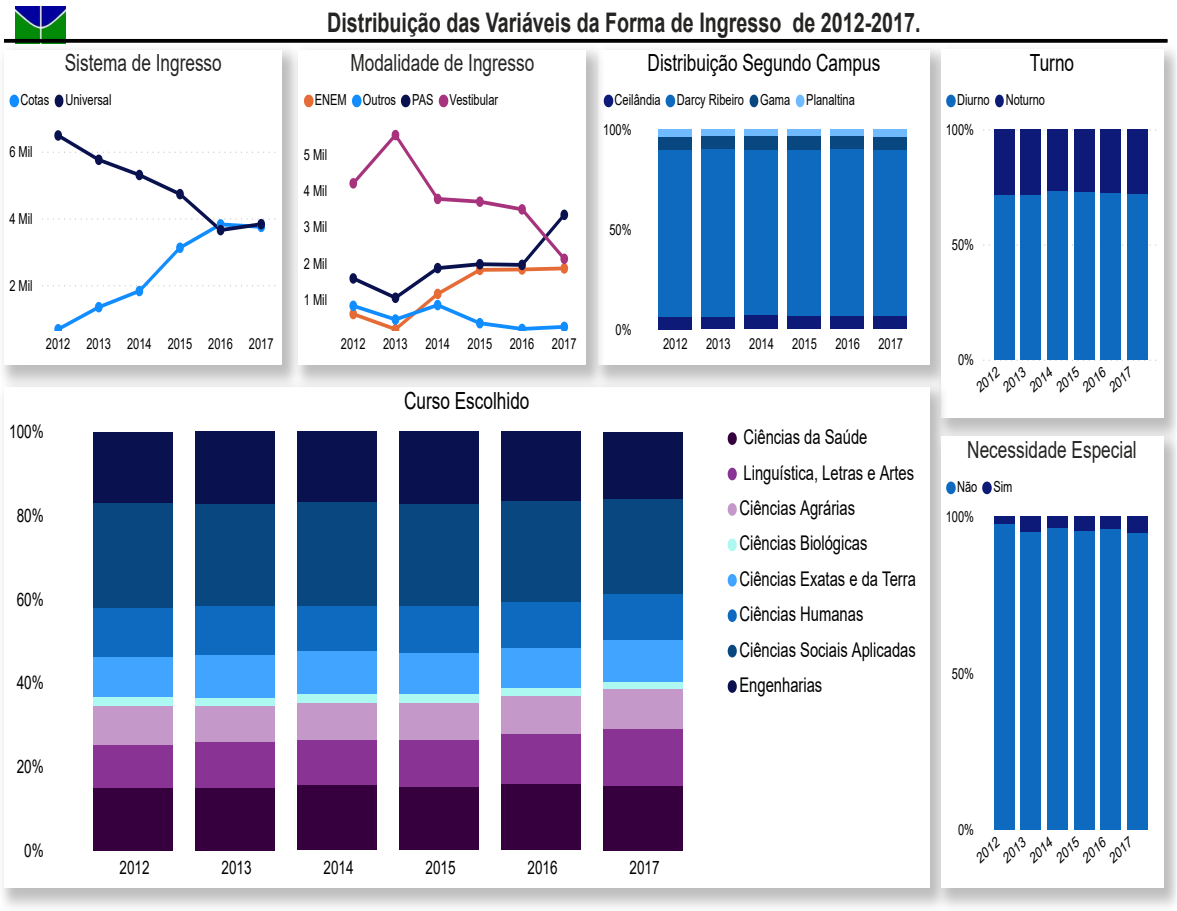
No apêndice deste trabalho, Seção 9, foram apresentadas tabelas com o total das observações para cada variável selecionada e o percentual de cada categoria ao longo do estudo, bem como agrupamentos, células coloridas com a mesma cor, realizados para análise dos dados.

Para análise da associação entre as variáveis e elaboração do **modelo de regressão logística** serão utilizados os *softwares* R e SAS.

4 Análise Descritiva

Nessa primeira etapa do estudo, o objetivo é estabelecer o perfil do estudante ingressante da Universidade de Brasília.

Figura 3: Perfil de ingresso na universidade.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Verifica-se uma grande mudança no sistema de ingresso (Figura 3) a partir do ano de 2012 em consonância com a implantação do sistema de cotas sociais[1]. Nota-se que nos anos de 2016 e 2017, a quantidade de estudantes ingressantes pelos sistemas universal e de cotas são equivalentes.

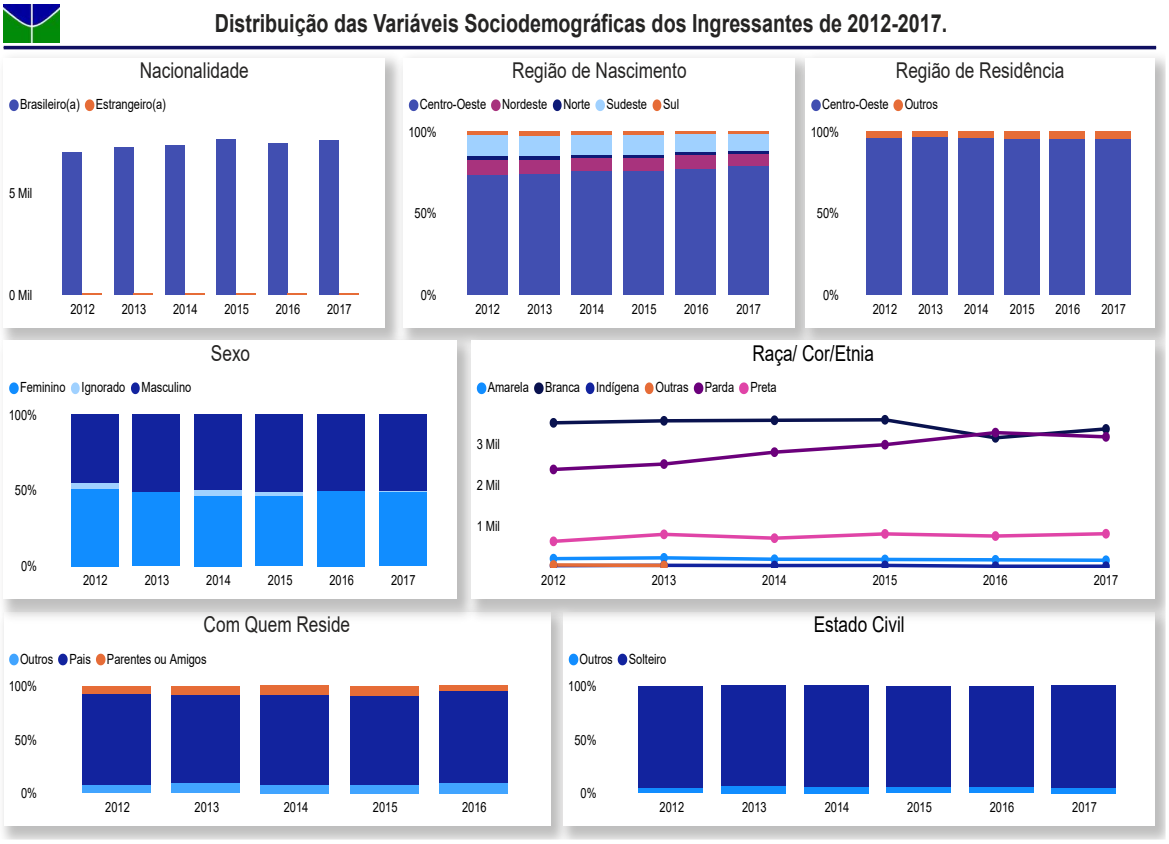
[1] A lei de cotas (Lei Nº 12.711) publicada em 29 de agosto de 2012, obriga as instituições federais de ensino superior a cumprir até 2016 a reserva de 50% das vagas para estudantes que cursaram o ensino médio em escolas públicas. Para mais informações acesse: `http://www.planalto.gov.br/CCIVIL_03/_Ato2011-2014/2012/Lei/L12711.htm`

Houve uma considerável mudança na modalidade de ingresso, que até 2013 era representada em sua maioria pela categoria vestibular, ocorrendo um substancial aumento para as modalidades PAS e ENEM. Quanto a modalidade outros, resultado do agrupamento das categorias (apêndice A - Tabela 28), sua representatividade foi reduzida ao longo do estudo.

Para campus, turno e curso escolhido, nota-se que não ocorreram mudanças ao longo do estudo. Cabe ressaltar que o perfil descritivo dos ingressantes nessas variáveis deve-se a oferta de vagas na instituição. Os cursos foram agrupados segundo o critério das grandes áreas do conhecimento utilizado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), o qual utiliza oito categorias.

Ao longo do estudo também foi verificado um aumento do número de alunos ingressantes portadores de necessidades especiais, que representa mudanças diretamente relacionadas a políticas afirmativas.

Figura 4: Perfil sociodemográfico dos ingressantes.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

A Figura 4 apresenta variáveis relacionadas ao perfil sociodemográfico. Mais de 99% dos ingressantes são brasileiros(as), cerca de 73% nasceram na região centro oeste, com destaque para as unidades federativas do DF e Goiás com aproximadamente 66% e 6% respectivamente dos ingressantes.

Acima de 95% dos ingressantes possuem residência no Centro-Oeste com relevância para o DF e Goiás que acomodam em torno de 88% e 6% respectivamente dos ingressantes da universidade.

No que se refere ao sexo nota-se que não houve mudanças no perfil dos estudantes ao longo dos anos. A distribuição dos estudantes segundo sexo tem sido equitativa, em torno dos 50%.

Em relação ao estado civil mais de 93%, em todos os anos, declararam-se solteiros. As categorias da variável raça/cor/etnia, revelam dois pontos importantes: 1) a alta diferença entre as proporções das categorias e 2) a tendência para a redução do número de alunos brancos e aumento dos alunos pardos e pretos.

A variável com quem reside não sofreu mudanças significativas ao longo do estudo apontando que mais de 82% dos ingressantes vivem com os pais, cerca de 8% vivem com parentes ou amigos(as) com exceção do ano de 2017 que foi aproximadamente 5%.

Pela Tabela 6 verifica-se que: embora as principais medidas descritivas para a idade dos ingressantes como, média, mediana, 1° quartil (25% dos valores estão abaixo dessa observação) e 3° quartil (75% dos valores estão abaixo dessa observação), não demonstraram grandes modificações, o Coeficiente de Variação (C.V), que analisa dispersão em termos relativos, assim, quanto menor for seu valor mais homogêneos estão os dados, indica uma dispersão média (maior que 15% e menor que 30%). Isso deve-se em grande parte pela presença de valores extremos nos dados (*outliers*) observados a partir de um certo limiar.

Quadro 6: Medidas estatísticas da idade dos ingressantes - UnB, 2012 - 2017.

Ano	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo	Variância	C.V
2012	15,00	17,00	18,00	19,72	20,00	65,00	25,97	0,26
2013	14,00	17,00	18,00	20,59	21,00	80,00	38,01	0,30
2014	14,00	17,00	18,00	20,14	20,00	74,00	33,23	0,29
2015	14,00	18,00	18,00	20,45	20,00	61,00	34,22	0,29
2016	0,00	18,00	18,00	20,31	20,00	65,00	33,16	0,28
2017	15,00	18,00	18,00	20,06	20,00	67,00	29,77	0,27

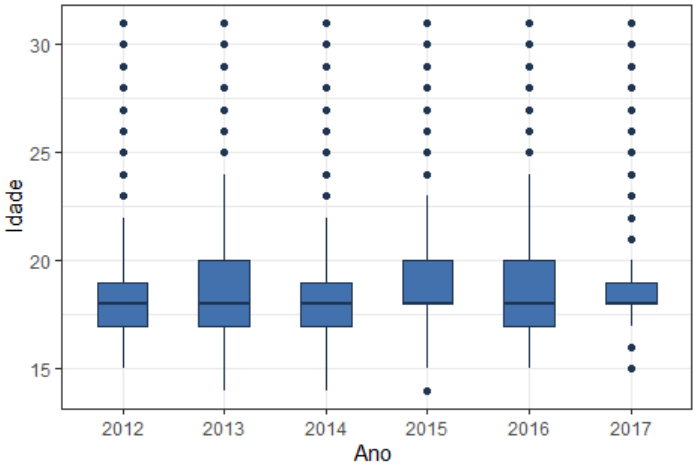
Para encontrar a observação que limita o número de *outliers* e representa de forma mais homogênea os dados foram calculados os quantis (Tabela 7) do período em análise (2012-2017).

Quadro 7: Quantis da idade dos ingressantes - UnB, 2012 - 2017.

0,25	0,5	0,75	0,85	0,9	0,95	0,99
17	18	20	23	26	32	47

Para verificar o comportamento da variável idade os dados foram limitados entre as observações iguais ou acima de 14, que representa a idade mínima dos ingressantes nos anos em estudo salvo o ano de 2016, e iguais ou abaixo de 32, cujo 95% das observações estão abaixo desse valor. A Figura 5 representa as principais medidas descritivas desse novo conjunto de dados.

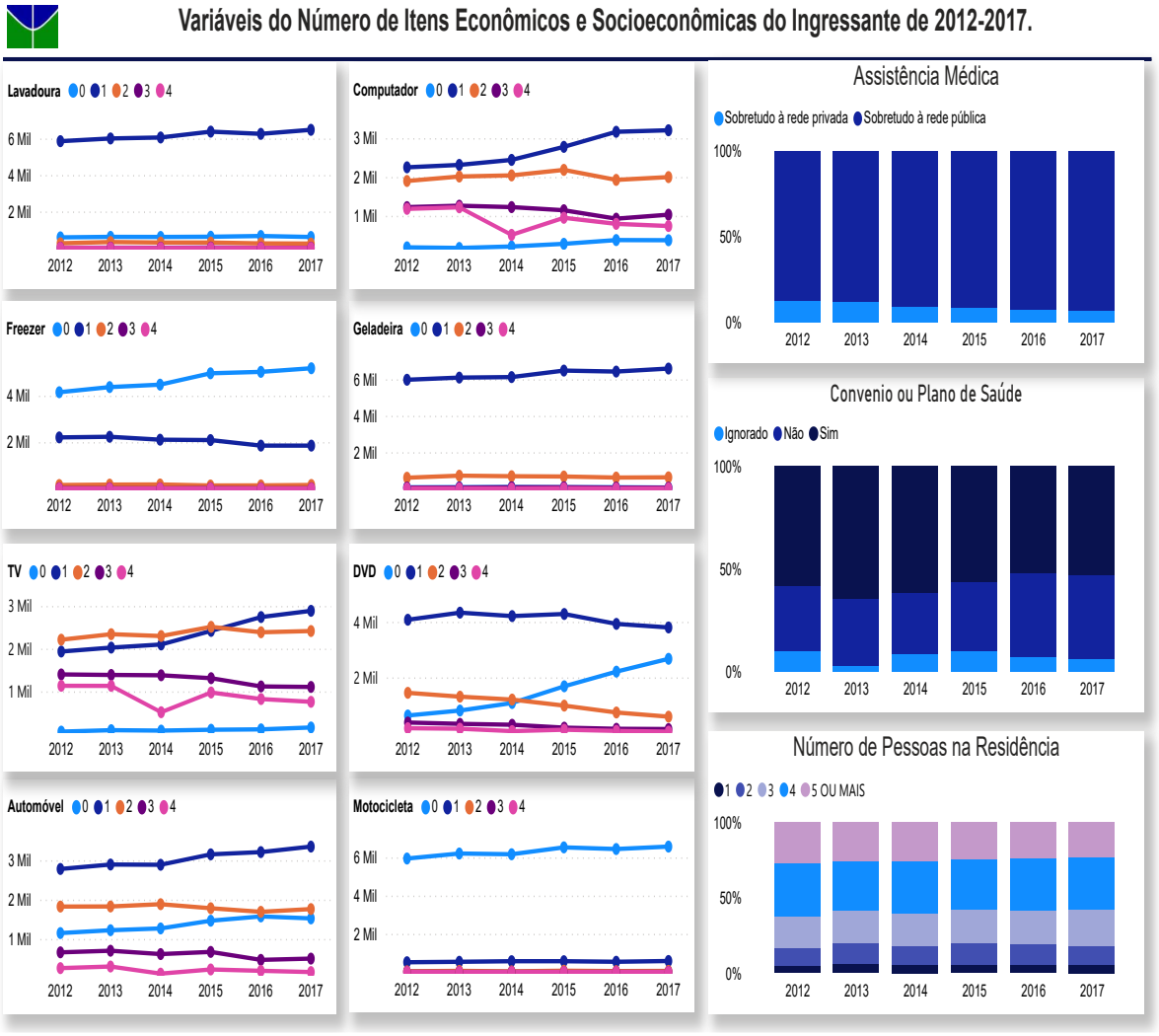
Figura 5: Distribuição da idade dos ingressantes - UnB, 2012 - 2017.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Observa-se diferenças no intervalo interquartil e limites inferiores e superiores do boxplot indicando que a exclusão dos *outliers* influenciou menos as medidas de posição e consequentemente foi possível uma melhor representação dos dados apontando para mudanças nessa variável, como o aumento da idade entre os ingressantes a partir de 2014.

Figura 6: Perfil socioeconômico dos ingressantes.

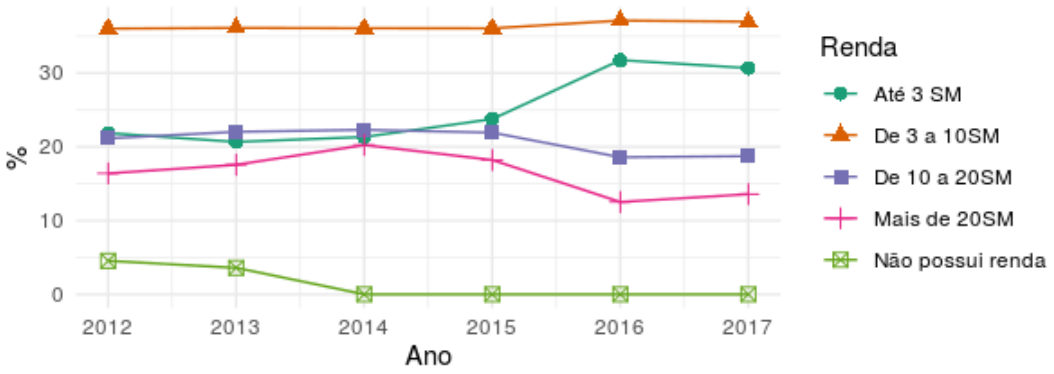


Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

A análise da posse de bens ao longo dos anos (Figura 6) traduz a mudança no perfil socioeconômico dos estudantes. Vale observar que alguns itens, como DVD e TV, sofreram forte queda no consumo da população em geral, o que explica o aumento da proporção de estudantes que não possuem esses itens. Por outro lado, itens que normalmente são prioridade na vida doméstica das famílias tais como geladeira, computador e máquina de lavar e que não sofreram grandes alterações no consumo com a inserção de novas tecnologias evidenciam essa alteração no perfil de forma mais clara, indicando um aumento de alunos que possuem 0 ou 1 e uma redução entre aqueles que tem mais de 1, isto está em consonância com a mudança na renda familiar dos ingressantes.

Quanto a assistência médica, houve um aumento, ao longo do estudo, entre aqueles que buscam a rede pública de saúde. Já convênio ou plano de saúde acima de 30% dos ingressantes não possuem, tendo um aumento expressivo nos últimos dois anos do estudo, acima de 40%.

Figura 7: Renda dos ingressantes - UnB, 2012 - 2017.

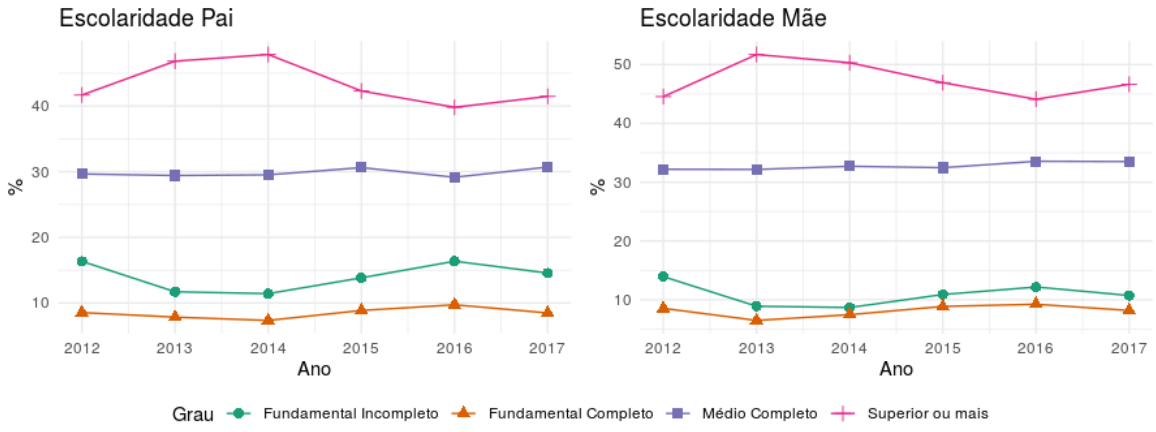


Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Na Figura 7 observa-se uma variação na renda dos ingressantes. Merece destaque as alterações quantitativas que ocorreram a partir do ano de 2014 nas categorias até 3, de 10 a 20 e mais de 20 salários-mínimos. Outra variável que está relacionada ao nível econômico familiar é a escolaridade dos pais.

A fim de melhor representar os níveis da escolaridade do pai/mãe (Figura 8) foram feitos os seguintes agrupamentos das categorias (apêndice A - Tabela 30): não sabe ler nem escrever e ensino fundamental incompleto em fundamental incompleto; ensino fundamental completo e ensino médio incompleto em fundamental completo ; ensino médio completo e ensino superior incompleto em médio completo ; ensino superior completo e pós-graduação em superior ou mais.

Figura 8: Escolaridade dos pais dos ingressantes - UnB, 2012 - 2017.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

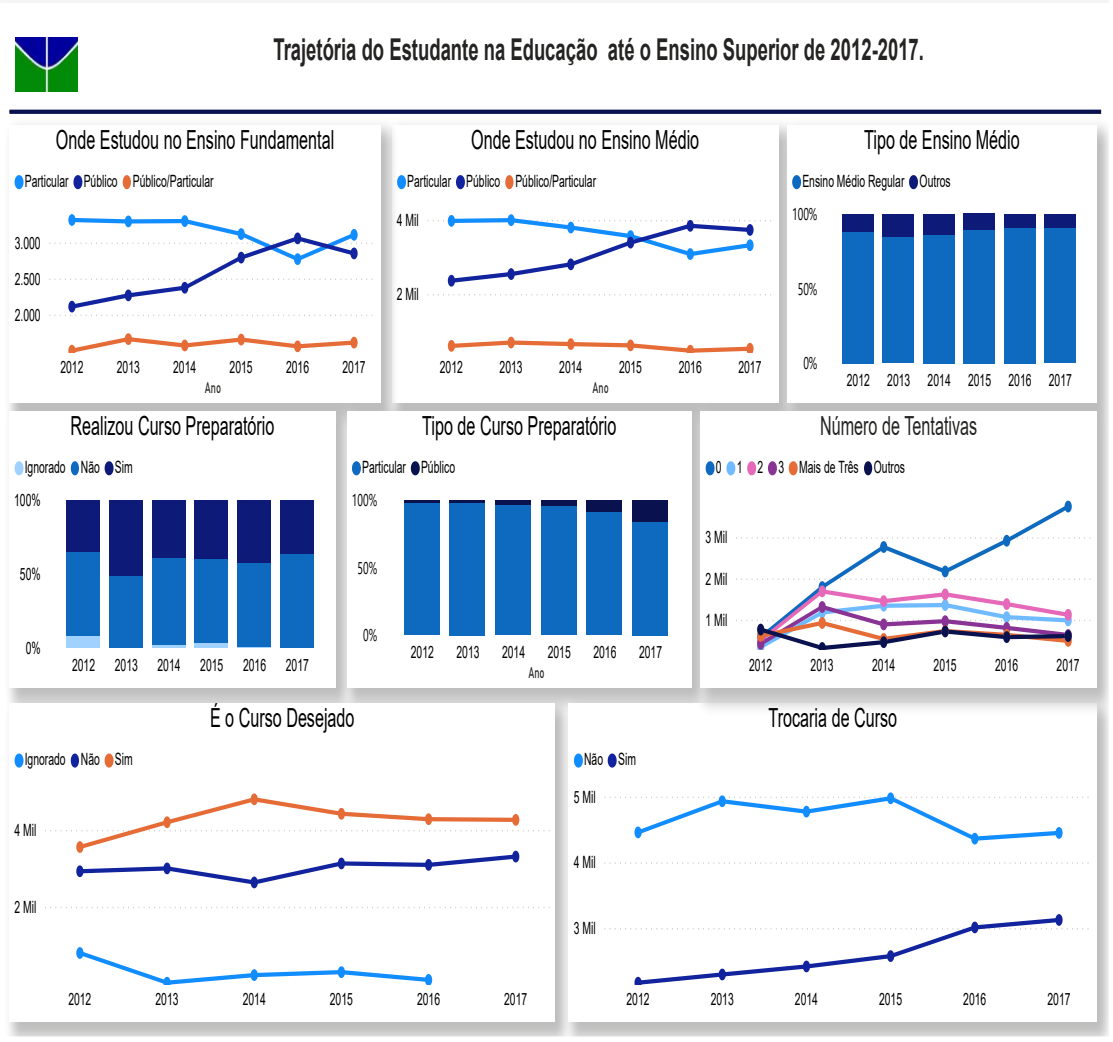
O nível de escolaridade dos pais dos ingressantes da UnB (Figura 8) mostra que mesmo após ações afirmativas para maior equidade no acesso à universidade, ainda continua elevado o número de estudantes cujos pais possuem ensino superior completo ou pós-graduação, o que representa ainda uma certa elitização familiar nos ingressantes. Esta variável diretamente relacionada com a renda é considerada um dos maiores determinantes do sucesso educacional dos filhos (BARROS et al, 2001).

A análise descritiva indica uma mudança no perfil econômico do ingressante da UnB revelando o surgimento de novos grupos dentro da universidade. Essa mudança foi verificada para todas as categorias das variáveis analisadas. No início do período, observa-se um perfil de estudantes com mais alta renda, maior proporção de brancos e pais com mais elevada escolaridade. Esse perfil começa a mudar, sobretudo, a partir de 2016, quando é possível verificar o acesso de estudantes com renda familiar menor que 3 SM.

Para os níveis das variáveis referentes a trajetória de ensino dos ingressantes (Figura 9) de modo análogo ao que foi realizado com as variáveis escolaridade dos pais foram feitos alguns agrupamentos (apêndice A - Tabela 31) com o objetivo de representar melhor essas categorias.

No que se refere ao ensino fundamental e ensino médio foram agrupadas as categorias: sobretudo em escolas públicas, escolas particulares com bolsa e escolas particulares na categoria Público/Particular, e somente em escolas particulares com bolsa e escolas particulares na categoria Particular. Os demais agrupamentos são autoexplicativos com o preenchimento das células.

Figura 9: Perfil da trajetória estudantil dos ingressantes.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

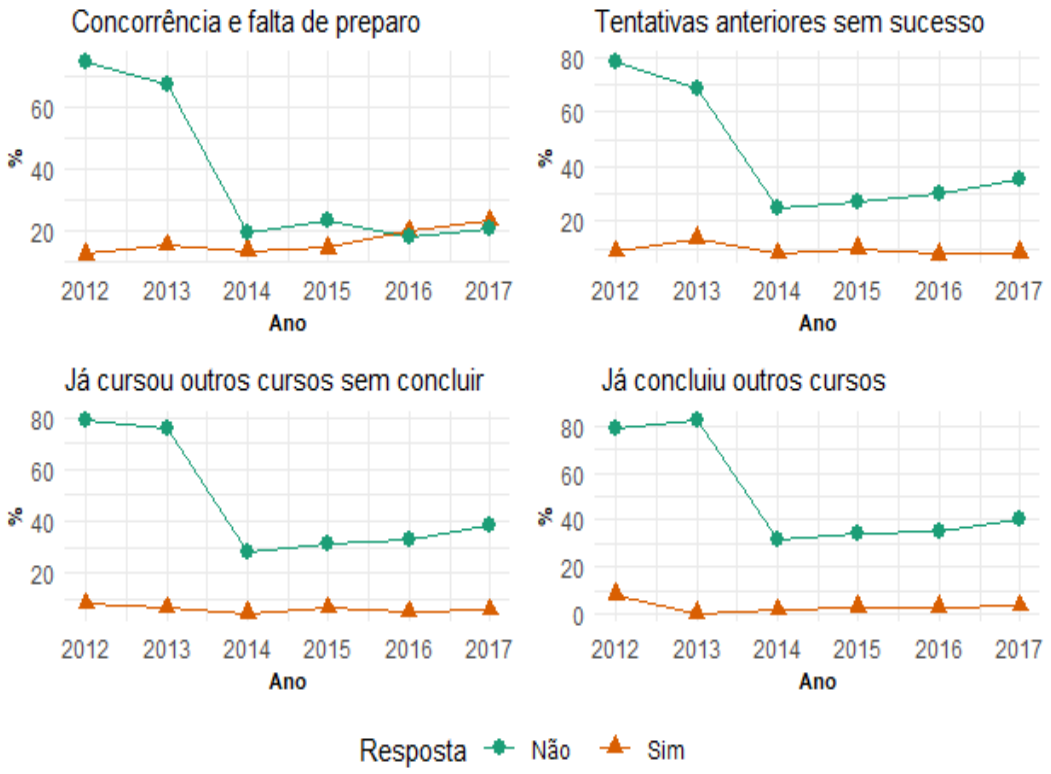
A trajetória do estudante (Figura 9) revela um crescente aumento dos alunos provenientes das escolas públicas a partir do ano de 2012. Esses dados corroboram com o esperado devido à forte implementação de políticas voltadas para inserção dos alunos do sistema público nas instituições de Ensino Superior realizada nos últimos anos.

Quanto a forma como se deu o ingresso na Universidade a partir de 2013 observa-se uma redução no número de alunos que precisaram realizar curso preparatório para vestibular e entre esses houve redução no número de estudantes que pagavam pelo cursinho e um aumento no acesso da categoria público ou gratuito.

No que se refere ao tipo de ensino foi verificado um aumento na categoria ensino médio regular, durante o estudo. O número de tentativas para acesso ao ensino superior, diminuiu e muitos desses alunos passaram a conquistar sua vaga na primeira tentativa. No ano de 2017 o percentual nessa categoria chegou a quase 50%.

Cerca de 40% dos estudantes não ingressaram no curso desejado. Verifica-se uma tendência crescente para essa categoria a partir de 2014. Quanto a trocar de curso, no início do estudo esse percentual era em torno de 30% ao final do estudo esse percentual ficou em torno de 40% indicando um aumento significativo.

Figura 10: Motivos do não ingresso no curso desejado de 2012 - 2017.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

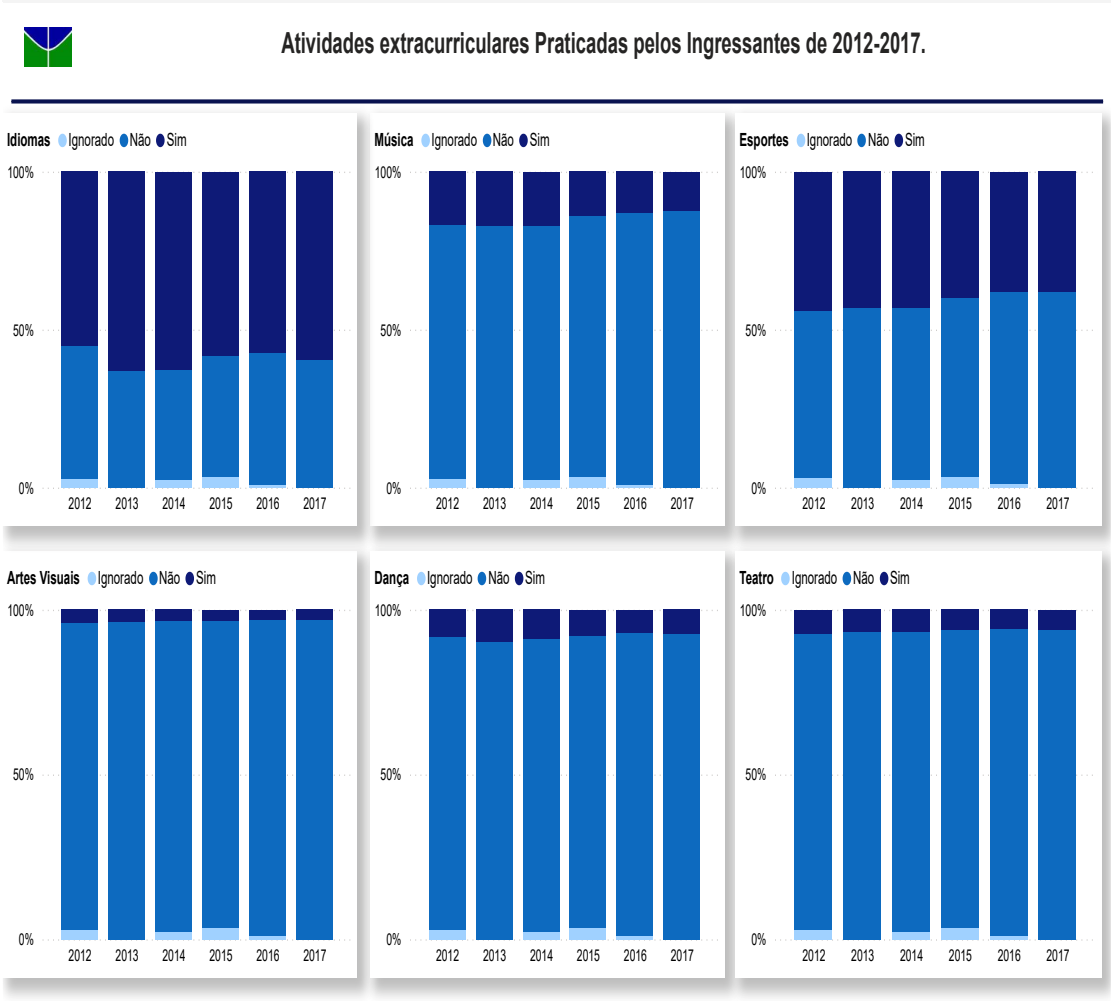
Dentre os motivos para o não ingresso no curso desejado (Figura 10) concorrência e falta de preparo apresentou tendência de aumento. Em 2012, cerca de 12% dos ingressantes evoluindo para mais de 23% em 2017. Tentativas anteriores sem sucesso e já cursou outros cursos sem concluir foi motivo para cerca de 8% e 5% dos ingressantes respectivamente.

Um dos motivos menos significativos em termos de observações para o não ingresso no curso desejado foi já conclui outros cursos que ficou abaixo de 4% desde 2013.

As atividades extracurriculares (Figura 11) indicam que os ingressantes disponibilizam mais competências para aquelas que direta ou indiretamente se relacionem com o mercado de trabalho.

Acima de 50% dos ingressantes fazem um curso de línguas e cerca de 40% praticam alguma atividade esportiva. Música também é uma categoria que aparece com representatividade, porém nos últimos anos sofreu uma queda. Esses percentuais caem consideravelmente ao analisar as demais categorias.

Figura 11: Perfil extracurricular do ingressante.

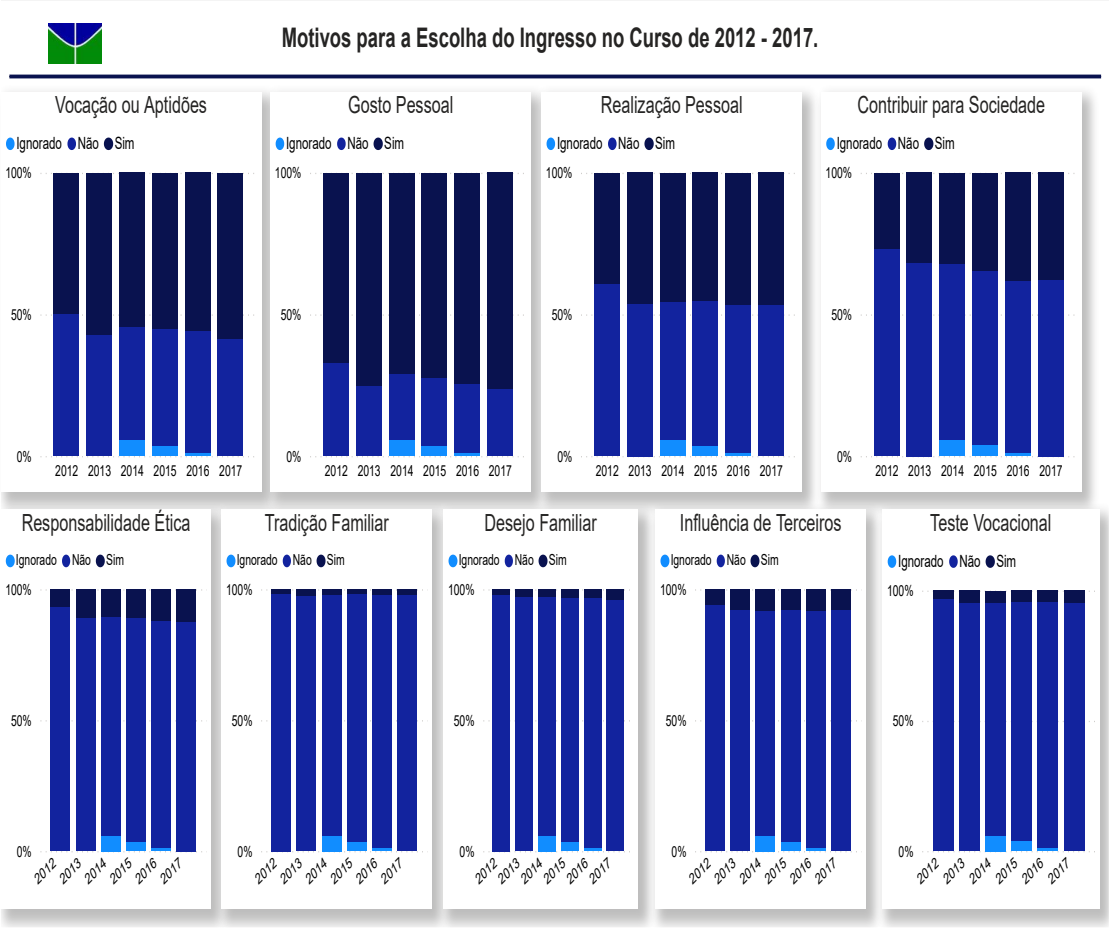


Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Dentre os motivos para ingresso no curso (Figura 12) gosto pessoal, vocação ou aptidões pessoais, realização pessoal e possibilidade de contribuir para a sociedade são os motivos mais relevantes para os ingressantes na escolha do curso com cerca de 70%, 55%, 45% e 30% respectivamente, sendo verificado a partir de 2014 a tendência de aumento nessas categorias.

Ao longo do estudo, nota-se que houve um aumento nos motivos responsabilidade ética e desejo familiar, porém ainda com percentuais baixos, assim como tradição familiar, influência de terceiros e teste vocacional.

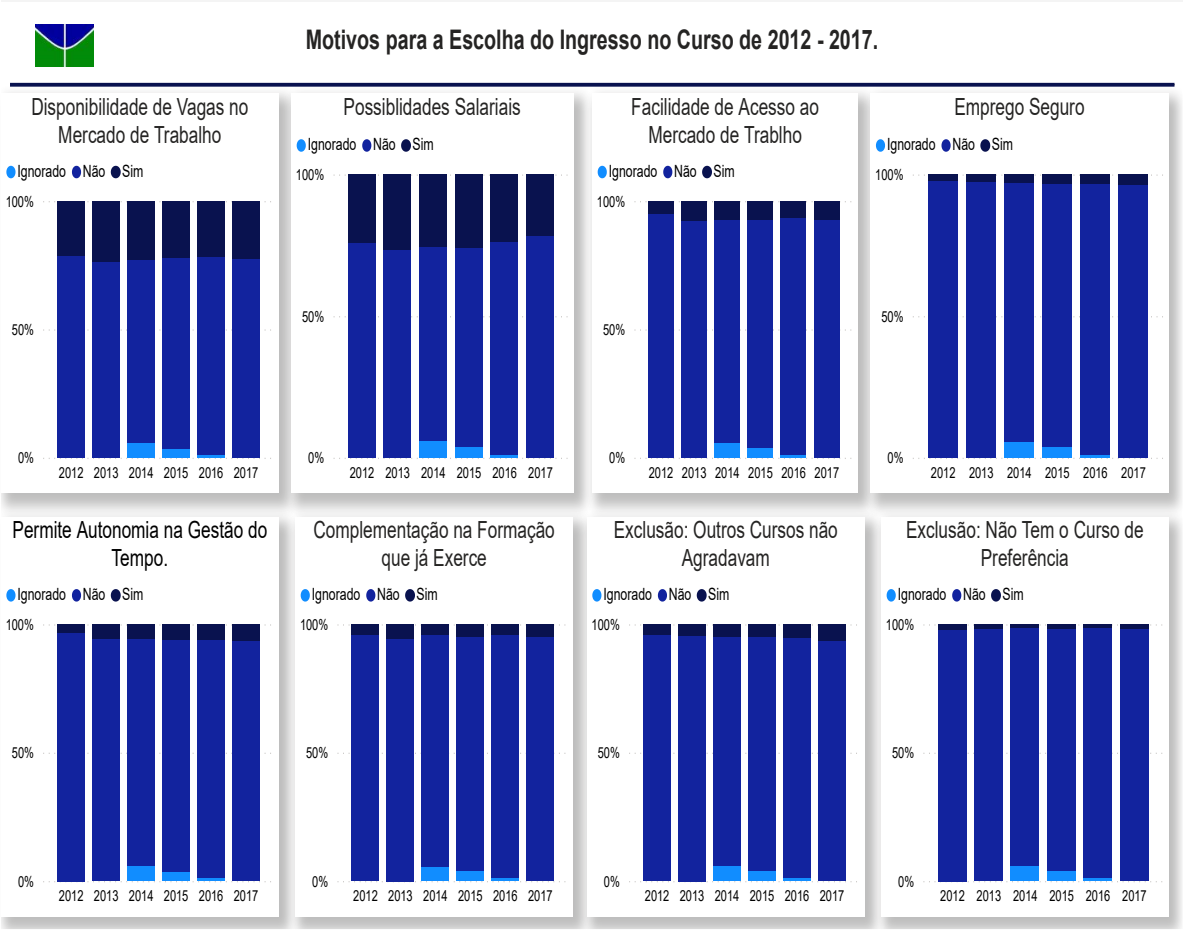
Figura 12: Motivos para o ingresso no curso.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Acima de 20% dos ingressantes (Figura 13) consideraram disponibilidade de vagas no mercado e possibilidades salariais como motivos relevantes para a escolha do curso. As demais categorias ficaram abaixo de 10%.

Figura 13: Motivos para o ingresso no curso.

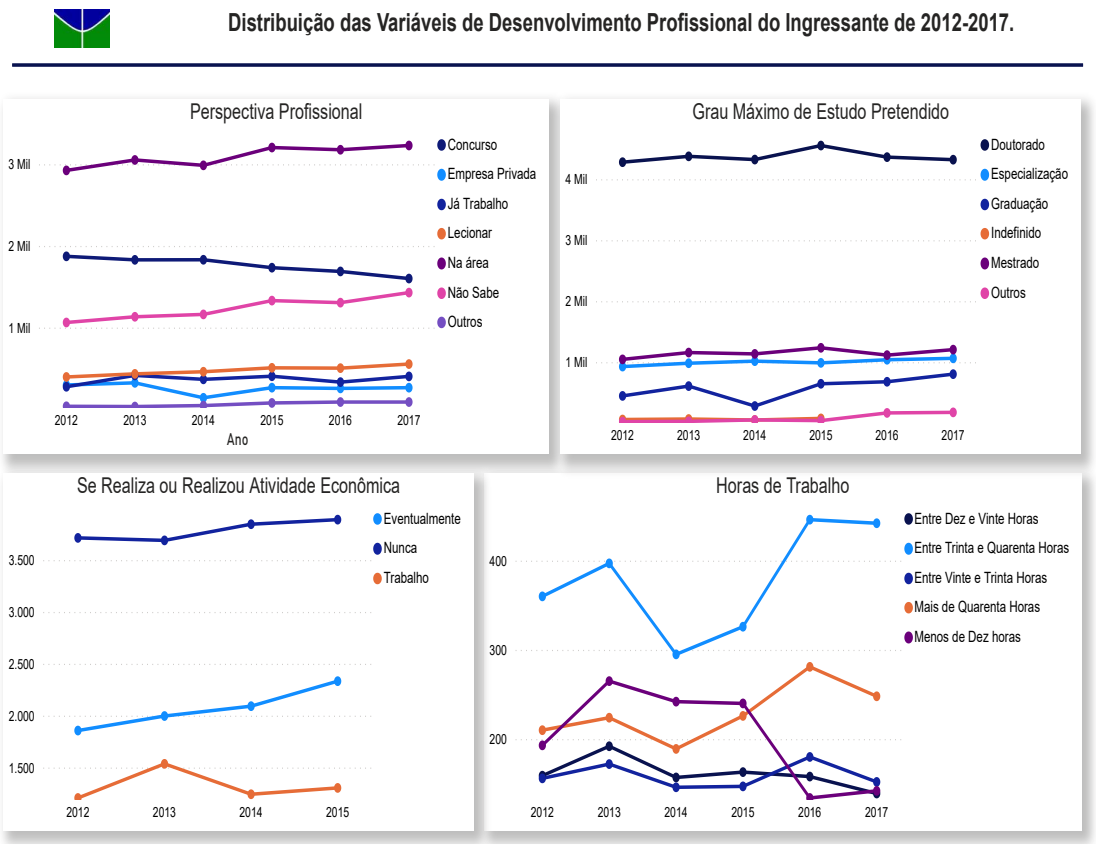


Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Pode-se concluir que embora exista um forte apelo pela sociedade para o estudante ao escolher um curso superior já direcioná-lo para o mercado de trabalho (SEVERINO, 2008), o que se verifica é a forte influência que fatores pessoais exercem nessa escolha.

Para perspectiva profissional (Figura 14) nota-se, ao longo do estudo, uma tendência de aumento na categoria não sabe e uma tendência de redução entre aqueles que planejam prestar concurso. Em torno de 42% dos ingressantes planejam exercer atividade na área de formação com tendência de aumento. Lecionar e já trabalho, resultados dos agrupamentos (Apêndice A - Tabela 32), correspondem a cerca de 6% e 5% dos ingressantes respectivamente.

Figura 14: Desenvolvimento profissional e expectativas



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Quanto as expectativas acadêmicas (Figura 14) acima de 97% dos ingressantes pretendem completar o ensino superior ou fazer uma pós-graduação. Em torno de 60% dos ingressantes pretende fazer um doutorado.

No que se refere as atividades econômicas cerca de 50% dos ingressantes não trabalham. A categoria eventualmente, resultado do agrupamento dos níveis: não trabalho no momento e exerço atividades remuneradas eventualmente, corresponde a aproximadamente 30% do ingressantes. Verifica-se um aumento nessas categorias nos últimos dois anos do estudo acompanhada de uma redução de ingressantes que exercem alguma atividade econômica.

Entre aqueles que trabalham houve a partir de 2014 um aumento no intervalo de 30 a 40 horas e mais de 40 horas trabalhadas, nos últimos dois anos do estudo (2016 e 2017) esses percentuais somados chegaram a mais de 60%, nos demais anos ficaram em torno de 50%.

5 Análise Bivariada

Tendo em visto a análise descritiva um ponto importante que chamou a atenção foi o elevado número de estudantes que não ingressou no curso desejado.

Durante o período estudado, em torno 40% dos alunos não ingressaram no curso desejado, com exceção do ano de 2014 (34,42%). Dentre os motivos para não ingressar no curso desejado destaca-se concorrência/falta de preparo e tentativas anteriores sem sucesso.

Com o objetivo de verificar possíveis fatores que possam explicar o não ingresso no curso desejado foram analisadas potenciais variáveis que se associariam a essa resposta. O percentual de estudantes que não ingressaram no curso desejado, valor da estatística qui-quadrado e os respectivos p-valores fornecidos pelo *software SAS* estão apresentadas nas tabelas a seguir.

Quadro 8: Teste de associação entre as variáveis explicativas de ingresso e a variável respota "Não ingressou no curso desejado" - UnB, 2012 - 2017.

Variáveis	**Categorias**	**Não ingresso no curso desejado (%)**	**Estatística**	**p-valor**
Semestre	2012	16,17	163,06	<,0001
	2013	16,58		
	2014	14,57		
	2015	17,3		
	2016	17,09		
	2017	18,29		
Sistema de Ingresso	Sistema de Cotas	31,23	48,33	<,0001
	Sistema Universal	68,77		
Modalidade de Ingresso	ENEM	18,48	510,46	<,0001
	PAS	20,84		
	Vestibular	53,04		
	Outros	7,63		
Campus	Darcy Ribeiro	81,86	381,52	<,0001
	Ceilândia	8,05		
	Gama	5,2		
	Planaltina	4,89		
Turno	Diurno	69,6	104,29	<,0001
	Noturno	30,4		
Curso	Ciências Exatas e da Terra	9,93	486,2	<,0001
	Ciências Biológicas	1,62		
	Engenharias	13,08		
	Ciências da Saúde	17,32		
	Ciências Agrárias	10,95		
	Ciências Sociais Aplicadas	23,53		
	Ciências Humanas	11,15		
	Linguística, Letras e Artes	12,42		

Quadro 9: Teste de associação entre as variáveis explicativas sociodemográficas e a variável resposta “Não ingressou no curso desejado” - UnB, 2012 - 2017.

Variáveis	**Categorias**	**Não ingresso no curso desejado (%)**	**Estatística**	**p-valor**
Sexo	Masculino	48,77	39,61	<,0001
	Feminino	51,23		
Nacionalidade	Brasileiro(a)	99,87	7,32	0,0068
	Estrangeiro(a)	0,13		
UF de Nascimento	Norte	1,97	17,43	0,0016
	Nordeste	7,96		
	Sudeste	10,85		
	Sul	2,14		
	Centro-Oeste	77,07		
Estado Civil	Solteiro(a)	93,54	24,27	<,0001
	Outros	6,46		
Raça/Cor/Etnia	Amarela	2.89	13,67	0,0084
	Branca	46,30		
	Indígena	0,54		
	Parda	39,73		
	Preta	10,54		
UF de Residência	Norte	0,31	97,86	<,0001
	Nordeste	0,92		
	Sudeste	1,67		
	Sul	0,18		
	Centro-Oeste	96,91		
Com quem Reside?	Pais	82,51	48,91	<,0001
	Com parentes	7,87		
	Outros	9,62		
Necessidade especial	Sim	4,57	7,53	0,0061
	Não	95,43		

Quadro 10: Teste de associação entre as variáveis explicativas socioeconômicas e a variável resposta “Não ingressou no curso desejado” - UnB, 2012 - 2017.

Variáveis	**Categorias**	**Não ingresso no curso desejado (%)**	**Estatística**	**p-valor**
Renda Mensal da Família	Até 3 SM	25.33	62,4683	<,0001
	De 3 a 10 SM	37,93		
	De 10 a 20 SM	21,69		
	Mais de 20 SM	15,05		
Escolaridade do Pai	Fundamental Incompleto	14,47	48,18	<,0001
	Fundamental Completo	8,63		
	Médio Competo	30,98		
	Superior ou Mais	41,37		
	Não Sabe	4,55		
Escolaridade da Mãe	Fundamental Incompleto	11,21	83,25	<,0001
	Fundamental Completo	8,07		
	Médio Completo	34,73		
	Superior ou Mais	44,98		
	Não Sabe	1,01		

Quadro 11: Teste de associação entre as variáveis explicativas da trajetória estudantil e a variável respota "Não ingressou no curso desejado" - UnB, 2012 - 2017.

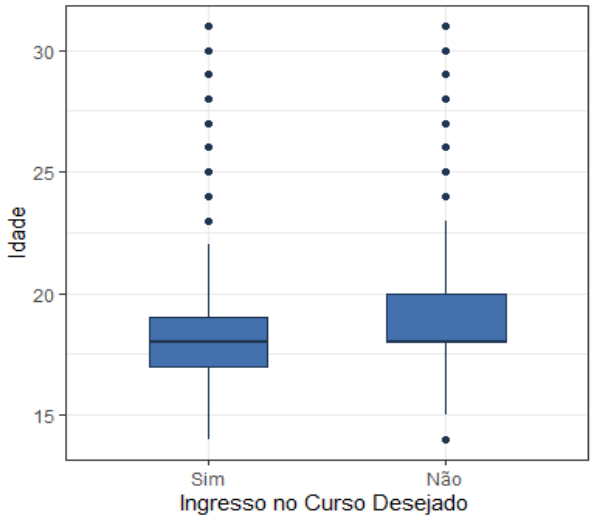
Variáveis	Categorias	Não ingresso no curso desejado (%)	Estatística	p-valor
Ensino Fundamental	Público	34,19	51,76	<,0001
	Público/Particular	23,42		
	Particular	42,4		
Ensino Médio	Público	40,76	78,12	<,0001
	Público/Particular	9,16		
	Particular	50,08		
Tipo de Ensino Médio	Ensino Médio Regular	88,93	6,31	0,012
	Outros	11,07		
Curso Preparatório?	Sim	42,97	30,93	<,0001
	Não	57,03		
Tipo de Curso Preparatório	Público	5,34	54,03	0,0201
	Particular	94,66		
Número de Tentativas	Primeira	23,05	2514,3	<,0001
	Uma	13,81		
	Duas	21,39		
	Três	15,54		
	Mais de três	13,82		
	Outros	12,4		
Trocaria de Curso?	Sim	52,13	3457,76	<,0001
	Não	47,87		

Quadro 12: Teste de associação entre as variáveis explicativas de movimento do ingressante e a variável respota "Não ingressou no curso desejado" - UnB, 2012-2017.

Variáveis	Categorias	Não ingresso no curso desejado (%)	Estatística	p-valor
Grau de Estudo Pretendido	Graduação	9,09	70,77	<,0001
	Doutorado	59,03		
	Especialização	14,00		
	Mestrado	15,82		
	Indefinido	0,66		
	Outros	1,39		
Atividade Econômica	Nunca	49,64	84,56	<,0001
	Eventuamente	29,79		
	Trabalho	20,57		
Perspectiva Profisional	Não Sabe	18,49	239,1	<,0001
	Dentro da Área	39,14		
	Lecionar	6,04		
	Empresa Privada	3,24		
	Concurso Público	26,42		
	Trabalho	5,63		
	Outros	1,04		
Horas de Trabalho	Menos de 10	17,63	7,79	0,0995
	Entre 10 e 20	13,26		
	Entre 20 e 30	14,02		
	Entre 30 e 40	34,81		
	Mais de 40	20,28		

Os testes qui-quadrados realizados entre as variáveis explicativas e a variável resposta, não ingresso no curso desejado na Tabela 8, Tabela 9, Tabela 10, Tabela 11 e Tabela 12 apresentam o mesmo resultado: Há evidências estatísticas para rejeitar H_0 e afirmar que existe associação entre o não ingresso no curso desejado e as variáveis respostas. O nível de significância adotado para tal conclusão foi de 0,1.

Figura 15: Distribuição da idade segundo o ingresso no curso desejado.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

A Figura 15 auxilia a investigação da associação entre não ingressar no curso desejado e a idade, porém será feito um modelo regressão logística simples no *software R* para verificar quantitativamente essa relação. A variável dependente é o não ingresso ou ingresso (1 ou 0) no curso desejado, associando-se com a idade dos ingressantes, os resultados estão apresentados na Tabela 13.

Quadro 13: Parâmetros estimados para o modelo proposto.

	Estimativa	Erro padrão	Valor Z	$Pr(>\|z\|)$
Intercepto	-0,830620	0,036046	-23,04	$< 2e-16^{***}$
Idade	0,023978	0,001711	14,01	$< 2e-16^{***}$

No modelo proposto a variável de interesse idade obteve um coeficiente de 0,023978. Pelo fato de ser positivo informa que quando a idade se eleva, elevam-se as chances de não ingressar no curso desejado. Nota-se que há significância estatística a $\alpha = 0,001$ na utilização da variável idade para o modelo regressão logística proposto.

A análise dos resultados indica associação entre a variável resposta, ingressou no curso desejado (sim, não) e as variáveis explicativas. A escolha das variáveis para o modelo de regressão logística levará em consideração essas associações e a literatura referente as variáveis em estudo. Um estudo mais completo será trabalhado na próxima seção.

6 Aplicação do modelo de regressão logística

A análise univariada do perfil do ingressante da UnB, foi relevante para identificar mudanças ocorridas em algumas variáveis ao longo do estudo, bem como verificar a ausência de alterações em outras.

Considerando o objetivo principal do estudo foi realizada a análise bivariada para verificar a presença de associação entre a variável resposta não ingresso no curso desejado e as demais variáveis estudadas na análise univariada. Os p-valores baixos indicaram a associação entre a variável dependente e as independentes.

Para a modelagem da variável curso que foi agrupada a princípio por grandes áreas do conhecimento, fez-se uma análise da demanda e da nota de corte dos processos seletivos anteriores resultando na classificação segundo o prestígio.

- Maior prestígio: Medicina, Direito, Relações Internacionais, as Engenharias, Odontologia, Psicologia, Comunicação Social e Ciências Econômicas.
- Baixo prestígio: Biblioteconomia, Pedagogia, Serviço Social, Letras e Matemática
- Médio prestígio: Demais cursos.

Caso seja necessário outros agrupamentos serão realizados, a fim de não desconsiderar uma variável que possa ser relevante para o modelo sob outra perspectiva de análise. As variáveis selecionadas são:

Quadro 14: Variáveis explicativas e seus respectivos níveis

Variável	**Níveis**
X_{i1} **- Ano de Ingresso**	1 - 2012 2 - 2013 3 - 2014 4 - 2015 5 - 2016 6 - 2017
X_{i2} **- Sistema de ingresso**	1 - Cotas 2 - Universal
X_{i3} **- Modalidade de ingresso**	1 - Outros (Diploma de Curso Superior, Vagas remanescentes, Transferência) 2 - ENEM 3 - PAS 4 - Vestibular
X_{i4} **- Campus**	1 - Darcy-Ribeiro 2 - Ceilândia 3 - Gama 4 - Planaltina
X_{i5} **- Turno**	1 - Diurno 2 - Noturno
X_{i6} **- Curso**	1 - Prestígio alto 2 - Prestígio baixo 3 - Prestígio médio
X_{i7} **- Sexo**	1 - Masculino 2 - Feminino
X_{i8} **- Estado Civil**	1 - Outros (Casado(a), Divorciado(a), Separado(a), União Estável, Viúvo(a)) 2 - Solteiro
X_{i9} **- Raça/Cor/Etnia**	1 - Amarela 2 - Branca 3 - Indígena 4 - Parda 5 - Preta 6 - Outros
X_{i10} **- Renda Mensal**	1 - Até 3 SM 2 - De 3 a 10 SM 3 - De 10 a 20 SM 4 - Mais de 20 SM 5 - Não sabe informar
X_{i11} **- Escolaridade do Pai**	1 - Fundamental Incompleto 2 - Fundamental Completo 3 - Médio Completo 4 - Superior ou Mais 5 - Não Sabe Informar
X_{i12} **- Escolaridade da Mãe**	1 - Fundamental Incompleto 2 - Fundamental Completo 3 - Médio Completo 4 - Superior ou Mais 5 - Não Sabe Informar
X_{i13} **- Ensino Médio**	1 - Público 2 - Público/Particular 3 - Particular
X_{i14} **- Realizou Curso Preparatório**	1 - Sim 2 - Não
X_{i15} **- Número de Tentativas**	1 - Primeira 2 - Uma 3 - Duas 3 - Três 4 - Mais de três 5 - Já realizei, sem concluir, outro(s) curso(s) de graduação na UnB 6 - Já conclui outro(s) curso(s) de graduação na UnB;
X_{i16} **- Trocaria de Curso**	1 - Sim 2 - Não
X_{i17} **- Perspectiva Profissional**	1 - Na área de formação 2 - Não sabe informar 3 - Lecionar(Ensino médio / Superior) 4 - Empresa privada 5 - Concurso 6 - Já trabalho 7 - Outros
X_{i18} **- Idade**	Anos completos (quantitativa)

Os dados foram organizados de maneira a serem mantidas somente as variáveis explicativas de interesse (X_{i1} a X_{i18}) e a variável resposta Y_i. Foram eliminadas as observações que apresentassem pelo menos um valor NA em alguma das variáveis, assim como as observações com valor ignorado.

Após alterações realizadas na base de dados original, a nova base de dados apresentou 38168 observações com 19 variáveis: 1 resposta, 17 explicativas (categóricas) e 1 explicativa quantitativa. Após investigar os níveis das variáveis selecionadas foi feito um primeiro ajuste do modelo de regressão logística múltiplo considerando todas as variáveis explicativas, sendo verificado que algumas das variáveis não eram significativas optou-se por reagrupar algumas categorias. Os novos níveis da variável explicativa que apresentou resultados significativos são:

- X_{i11}**:** Escolaridade do pai;
  1. Superior ou mais;
  2. Até o ensino médio: Fundamental incompleto, fundamental completo e médio completo;
  3. Não sabe informar.
- X_{i12}**:** Escolaridade da mãe;
  1. Superior ou mais;
  2. Até o ensino médio: Fundamental incompleto, fundamental completo e médio completo;
  3. Não sabe informar.

Para viabilizar a validação do modelo, foi realizado um processo de amostragem aleatório, no *software SAS*, dividindo os dados em dois grupos, um para o desenvolvimento do modelo com 23357 observações e outro para a validação com 14811 observações.

6.1 Modelo completo

Foi testado um modelo inicial, considerando as 18 variáveis explicativas, $X_{i1}, \cdots, X_{i18}$, apresentadas no Quadro 12, assim X_{i2} é a variável sistema de ingresso com níveis: i=1 (Cotas), i=2 (Universal). De forma semelhante as demais variáveis foram representadas com seus respectivos níveis em relação a variável resposta não ingresso no curso desejado.

6.2 Seleção das variáveis explicativas

Para a seleção das variáveis explicativas que estarão presentes no modelo final, foram utilizados os métodos de seleção, automáticos e manuais vistos na seção 2.8.

6.2.1 Seleção automática

Os três métodos de seleção automática, *backward, forward* e *stepwise*, selecionaram as mesmas variáveis explicativas para o modelo: Ano de ingresso, modalidade de ingresso, campus, curso, estado civil, renda, escolaridade dos pais, tipo de instituição de ensino médio, número de tentativas , trocaria de curso, perspectiva profissional e idade. Portanto, não foram consideradas estatisticamente significantes as variáveis sistema de ingresso, turno, sexo, raça/cor/etnia e realizou curso preparatório.

6.2.2 Seleção manual

Nesse método, foi considerado como inicial o modelo completo (todas as variáveis estão presentes), a partir do qual foram retiradas, uma a uma, as variáveis explicativas com menor significância, até que todas que restassem no modelo fossem significativas.

Do modelo inicial, os resultados de significância podem ser verificados na Tabela 16 a qual indica que a variável sistema de ingresso (X_{i2}) apresenta a menor significância (p-valor=0,791477), portanto, será retirada.

Do modelo sem a variável X_{i2}, após testadas as respectivas significâncias, a variável turno (X_{i5}) foi a menos significante (p-valor=0.979548), sendo retirada do modelo.

Do modelo sem as variáveis X_{i2} e X_{i5}, também testadas as respectivas significâncias, a variável realizou curso preparatório (X_{i14}) foi a menos significante (p-valor=0.456440), sendo retirada do modelo.

Do modelo sem as variáveis X_{i2}, X_{i5} e X_{i14}, testadas as respectivas significâncias, a variável sexo (X_{i7}) foi a menos significante (p-valor=0.464111), sendo retirada do modelo.

No próximo modelo sem as variáveis X_{i2}, X_{i5}, X_{i14} e X_{i7}, testadas as respectivas significâncias, a variável raça/cor/etnia (X_{i9}), apresentou níveis com p-valores no geral elevados (> 0.86) sendo retirada do modelo. Todas as variáveis que restaram mostraram-se significativas ao modelo e, portanto, comporão o modelo candidato.

Tabela 16: Resultados da significância dos parâmetros do modelo completo.

Variáveis	**Estimativa**	**Erro-Padrão**	**Valor Z**	**p-valor**[2]
Intercepto	-1,355465	0,189575	-7,150	8,68e-13***
Ano				
2013	-0,30709	0,068465	-4,485	7,28e-06***
2014	-0,537967	0,070693	-7,610	2,74e-14***
2015	-0,358646	0,069522	-5,159	2,49e-07***
2016	-0,293685	0,070958	-4,139	3,49e-05***
2017	-0,152143	0,073594	-2,067	0,038704*
Sistema de Ingresso				
Universal	0,011736	0,044387	0,264	0,791477
Modalidade de Ingresso				
ENEM	0,339264	0,075514	4,493	7,03e-06***
PAS	0,195468	0,075288	2,596	0,009425**
Vestibular	0,251161	0,06924	3,627	0,000286***
Campus				
Ceilândia	0,090007	0,061492	1,464	0,143272
Gama	-0,128973	0,069669	-1,851	0,064138,
Planaltina	0,523347	0,088548	5,910	3,41e-09***
Turno				
Noturno	0,000415	0,034893	0,012	0,99051
Curso por prestígio				
Baixo	0,435027	0,054152	8,033	9,48e-16***
Médio	0,299621	0,041114	7,288	3,16e-13***
Sexo				
Feminino	0,023911	0,030834	0,775	0,438056
Estado Civil				
Solteiro	0,192917	0,079941	2,413	0,01581*
Raça/Cor/Etnia				
Branca	0,016064	0,08981	0,179	0,858047
Indígena	0,014203	0,21579	0,066	0,947521
Parda	0,018599	0,090321	0,206	0,83685
Preta	0,007663	0,099527	0,077	0,938629
Outras	0,658792	0,304945	2,160	0,030745*
Renda				
3-10 SM	0,135225	0,040441	3,344	0,000827***
10-20 SM	0,173198	0,050416	3,435	0,000592***
>20 SM	0,086537	0,057568	1,503	0,132786
Não sabe	0,417788	0,138925	3,007	0,002636**
Escolaridade Pai				
Até o ensino médio	0,085031	0,035005	2,429	0,015135*
Não sabe	0,060319	0,076524	0,788	0,430557
Escolaridade Mãe				
Até o ensino médio	0,111177	0,034304	3,241	0,001191**
Não sabe	0,336289	0,162434	2,070	0,038424*
Ensino Médio				
Público/Particular	0,239619	0,060122	3,986	6,73e-05***
Particular	0,250105	0,043821	5,707	1,15e-08***
Curso preparatório				
Não	0,023859	0,031894	0,748	0,454416
Tentativas				
Uma	0,538867	0,045774	11,772	<2e-16***
Duas	0,909807	0,043546	20,893	<2e-16***
Três	1,183304	0,050686	23,346	<2e-16***
>Três	1,404236	0,056174	24,998	<2e-16***
Não concluinte	1,477335	0,064471	22,915	<2e-16***
Concluinte	1,80622	0,111232	16,238	<2e-16***
Trocaria de Curso				
Não	-1,276649	0,031852	-40,081	<2e-16***
Perspectiva profissional				
Não sabe	0,195691	0,041633	4,700	2,60e-06***
Lecionar	-0,198846	0,062318	-3,191	0,001419**
Empresa privada	0,01777	0,082994	0,214	0,830459
Concusrso	0,101093	0,037967	2,663	0,007753**
Já trabalho	0,073325	0,071003	1,033	0,301741
Outros	0,166403	0,153005	1,088	0,276787
Idade	0,01442	0,003567	4,043	5,28e-05***

[2]Significância dos parâmetros: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

A contribuição de cada variável preditora para o modelo estão apresentadas na Tabela 17. Existe uma relação entre o p-valor da variável e sua contribuição para o modelo, variáveis com p-valores menores apresentam um valor de contribuição maior, as variáveis trocaria de curso seguida do número de tentativas (e seus níveis) foram as que apresentaram as maiores contribuições para o modelo (p-valores também foram os menores).

Tabela 17: Análise da contribuição de cada variável.

Variáveis	Contribuição	Variáveis	Contribuição
Ano		**Escolaridade Pai**	
2013	4,6779	Até o ensino médio	2,4522
2014	7,8963	Não sabe	0,8367
2015	5,4919	**Escolaridade Mãe**	
2016	4,5088	Até o ensino médio	3,3194
2017	2,3643	Não sabe	2,0406
Modalidade de Ingresso		**Ensino Médio**	
ENEM	4,5123	Público/Particular	4,2934
PAS	2,6651	Particular	6,9278
Vestibular	3,6366	**Tentativas**	
Campus		Uma	11,7632
Ceilândia	1,6079	Duas	21,0391
Gama	1,9205	Três	23,6250
Planaltina	5,9824	>Três	25,3704
Curso por prestígio		Não concluinte	22,9055
Baixo	8,4347	Concluinte	16,2533
Médio	7,5326	**Trocaria de Curso**	
Estado Civil		Não	40,1176
Solteiro	2,3778	**Perspectiva profissional**	
Renda		Não sabe	4,7220
3-10 SM	3,3198	Lecionar	3,2433
10-20 SM	3,4156	Empresa privada	0,1552
>20 SM	1,4752	Concusrso	2,6741
Não sabe	3,0251	Já trabalho	1,0565
Idade	4,1422	Outros	1,1340

Foram testadas possíveis interações entre as variáveis selecionadas para o modelo de regressão logística.

- Ano ($X_{i,1}$) versus todas as variáveis significativas para o modelo.
- Renda ($X_{i,10}$) versus prestígio do curso, grau de escolaridade do pai/mãe e onde estudou no ensino médio.
- Trocaria de curso ($X_{i,16}$) versus campus e número de tentativas.
- Número de tentativas ($X_{i,15}$) versus curso

  As interações que mostraram-se significativas foram:

  – Ano versus Curso ($X_{i,1} * X_{i,6}$);

  – Ano versus Ensino Médio ($X_{i,1} * X_{i,13}$);

  – Ano versus Trocaria de Curso ($X_{i,1} * X_{i,16}$)

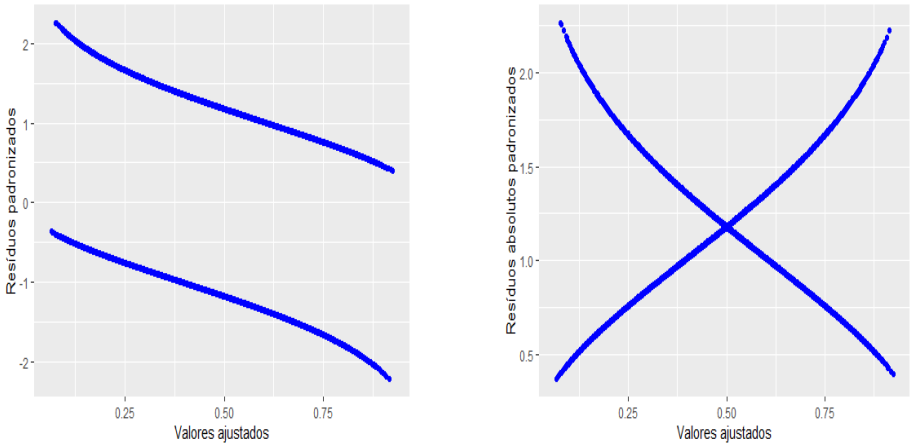
Ajustando-se o novo modelo a medida de ajuste do modelo foi *deviance*= 26558, ou seja, a inserção da interação entre as variáveis, o que torna o modelo ainda mais complexo, não reduziu de maneira significativa a variabilidade, além do mais, o nível de significância das variáveis reduziu (aumento do p-valor) causando a perda de significância para determinadas categorias e algumas das variáveis como modalidade de ingresso (X_{i3}) e escolaridade do pai ($X_{i,13}$), que antes tinham efeito importante para o modelo deixaram de ter.

Assim foi escolhido trabalhar com o modelo aditivo que embora apresente maior variabilidade possui menos variáveis favorecendo a sua interpretação, além de torná-lo mais parcimonioso, com menor erro e menos dependente dos dados observados.

6.3 Análise de resíduos

Como observado na Subseção 2.3 a regressão logística não assume que os resíduos sejam normalmente distribuídos nem que a variância seja constante.

Figura 16: Valores ajustados e resíduos.



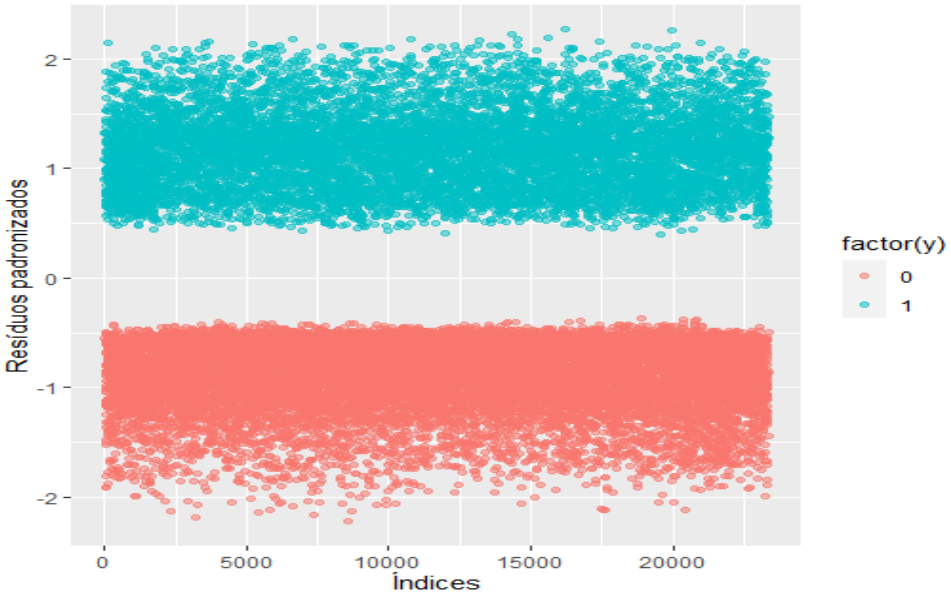
(a) Resíduos padronizados.

(b) Valor absoluto dos resíduos padronizados.

A Figura 16 representa os valores ajustados versus resíduos padronizados e valores ajustados versus módulo dos resíduos padronizados, os mesmos gráficos foram feitos considerando apenas os resíduos e a representação gráfica ficou similar, logo optou-se por utilizar apenas os gráficos considerando os resíduos padronizados.

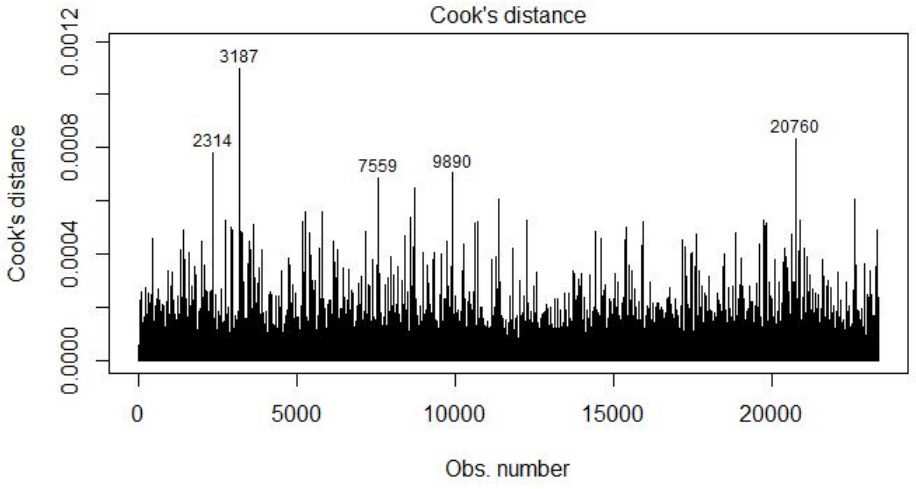
O desvio residual é útil para determinar se os pontos individuais não estão bem ajustados pelo modelo. Observa-se na Figura 17 que os resíduos estão bem distribuídos e não existe a presença de *outliers*. Pode-se também ajustar os resíduos para ver quantos excedem um determinado valor.

Figura 17: Distribuição dos resíduos.



Semelhante à regressão linear, também é possível identificar observações influentes com os valores de distância de Cook. A Figura 18 identificou os 5 maiores valores.

Figura 18: Cinco observações mais influentes no modelo.



Investigando essas observações mais detalhadamente verifica-se que esses pontos influentes incluem:

– Ingresso no curso desejado, ano de ingresso 2012; modalidade de ingresso outros (diploma de curso superior, vagas remanescentes e transferência); campus Dary-Ribeiro; curso de prestígio médio; estado civil solteiro e trocariam de curso. Os demais preditores não apresentaram nenhum nível com destaque.

6.4 Análise das estatíticas da qualidade do ajuste

Para verificar a qualidade do ajuste do modelo selecionado serão utilizadas as estatísticas e a hipótese apresentada na Subseção 2.10. Os resultados foram obtidos por meio do *software R*.

A Tabela 18 apresenta as principais estatísticas para a capacidade do modelo explicar a variabilidade dos dados. Quanto mais próximo de 1 melhor o ajuste do modelo. Cabe ressaltar que essa proximidade é teórica e os valores podem sofrer alterações ao se trabalhar com dados reais.

Tabela 18: Principais estatísticas para o modelo.

Cox Snell	Nagelkerke	McFadden	Tjur	χ^2 de Pearson
0,162	0,218	0,130	0.167	0,167

Para verificar a contribuição de cada parâmetro no modelo para redução da variância (Tabela 19) foi feito o teste *Anova*. A variância com o modelo nulo foi 31741 e à medida que as variáveis explicativas foram sendo incluídas, estas reduziram a variância do modelo para 27607, contribuindo para o mesmo.

Tabela 19: Análise da redução da *deviance*

	DF	**Deviance**	**g.l resíduos**	**Deviance resíduos**	**p-valor**[2]
Intercepto			23356	31741	
Ano	5	200,69	23351	31540	<2,2e-16***
Modalidade de ingresso	3	243,50	23348	31296	<2,2e-16***
Campus	3	167,68	23345	31129	<2,2e-16***
Curso por prestígio	2	270,48	23343	30858	<2,2e-16***
Estado Civil	1	0,00	23342	30858	0,949647
Renda	4	72,00	23338	30786	8,595e-15***
Escolaridade pai	2	7,67	23336	30778	0,021594*
Escolaridade mãe	2	13,68	23334	30765	0,001068**
Ensino médio	2	87,22	23332	30678	<2,2e-16***
Número de tentativas	6	1259,31	23326	29418	<2,2e-16***
Trocaria de curso	1	1748,59	23325	27670	<2,2e-16***
Perspectiva profissional	6	45,55	23319	27624	3,646e-08***
Idade	1	17,15	23318	27607	3,453e-05***

[2]Significância dos parâmetros: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

6.5 Validação do modelo

Ao desenvolver modelos para predição, um fator a ser considerado é quão bem o modelo se sai em predizer a variável resposta em observações fora da amostra.

Considerando que a análise de diagnóstico indicou a adequabilidade do modelo proposto em relação aos pressupostos, as outras 14811 observações foram utilizadas para o procedimento de validação do modelo escolhido baseado nas primeiras 23357 observações do banco de dados.

Para verificar a capacidade preditiva do modelo será utilizada a matriz de classificação vista na Subseção 2.12, bem como, os conceitos apresentados na Subseção 2.11, considerando 0.5 como ponte de corte.

Tabela 20: Matriz de classificação.

Predição	**Referência**	
	0	1
FALSO	0,449 (6646)	0,204 (3017)
VERDADEIRO	0,125 (1852)	0,223 (3296)

O modelo consegue acertos de 64% na predição de valores positivos ou dos "eventos" 3296/(1852 + 3296) e 68,8% na predição de valores negativos ou os "não eventos" 6646/(6646 + 3017). Esses valores também estão presentes na Tabela 21 como Verdadeiro Preditivo Positivo e Verdadeiro Preditivo Negativo respectivamente.

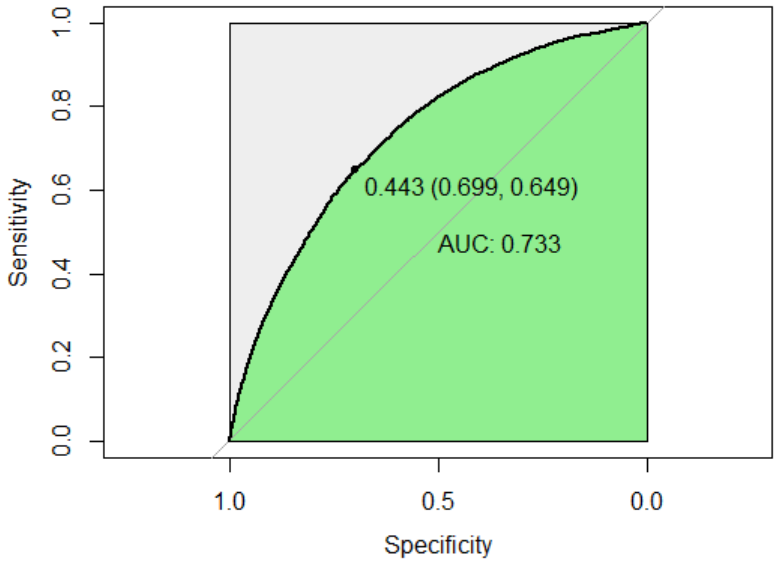
Tabela 21: Valores obtidos da matriz de classificação.

Estimativas	**Treino**	**Validação**
Sensibilidade	0,5162	0,5221
Especificidade	0,8041	0,7821
Verdadeiro Preditivo Positivo	0,6537	0,6402
Verdadeiro Preditivo Negativo	0,6987	0,6878
Precisão (Acurácia)	0,6839	0,6713

O modelo apresentou uma excelente acurácia total 67%, Tabela 21, com intervalo de (0.6636, 0.6788) com 95% de confiança para o verdadeiro valor do parâmetro. A capacidade do modelo em fazer previsão (sensibilidade), isto é, avaliar o evento predito como verdadeiro dado que ele é verdadeiro está em torno de 52%. Já a especificidade em avaliar um não evento dado que ele é não evento é aproximadamente 78%.

A Figura 19 representa a Área sobre a Curva ROC (AUC - Area Under the ROC Curve) cujo objetivo é comparar os classificadores a partir da performance da curva em um único valor FAWCETT (2006). Este conceito foi abordado na Subseção 2.13 e de acordo com a classificação de HOSMER e LEMESHOW (2000) quando $0.7 < AUC < 0.8$ o modelo é aceitável.

Figura 19: Especificidade versus sensibilidade.



Para o modelo em estudo AUC=0.733, o que torna o modelo aceitável, isso significa que se forem feitas duas observações quaisquer elas serão corretamente ordenadas com probabilidade de 73%.

A Tabela 22 apresenta as estimativas dos parâmetros e seus respectivos p-valores para as variáveis preditoras usando o banco de dados para a construção do modelo (Treino) e o banco de dados para a validação do modelo. Nota-se que os parâmetros estimados estão bem próximos. Já os p-valores da validação (V) com exceção das variáveis estado civil ($X_{i,8}$) e escolaridade do pai ($X_{i,11}$) apresentaram valores significativos.

Tabela 22: Estimativas e p-valores para o modelo de treino e validação.

Variáveis	**Estimativa (T)**	**Estimativa (V)**	**p-valor (T)**[2]	**p-valor (V)**[2]
Intercepto	-1,295659	-0,978867	3,48e-15***	1,16e-06***
Ano				
2013	-0,318147	-0,461356	2,90e-06***	6,37e-10***
2014	-0,553397	-0,525344	2,87e-15***	1,24e-11***
2015	-0,375451	-0,311076	3,98e-08***	4,36e-05***
2016	-0,312091	-0,29744	6,52e-06***	0,000113***
2017	-0,170146	-0,053006	0,018065*	0,483529
Modalidade de Ingresso				
ENEM	0,34009	0,14949	6,41e-06***	0,090264.
PAS	0,200369	0,197866	0,007696**	0,025078*
Vestibular	0,251481	0,315698	0,000276***	8,74e-05***
Campus				
Ceilândia	0,095672	0,212713	0,107848	0,004525**
Gama	-0,132776	-0,101324	0,054797.	0,222996
Planaltina	0,527861	0,501138	2,20e-09***	3,83e-06***
Curso por prestígio				
Baixo	0,44398	0,465971	<2e-16***	7,46e-13***
Médio	0,302845	0,256264	4,97e-14***	2,54e-07***
Estado Civil				
Solteiro	0,189993	0,030988	0,017417*	0,760721
Renda				
3-10 SM	0,13361	0,110315	0,000901***	0,027440*
10-20 SM	0,170603	0,273633	0,000637***	1,18e-05***
>20 SM	0,08371	0,264102	0,140159	0,000167***
Não sabe	0,41999	0,36459	0,002485**	0,019232*
Escolaridade Pai				
Até o ensino médio	0,085442	0,035554	0,014199*	0,409354
Não sabe	0,063848	0,055888	0,402734	0,563498
Escolaridade Mãe				
Até o ensino médio	0,113324	0,142497	0,000902***	0,000780***
Não sabe	0,331212	0,028915	0,041291*	0,882355
Ensino Médio				
Público/Particular	0,24376	0,183204	1,76e-05***	0,009829**
Particular	0,256444	0,228005	4,28e-12***	7,15e-07***
Tentativas				
Uma	0,537173	0,390368	<2e-16***	1,42e-11***
Duas	0,906274	0,763083	<2e-16***	<2e-16***
Três	1,17929	1,008887	<2e-16***	<2e-16***
>Três	1,396256	1,310128	<2e-16***	<2e-16***
Não concluinte	1,475464	1,445786	<2e-16***	<2e-16***
Concluinte	1,807342	1,430088	<2e-16***	<2e-16***
Trocaria de Curso				
Não	-1,27592	-1,27609	<2e-16***	<2e-16***
Perspectiva profissional				
Não sabe	0,196506	0,172715	2,34e-06***	0,000889***
Lecionar	-0,201735	0,009199	0,001181**	0,907017
Empresa privada	0,012863	0,077147	0,876633	0,44334
Concusrso	0,101094	0,129841	0,007493**	0,005826**
Já trabalho	0,074856	0,206785	0,290719	0,016114*
Outros	0,173277	0,39564	0,256786	0,042305*
Idade	0,014531	0,009224	3,44e-05***	0,040953*

[2]Significância dos parâmetros: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

6.6 Resultados

Com o modelo de desenvolvimento validado, os parâmetros serão estimados considerando a base de dados completa, ou seja, as 38168 observações.

Tabela 23: Parâmetros estimados considerando a base de dados completa.

Variáveis	Estimatica	Erro Padrão	Valor Z	P-valor[2]
Intercepto	-1,17304	0,126918	-9,243	<2e-16***
Ano				
2013	-0,36857	0,049845	-7,394	1,42e-13***
2014	-0,54267	0,051569	-10,523	<2e-16***
2015	-0,34987	0,050327	-6,952	3,60e-12***
2016	-0,3048	0,050965	-5,981	2,22e-09***
2017	-0,1174	0,051932	-2,261	0,023782*
Modalidade de Ingresso				
ENEM	0,258566	0,057211	4,52	6,20e-06***
PAS	0,195523	0,05713	3,422	0,000621***
Vestibular	0,271903	0,052351	5,194	2,06e-07***
Campus				
Ceilândia	0,139097	0,046545	2,988	0,002804**
Gama	-0,11828	0,053108	-2,227	0,025938*
Planaltina	0,514961	0,068382	7,531	5,05e-14***
Curso por prestígio				
Baixo	0,452006	0,040852	11,065	<2e-16***
Médio	0,285375	0,031226	9,139	<2e-16***
Estado Civil				
Solteiro	0,12983	0,06273	2,07	0,038483*
Renda				
3-10 SM	0,123559	0,031323	3,945	7,99e-05***
10-20 SM	0,210593	0,038959	5,406	6,46e-08***
>20 SM	0,153513	0,044069	3,483	0,000495***
Não sabe	0,393974	0,103404	3,81	0,000139***
Escolaridade Pai				
Até o ensino médio	0,066339	0,027062	2,451	0,014231*
Não sabe	0,058275	0,059848	0,974	0,330193
Escolaridade Mãe				
Até o ensino médio	0,124546	0,026559	4,689	2,74e-06***
Não sabe	0,212356	0,124463	1,706	0,087976,
Ensino Médio				
Público/Particular	0,219925	0,044289	4,966	6,85e-07***
Particular	0,246578	0,028801	8,562	<2e-16***
Tentativas				
Uma	0,478052	0,035781	13,361	<2e-16***
Duas	0,846197	0,033614	25,174	<2e-16***
Três	1,112681	0,038768	28,701	<2e-16***
>Três	1,357695	0,042726	31,777	<2e-16***
Não concluinte	1,458342	0,050394	28,939	<2e-16***
Concluinte	1,644385	0,081856	20,089	<2e-16***
Trocaria de Curso				
Não	-1,2742	0,024792	-51,396	<2e-16***
Perspectiva profissional				
Não sabe	0,184825	0,032445	5,697	1,22e-08***
Lecionar	-0,12196	0,048731	-2,503	0,012327*
Empresa privada	0,037702	0,063883	0,59	0,555071
Concusrso	0,110953	0,029446	3,768	0,000165***
Já trabalho	0,124664	0,054595	2,283	0,022406*
Outros	0,252378	0,119856	2,106	0,035233*
Idade	0,012681	0,002765	4,586	4,53e-06***

[2]Significância dos parâmetros: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

O modelo de regressão logística, porém, traz os resultados dos estimadores na forma logarítmica, ou seja, o log das chances da variável idade no modelo é 0,012681. Para uma melhor interpretação da relação de determinada variável com o não ingresso no curso desejado mantida todas as demais variáveis constantes é necessária a transformação deste coeficiente, utilizando o *antilog* ($e^{\hat{\beta}_{18}}$) nas variáveis de regressão. Assim, obtém-se a razão das chances (*OR-Odds Ratio*) para as variáveis independentes, vide Subsubseção 2.4.1 e Subsubseção 2.6.1.

A razão de chances e seu intervalo de confiança, Tabela 24, serão obtidos realizando procedimentos análogos a teoria apresentada na Subseção 2.6 e Subsubseção 2.6.2 respectivamente.

Tabela 24: Razão de chances e intervalo de confiança.

Variáveis	**OR**	**2.5%**	**97.5%**	**Variáveis**	**OR**	**2.5%**	**97.5%**
Intercepto	0,309	0,241	0,397	**Escolaridade Pai**			
Ano				Superior ou mais	-	-	-
2012	-	-	-	Até o ensino médio	1,069	1,013	1,127
2013	0,692	0,627	0,763	Não sabe	1,06	0,943	1,192
2014	0,581	0,525	0,643	**Escolaridade Mãe**			
2015	0,705	0,639	0,778	Superior ou mais	-	-	-
2016	0,737	0,667	0,815	Até o ensino médio	1,133	1,075	1,193
2017	0,889	0,803	0,985	Não sabe	1,237	0,969	1,578
Modalidade de Ingresso				**Ensino Médio**			
Outros	-	-	-	Público	-	-	-
ENEM	1,295	1,158	1,449	Público/Particular	1,246	1,142	1,359
PAS	1,216	1,087	1,36	Particular	1,28	1,209	1,354
Vestibular	1,312	1,184	1,454	**Tentativas**			
Campus				Nenhuma	-	-	-
Darcy Ribeiro	-	-	-	Uma	1,613	1,504	1,73
Ceilândia	1,149	1,049	1,259	Duas	2,331	2,182	2,489
Gama	0,888	0,801	0,986	Três	3,043	2,82	3,283
Planaltina	1,674	1,464	1,914	>Três	3,887	3,575	4,227
Curso por prestígio				Não concluinte	4,299	3,895	4,745
Alto	-	-	-	Concluinte	5,178	4,41	6,079
Baixo	1,571	1,451	1,702	**Trocaria de Curso**			
Médio	1,33	1,251	1,414	Sim	-	-	-
Estado Civil				Não	0,28	0,266	0,294
Outros	-	-	-	**Perspectiva profissional**			
Solteiro	1,139	1,007	1,288	Na área de formação	-	-	-
Renda				Não sabe	1,203	1,129	1,282
Até 3 SM	-	-	-	Lecionar	0,885	0,805	0,974
3-10 SM	1,132	1,064	1,203	Empresa privada	1,038	0,916	1,177
10-20 SM	1,234	1,144	1,332	Concusrso	1,117	1,055	1,184
>20 SM	1,166	1,069	1,271	Já trabalho	1,133	1,018	1,261
Não sabe	1,483	1,211	1,816	Outros	1,287	1,018	1,628
				Idade	1,013	1,007	1,018

Considerando a categoria Ceilândia da variável campus ($X_{2,4}$) a chance de não ingressar no curso desejado é 1,14 vezes maior comparado com nível Darcy Ribeiro mantidas todas as demais variáveis constantes. Ou também, as chances de não ingressar no curso desejado é 14,9% maior para quem ingressou no campus Ceilância comparado com o campus Darcy Riberio mantidas todas as demais variáveis constantes.

Interpretações análogas a anterior podem ser feitas considerando os parâmetros estimados para cada nível das variáveis mantidas as demais constantes. Quando o valor do parâmetro estimado é menor que 1 a interpretação é um pouco diferente, por exemplo, para a variável ano de ingresso possui *odds ratio* < 1.

Considerando a categoria 2013 da variável ano de ingresso a chance de não ingressar no curso desejado é 0,69 vezes menor comparado com o nível 2012, mantidas todas as demais variáveis constantes.

Dito de outra forma a afirmação do parágrafo anterior, a chance não ingressar no curso desejado é 30,8% menor para os ingressantes do ano de 2013 comparado com ano de 2012. Nota-se que a cada ano, a partir de 2014, a razão de chances < 1 de não ingressar no curso desejado foi aumentando comparada com a categoria 2012. Consequentemente foi reduzindo o quanto a chance de não ingressar no curso desejado é menor do que a chance de ingressar no curso desejado: 29,5% (2015), 26,3% (2016) e 11,1% (2017), comparado ao ano de 2012.

Considerando a análise descritiva e bivariada das variáveis, e estudos referenciais, dos resultados apresentados na Tabela 24 alguns já eram esperados como, por exemplo:

- Modalidade de ingresso (X_{i3}): Alunos ingressantes por ENEM, PAS e vestibular tem mais chances de não ingressar no curso desejado comparado com a categoria outros que corresponde a diploma de curso superior, vagas remanescente e transferência, pois, além da concorrência ser menor, estes processos seletivos (outros) são divulgados para um público específico e em alguns casos muitos alunos não ficam sabendo devido principalmente a grande diversidade das datas de publicação dos editais, o que de certa forma facilita o ingresso no curso desejado.
- Campus (X_{i4}): Esses resultados estão diretamente relacionados as áreas de formação, pois, enquanto Ceilândia oferece cursos ligados à área da saúde (Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Saúde Coletiva e Terapia Ocupacional) e Planaltina-DF a área das ciências agrárias (Ciências Naturais, Educação do Campo, Gestão Ambiental e Gestão do Agronegócio), já o campus Gama oferece cursos ligado a área das engenharias (Engenharia Aeroespacial, Engenharia Automotiva, Engenharia de Energia, Engenharia de Software e Engenharia Eletrônica).

Assim para os campus Ceilândia ($X_{2,4}$) e Planaltina-DF ($X_{4,4}$) as chances do aluno não ingressar no curso desejado é maior comparado com o campus Darcy Ribeiro ($X_{1,4}$). Já a chance de o aluno não ingressar no curso desejado é menor no campus Gama ($X_{3,4}$) comparado com o campus Darcy Ribeiro.

- Curso ($X_{i,6}$): Alunos que ingressaram em cursos de baixo prestígio ($X_{2,6}$) têm 57,1% mais chances de não ingressar no curso desejado comparado com alunos que ingressaram em cursos de alto prestígio ($X_{3,6}$) mantidas todas as demais categorias constantes. Já alunos que ingressaram em cursos de médio prestígio ($X_{3,6}$) tem 33% mais chances de não ingressar no curso desejado comparado com alunos que ingressaram em cursos de alto prestígio ($X_{1,6}$) mantidas todas as demais categorias constantes.
- Estado civil ($X_{i,8}$): É de se esperar que os ingressantes que pertencem a categoria 1 (casado(a), divorciado(a), separado(a), união estável e viúvo(a)) tem mais maturidade nas suas escolhas, já tiveram alguma experiência além do relacionamento familiar/amigos.
- Escolaridade do pai e da mãe ($X_{i,11}$) e ($X_{i,12}$): Alunos cujos pais possuem grau de escolaridade até o ensino médio ($X_{2,11}$) e ($X_{2,12}$) têm mais chance de não ingressar no curso desejado comparados com os alunos com pais que possuem ensino superior ou mais ($X_{1,11}$) e ($X_{1,12}$).
- Trocaria de curso ($X_{i,16}$): Existe uma ligação direta entre o não ingresso no curso desejado e querer trocar de curso, esta variável foi bem significativa no modelo. Assim alunos que não querem trocar de curso ($X_{2,16}$) tem uma menor chance de não ingressar no curso desejado comparado com aqueles que querem trocar de curso.
- Perspectiva profissional ($X_{i,17}$): Ao comparar as categorias (níveis: 2,3,4,5, 6 e 7) com a categoria deseja atuar na área de formação os resultados também são autoexplicativos, espera-se que para essas outras categorias a chance de ingresso no curso não desejado seja maior comparada com a categoria de referência. A exceção ocorre na categoria lecionar (ensino médio/superior), nível 3 ($X_{3,17}$), os estudantes que ingressam com esse objetivo profissional têm menos chances de não ingressar no curso desejado.

Outros resultados merecem mais atenção como:

- Renda ($X_{i,10}$): Rendas maiores tem mais chances de não ingresso no curso desejado comparadas individualmente com a categoria base até 3 salários mínimos ($X_{1,10}$).
- Ensino médio ($X_{i,13}$): Alunos que estudaram na rede particular/pública ($X_{2,13}$) e particular ($X_{3,11}$) a chance de não ingressar no curso desejado é 24,6% e 28% respectivamente maior comparado com quem estudou apenas na rede pública ($X_{1,13}$).

As causas para estes resultados podem estar diretamente relacionadas as ações de políticas afirmativas mencionadas na parte descritiva, Seção 4. Tais variáveis estão ligadas ao perfil socioeconômico do estudante que tem sido um dos critérios para seleção no ensino superior dentro das cotas.

A variável número de tentativas ($X_{i,15}$) também apresentou resultados singulares, todas as *odds ratio* comparadas individualmente com a categoria base primeira tentativa ($X_{1,15}$) apresentaram chances > 1. Assim em todos os casos as chances de ingresso no curso não desejado são maiores quando comparados com o nível 1, primeira tentativa ($X_{1,15}$).

No entanto, pode-se pensar que estudantes que não conseguiram ingressar no curso desejado na primeira tentativa podem querer mudar a escolha do curso. Nota-se que a medida que o número de tentativas aumenta as chances de não ingressar no curso desejado também aumentam, por exemplo entre os ingressantes que já fizeram mais de três tentativas ($X_{4,15}$) a chance de não ingressar no curso desejado 3 vezes maior comparado com aqueles que ingressaram na primeira tentiva.

Para a variável idade as chances de não ingresso no curso desejado aumenta em 1,3% para cada aumento de 1 ano (variação unitária). Observa-se que o valor da *odds ratio* é diferente do encontrado no modelo de regressão logística considerando apenas essa variável 2,4%.

6.6.1 Predição de probabilidade

Com base no modelo de regressão logística já consolidado é possível estimar a probabilidade de um ingressante com determinadas características não ingressar no curso no curso desejado, essa é considerada uma das aplicações mais relevantes para o modelo logístico em diferentes áreas de estudo.

Como existem muitas possibilidades para análise será verificada algumas probabilidades, porém quanto maior as chances de não ingressar no curso desejado maior a probabilidade de predição, assim é possível obter probabilidades elevadas ou baixas considerando categorias com *odds ratio* OR maior ou menor respectivamente. Algumas possibilidades são:

- Pode-se pensar que o fato do aluno já almejar trocar de curso influencia o aumento da probabilidade para o não ingresso no curso desejado.

Assim para cada ano foi alterado o nível (sim e não) da variável trocaria de curso ($X_{i,16}$), mantendo-se os níveis das demais categorias constantes para os fatores do modelo de regressão.

A Tabela 25 apresenta as probabilidades de não ingressar no curso desejado para um perfil de estudante que fez o vestibular para concorrer a uma vaga no campus Darcy Ribeiro em um curso de médio prestígio, solteiro, com renda familiar de 3 a 10 salários-mínimos, com pais que cursaram até o ensino fundamental, que estudou apenas em instituições de ensino particulares e já tinha feito uma tentativa de ingresso, com idade de 20 anos.

Tabela 25: Probabilidades de não ingressar no curso desejado.

Ano	**Trocaria de curso**	**Não Trocaria de curso**
2012	0.704	0,43
2013	0.622	0,343
2014	0.581	0,305
2015	0.627	0,347
2016	0.637	0,357
2017	0.679	0,401

Os resultados apresentados indicam que modificando o nível da variável trocaria de curso para não ($X_{2,16}$), a probabilidade de não ingressar no curso desejado reduz de forma considerável, mostrando a importância dessa variável para o modelo.

- Foram realizadas análises da probabilidade do aluno não ingressar no curso desejado considerando o último ano do estudo (2017) e mudanças de algumas categorias que tenham maior efeito na razão de chances.

Assim para os alunos com idade de 20 anos, que ingressaram pelo vestibular, para um curso de médio prestígio, solteiro, com renda familiar de 3 a 10 salários-mínimos, que estudou apenas em instituições de ensino particulares, cujo nível de escolaridade dos pais é até ensino fundamental, já fez uma tentativa anterior, não trocaria de curso e a perspectiva profissional é atuar na área

de formação. A probabilidade de não ingressar no curso desejado modificando os níveis do fator campus está apresentada na Tabela 26.

Tabela 26: Probabilidades de não ingressar no curso desejado.

Campus			
Darcy-Ribeiro	Ceilândia	Gama	Planaltina-DF
0,3578	0,3903	0,3311	0,4825

A Tabela 27 contêm a probabilidade de não ingressar no curso desejado modificando os níveis do fator número de tentativas para os alunos com idade de 20 anos, que ingressaram pelo vestibular, no campus Darcy-Ribeiro, para um curso de médio prestígio, solteiro, com renda familiar de 3 a 10 salários mínimos, que estudou apenas em instituições de ensino particulares, cujo o nível de escolaridade dos pais é até ensino fundamental, não trocaria de curso e a perspectiva profissional é atuar na área de formação.

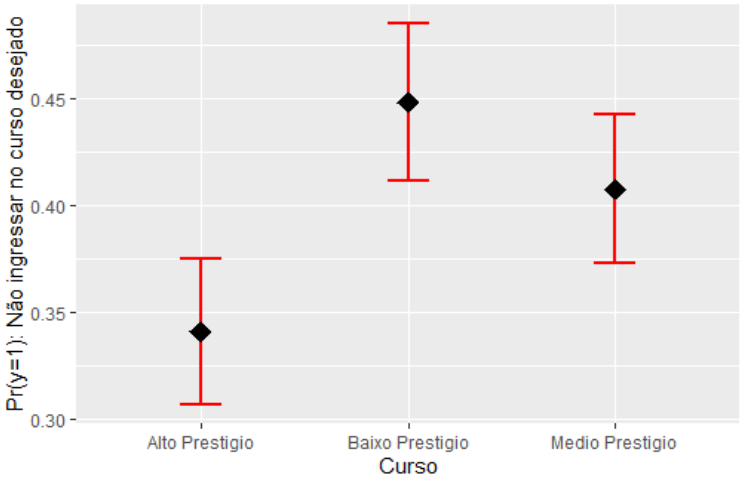
Tabela 27: Probabilidades de não ingressar no curso desejado.

Número de Tentativas						
Nenhuma	Uma	Duas	Três	Mais de três	Não concluinte UnB	Concluinte UnB
0,2567	0,3578	0,446	0,5124	0,573	0,5975	0,6414

Verifica-se com base na Tabela 26 e Tabela 27 que mesmo o aluno não tendo interesse em trocar de curso existe uma alta probabilidade de não ter ingressado no curso desejado considerado variáveis com categorias que apresentaram *odds ratio* elevadas.

A Figura 20 indica alterações na probabilidade de não entrar no curso desejado para um ingressante no ano 2017, com idade de 20 anos, que prestou vestibular, para o campus Darcy-Ribeiro, solteiro, com renda familiar de 3 a 10 salários-mínimos, que estudou apenas em instituições de ensino particulares, cujo o nível de escolaridade dos pais é até ensino fundamental, que já fez duas tentativas, e não trocaria de curso e a perspectiva profissional é prestar concurso.

Figura 20: Probabilidades preditas de não ingressar no curso pretendido.

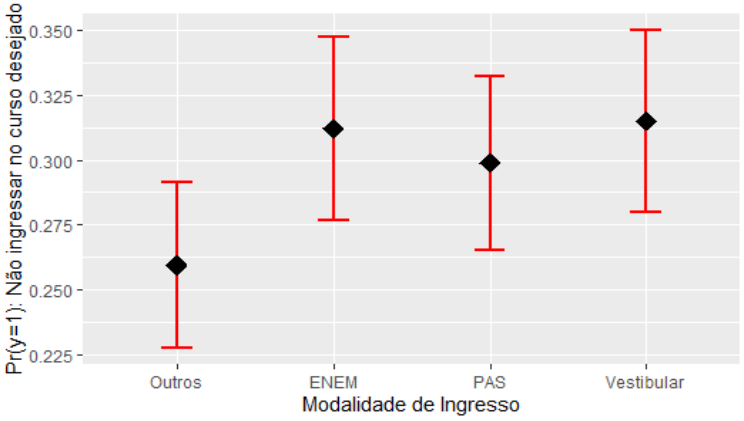


Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

As barras verticais (Figura 20) correspondem ao intervalo de confiança para a probabilidade predita de não ingressar no curso desejado, já o ponto triangular é a probabilidade estimada.

A Figura 21 indica alterações na probabilidade de não entrar no curso desejado para um ingressante no ano 2017, com 17 anos, para o campus Planaltina-DF, em um curso de médio prestígio, solteiro, com renda familiar de 3 a 10 salários-mínimos, que estudou apenas em instituições de ensino particulares, cujo o nível de escolaridade dos pais é superior ou mais, com nenhuma tentativa anterior, não trocaria de curso e a perspectiva profissional é atuar na área.

Figura 21: Probabilidades preditas de não ingressar no curso pretendido.



Fonte: Elaboração própria. Dados - OVE, 2012 - 2017.

Os resultados são coerentes com a teoria bem com outras análises realizadas anteriormente, assim a probabilidade de não ingressar no curso desejado mantidas todas as demais variáveis constantes é maior entre cursos de baixo prestígio. Já entre as modalidades de ingresso essa probabilidade é maior para os ingressantes do ENEM e vestibular mantidas todas as demais variáveis constantes.

7 Conclusões

Os resultados apontam para uma mudança no perfil do ingressante da Universidade de Brasília relacionada principalmente a políticas sociais de ação afirmativa para a ampliação do acesso de alunos cotistas, tornado o sistema de ingresso mais acessível e igualitário.

O modelo de regressão logística identificou as principais variáveis relacionadas ao não ingresso no curso desejado sendo elas: ano de ingresso, modalidade de ingresso, campus, curso segundo o prestígio, estado civil, renda, grau de escolaridade dos pais, onde estudo no ensino médio, número de tentativas, trocaria de curso, perspectiva profissional e idade.

Verificado o ajuste do modelo, foi elaborada a matriz de classificação a fim de obter medidas que descrevessem a capacidade preditiva, acurácia, sensibilidade e especificidade para a validação do modelo. Todas as medidas revelaram o bom ajuste do modelo e a curva ROC apresentou área em torno de 0.73, que de acordo com a literatura indica a aceitabilidade do modelo.

Para as variáveis modalidade de ingresso, campus, curso, estado civil, escolaridade dos pais trocaria de curso, perspectiva profissional e idade os resultados da razão de chances apresentaram valores convenientes ao serem comparados com a categoria base.

Para as variáveis renda e onde estudou no ensino médio os resultados necessitam de análises em um contexto mais amplo, levando em consideração justamente as mudanças ocorridas nos últimos anos principalmente no perfil socioeconômico dos ingressantes, assim estudantes com maior renda, que estudaram em escolas particulares tem menos chances de ingressar no curso desejado, ou seja, aumentam as chances de não ingressar no curso desejado ao comparar com alunos com renda menor e que estudaram somente em instituições públicas de ensino.

Outro ponto relevante no modelo é a indicação do aumento das chances de não ingressar no curso desejado quando aumenta o número de tentativas. Assim, por exemplo, entre aqueles tentaram três vezes as chances de não ingressar no curso desejado é 3 vezes maior comparado com aqueles que ingressaram na primeira tentativa. Esses resultados estão em consonância com a variável motivos para não ingressar no curso desejado cuja categoria tentativas anteriores sem sucesso foi a segunda mais apontada pelos ingressantes em primeiro lugar apareceu concorrência e falta de preparo.

Com base no modelo selecionado foi possível determinar a probabilidade de um ingressante com determinadas características não ingressar no curso desejado e sua relação direta com a *odds ratio*, pois quanto maior for a mesma para determinado atributo do estudante maiores as chances de não ingresso.

Levando em consideração o referencial teórico o presente estudo mostrou-se de relevante importância para compreensão do perfil do discente da Universidade de Brasília, compreendendo variáveis que englobam diversos aspectos dos ingressantes o que tornou possível o conhecimento do elevado número de alunos que não ingressam no curso desejado e possíveis causas relacionadas.

O modelo desenvolvido identificou fatores relacionados com a variável resposta, e com base nas probabilidades de determinado estudante não ingressar no curso escolhido pode auxiliar na elaboração de programas, que promovam não apenas o acesso do estudante, mas também sua permanência na instituição.

8 Referências Bibliográficas

AGRESTI, A. **Categorical Data Analysis**. Wiley-Interscience, New York. 1990.

AGRESTI, A. **An Introduction to Categorical Data Analysis**.Wiley-Interscience, 1996.

BARDAGI, M. P. **Evasão e comportamento vocacional de universitários: estudos sobre o desenvolvimento de carreira na graduação**. Tese (Doutorado) - Instituto de Psicologia, Universidade Federal do Rio Grande do Sul, 2007.

BARROS, R. P. et al. Determinantes do desempenho educacional no Brasil. **Instituto de Pesquisa Econômica e Aplicada - IPEA**, 2001. Disponível em: `http://repositorio.ipea.gov.br/bitstream/11058/2160/1/TD_834.pdf`. Visitado em: 21/06/2021

BUSSAB, W. O.; MORETTIN P. A. **Estatística básica**. Sétima Edição. Editora Saraiva. 2012.

CESAR, L. J. T. Mecanismos de seleção para o ensino superior e desigualdade educacional: um estudo sobre o PAS e o vestibular na Universidade de Brasília. **Sociedade e Estado**, v.28, n.3, 2013. Disponível em: `https://periodicos.unb.br/index.php/sociedade/article/view/5856`. Visitado em: 21/06/2021.

COSTA, B. S. P.; VASCONCELOS, A. M. N.; COSTA, M. T. L. Medidas de desigualdade: uma análise de segregação socioespacial na área metropolitana de Brasília. In: 22º SINAPE - Simpósio Nacional de Probabilidade e Estatística, Porto Alegre, 2016.

CUNHA, A. M.; TUNES, E.; SILVA, R. R. Evasão do curso de química da Universidade de Brasília: a interpretação do aluno evadido. **Química Nova**, São Paulo, v. 24, n. 2, p. 262-280, abr. 2001.

DIAS, E. C. M.; THEÓPHILO, C. R.; LOPES, M. A. S. Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros–Unimontes-MG. Disponível em: https://congressousp.fipecafi.org/anais/artigos102010/419.pdf. Visitado em: 06/10/2020.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**. 2006. Disponível em: https://doi.org/10.1016/j.irbm.2014.09.001. Visitado em: 04/05/2020.

GISI. L. M. **A Educação Superior No Brasil e o Caráter de Desigualdade do Acesso e da Permanência**. Revista Diálogo Educacional. Curitiba. V. 6. N.17. p.97-112. Jan./abr.2006.

HAIR, J. F.; WILLIAM C. B.; BARRY, J. B.; TATHAM R. L. **Análise multivariada de dados**. Sexta edição. São Paulo: Bookman.

HOSMER, D. W. ; LEMESHOW, S. **Applied Logistic Regression**. Segunda edição. New York: Wiley. 2000.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP) /Ministério da Educação. **Censo da Educação Superior (2016)**. Brasília: INEP/Ministério da Educação. Outubro de 2017. Disponível em: http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior. Visitado em 04/05/2020.

KUTNER M. H.; NACHTSCHEIM C. J.; NETER J.; LI W. **Applied Linear Statistical Models**. Quarta edição. McGraw-Hill Irwin, 2005.

OLIVEIRA, M.B. **Identificação de tipologias por meio do métodoGrade of Membership: Uma aplicação aos estudantes da Universidade de Brasília**. 2016. Trabalho de Conclusão de Curso. Universidade de Brasília. 2020. Disponível em:https://bdm.unb.br/handle/10483/24568. Visitado em

02/04/2020.

PROVETE, D. B. et al. Estatística aplicada à ecologia usando o R. Universidade Estadual Paulista (UNESP). Programa de Pós-Graduação Biologia Animal, 2011. Disponível em: `https://cran.r-project.org/doc/contrib/Provete-Estatistica_aplicada.pdf`. Visitado em: 21/06/2021.

RAWLINGS, J. O.; SASTRY, G. P. e DICKEY, D. A. **Applied Regression Analysis: A Research Tool**. Segunda edição. New York: Springer. 1998

SEVERINO. J. A. **O ensino superior brasileiro: novas configurações e velhos desafios**. Educar. Curitiba. n. 31. p. 39-52. 2008. Editora UFPR.

SHEATHER S. J. **A Modern Approach to Regression with R**. Springer, 2009

SIMONOFF, J. S. **Smoothing methods in statistics** . Springer, New York. 1996.

SIMONOFF, J.S. **Analyzing categorical data**. Springer, New York. 2003

9 APÊNDICE – Tabelas descritivas.

Tabela 28: Perfil de Ingresso do Estudante na Universidade.

Características	2012	2013	2014	2015	2016	2017
Sistema de ingresso	**7161**	**7093**	**7122**	**7846**	**7470**	**7571**
Não se aplica	-	-	-	-	0,03	-
Cotas	9,43	7,56	-	-	-	-
Cotas para Escolas Públicas	-	3,58	20,47	8,08	10,47	10,22
Cotas Raciais (UnB)	-	7,71	5,10	6,11	6,35	5,92
Universal	90,57	81,15	74,43	60,25	48,84	50,53
Cotas para Escolas Públicas e Raciais (<=1,5 SM)	-	-	-	8,93	13,24	13,00
Cotas para Escolas Públicas (<=1,5 SM)	-	-	-	6,69	7,08	7,73
Cotas para Escolas Públicas e Raciais	-	-	-	9,94	14,00	12,60
Modalidade de Ingresso	**7222**	**7223**	**7649**	**7844**	**7470**	**7571**
Portador de Diploma de Curso Superior (DCS)	0,15	0,69	0,04	0,04	0,07	0,03
ENEM	8,40	2,63	15,15	23,20	24,54	24,57
PAS	21,95	14,59	24,42	25,19	26,20	44,01
Vagas remanescentes	9,04	5,55	11,07	4,44	2,52	3,24
Transferência	2,34	0,08	0,04	0,05	0,07	0,11
Vestibular	58,11	76,45	49,27	47,08	46,61	28,05
Campus	**7278**	**7223**	**7649**	**7845**	**7470**	**7571**
Darcy Ribeiro	83,76	83,77	82,64	83,08	83,33	83,00
Ceilândia	6,13	6,34	7,32	6,78	6,96	6,86
Gama	6,39	6,88	6,84	6,82	6,33	6,47
Planaltina	3,72	3,00	3,20	3,31	3,37	3,67
Turno	**7278**	**7223**	**7649**	**7845**	**7470**	**7571**
Diurno	71,75	71,38	73,13	73,05	72,28	71,89
Noturno	28,25	28,62	26,87	26,95	27,72	28,11
Área de Ingresso	**7275**	**7221**	**7648**	**7838**	**7469**	**7571**
Ciências Exatas e da Terra	9,62	10,29	10,13	9,79	9,49	10,05
Ciências Biológicas	2,05	1,99	2,14	1,94	1,91	1,86
Engenharias	16,95	17,30	16,80	17,33	16,49	15,97
Ciências da Saúde	14,90	15,07	15,70	15,26	15,92	15,56
Ciências Agrárias	9,33	8,66	8,98	9,05	9,25	9,52
Ciências Sociais Aplicadas	25,25	24,30	24,76	24,23	24,10	22,81
Ciências Humanas	11,56	11,62	10,90	11,37	11,06	10,82
Linguística, Letras e Artes	10,34	10,77	10,56	11,05	11,77	13,41
Necessidade Especial	**7202**	**6505**	**2982**	**6307**	**7003**	**7108**
Não	97,61	95,13	96,28	95,31	95,96	94,70
Sim	2,39	4,87	3,72	4,69	4,04	5,30

[1]Fonte elaboração própria, dados Observatório da Vida Estudantil.
[2]Células coloridas, com cores equivalentes em uma variável foram agrupadas, em uma mesma categoria.

Tabela 29: Perfil sociodemográfico.

Características	2012	2013	2014	2015	2016	2017
Sexo	**7278**	**7223**	**7653**	**7848**	**7414**	**7571**
Masculino	45,15	50,81	49,71	50,93	50,18	50,59
Feminino	51,53	49,19	46,36	46,46	49,54	49,12
Ignorado	3,33	-	3,93	2,61	0,28	0,29
Nacionalidade	**6993**	**7223**	**7345**	**7641**	**7413**	**7571**
Brasileiro(a)	99,73	99,71	99,88	99,80	99,84	99,76
Estrangeiro	0,27	0,29	0,12	0,20	0,16	0,24
Região de Nascimento	**6919**	**6929**	**6968**	**7288**	**7295**	**7445**
Norte	1,89	2,38	1,95	2,11	2,08	1,95
Nordeste	9,25	8,34	8,38	7,93	7,98	7,21
Sudeste	12,66	12,53	11,80	11,91	10,57	10,03
Sul	2,40	2,57	2,18	2,03	1,82	1,83
Centro-Oeste	73,80	74,18	75,69	76,02	77,55	78,98
Estado Civil	**6964**	**7223**	**7344**	**7640**	**7413**	**7571**
Solteiro(a)	94,82	93,37	94,38	93,89	93,84	95,02
Casado(a)	3,49	4,33	3,73	4,06	3,95	3,29
União Estável	1,19	1,52	1,18	1,02	1,36	1,10
Divorciado(a)	0,45	0,64	0,59	0,84	0,63	0,52
Separado(a)	0,04	0,11	0,10	0,14	0,19	0,08
Viúvo(a)	0,01	0,03	0,03	0,05	0,03	-
Raça/Cor/Etnia	**6857**	**7223**	**7343**	**7639**	**7413**	**7571**
Branca	51,28	49,36	48,73	47,00	42,52	44,49
Parda	34,75	34,81	38,17	39,08	44,22	41,94
Preta	9,28	11,16	9,71	10,70	10,32	10,84
Amarela	3,14	3,25	2,74	2,57	2,52	2,32
Indígena	0,66	0,73	0,65	0,67	0,42	0,41
Outros	0,90	0,69	-	-	-	-
Região de Residência	**7009**	**7161**	**7227**	**7581**	**7364**	**7522**
Norte	0,50	0,31	0,40	0,49	0,49	0,48
Nordeste	1,36	0,88	1,19	1,08	1,28	1,34
Sudeste	2,24	1,94	2,27	2,78	2,65	2,33
Sul	0,16	0,15	0,17	0,30	0,24	0,15
Centro-Oeste	95,75	96,72	95,97	95,34	95,34	95,71
Com Quem Reside	**6931**	**7111**	**7112**	**7453**	**7229**	**7571**
Com os pais (pai e/ou mãe)	85,50	82,13	83,86	83,08	85,53	87,58
Com parentes ou amigos(as)	7,08	8,00	8,56	9,10	4,99	4,35
Cônjuge	1,30	-	-	0,03	-	-
Cônjuge, com filhos	1,99	2,98	2,11	0,66	3,18	2,42
Cônjuge, sem filhos	1,28	2,32	1,63	0,81	2,13	1,82
Em pensão	0,16	0,28	0,08	0,04	0,14	0,1
Em república	0,39	0,48	0,41	0,55	0,40	0,43
Em Residência Estudantil - CEU	0,03	0,04	0,04	0,05	0,08	0,07
Sozinho	2,21	3,49	2,78	3,60	2,96	2,88
Outros	0,06	0,28	0,52	2,09	0,00	-
Com filhos(as)	-	-	-	-	0,58	0,35

Tabela 30: Perfil Econômico Familiar do Estudante.

Características	2012	2013	2014	2015	2016	2017
Renda Mensal Familiar	**6712**	**7222**	**7309**	**7611**	**7406**	**7571**
Até 3 SM	21,86	20,67	21,34	23,77	31,76	30,68
De 3 a 10 SM	36,04	36,13	36,09	36,08	37,13	36,98
De 10 a 20 SM	21,16	22,04	22,31	21,93	18,59	18,74
Mais de 20 SM	16,40	17,57	20,25	18,22	12,52	13,59
Não possui renda mensal	4,54	3,59	-	-	-	-
N° de Pessoas que Vivem da Renda	**6884**	**7223**	**7299**	**7587**	**7407**	**7534**
1	4,56	6,44	5,44	5,85	5,86	5,46
2	12,35	13,58	12,86	13,87	13,30	12,97
3	20,74	21,49	21,22	22,63	22,45	23,64
4	35,28	32,99	34,99	33,28	34,31	35,00
5 OU MAIS	27,06	25,50	25,48	24,37	24,09	22,94
Escolaridade do Pai	**7114**	**7212**	**7397**	**7473**	**7252**	**7408**
Não sabe ler nem escrever	1,55	0,62	0,82	1,10	1,14	0,86
Ensino fundamental incompleto	14,79	11,08	10,60	12,73	15,22	13,69
Ensino fundamental completo	4,79	3,94	3,58	4,64	4,91	3,94
Ensino médio incompleto	3,74	3,90	3,76	4,22	4,81	4,54
Ensino médio completo	23,56	23,32	23,25	25,16	24,13	25,11
Ensino superior incompleto	6,09	6,10	6,27	5,46	5,02	5,59
Ensino superior completo	25,63	28,30	27,63	24,92	23,68	24,38
Pós-graduação	16,05	18,54	20,22	17,38	16,13	17,10
Não sabe informar;	3,81	4,20	3,85	4,40	4,95	4,79
Escolaridade da Mãe	**7176**	**7220**	**7411**	**7475**	**7292**	**7447**
Não sabe ler nem escrever	1,52	0,58	0,46	0,62	0,70	0,64
Ensino fundamental incompleto	12,40	8,31	8,22	10,27	11,46	10,08
Ensino fundamental completo	4,46	2,81	3,12	4,00	3,74	3,50
Ensino médio incompleto	4,07	3,67	4,34	4,86	5,49	4,67
Ensino médio completo	25,78	26,22	25,62	26,15	27,72	27,30
Ensino superior incompleto	6,41	5,94	7,08	6,31	5,83	6,20
Ensino superior completo	25,31	27,83	26,58	25,23	23,68	25,73
Pós-graduação	19,24	23,88	23,73	21,69	20,42	20,92
Não sabe informar;	0,81	0,76	0,84	0,87	0,96	0,94
Convenio ou Plano de Saúde	**7278**	**7223**	**7653**	**7848**	**7470**	**7571**
Sim	58,24	64,41	61,60	56,33	52,21	52,74
Não	31,48	32,88	29,71	33,72	40,39	40,95
Ignorado	10,28	2,71	8,69	9,95	7,40	6,31
Assistência Médica	**1987**	**2212**	**2198**	**2491**	**2827**	**2898**
Sobretudo à rede pública	87,02	88,02	90,63	91,49	92,47	92,82
Sobretudo à rede privada	12,98	11,98	9,37	8,51	7,53	7,18

Tabela 31: Trajetória escolar e características do ingresso na UnB.

Características	2012	2013	2014	2015	2016	2017
Ensino Fundamental	**6927**	**7223**	**7246**	**7566**	**7391**	**7571**
Somente em escolas públicas	30,49	31,40	32,76	36,90	41,39	37,64
Sobretudo em escolas públicas	9,04	9,65	9,63	10,15	10,11	9,72
Sobretudo em escolas particulares com bolsa	1,89	1,98	2,28	2,33	2,08	2,55
Sobretudo em escolas particulares	10,74	11,37	9,80	9,40	8,93	9,03
Somente em escolas particulares com bolsa	3,10	2,95	3,28	3,15	2,94	3,29
Somente em escolas particulares	44,74	42,66	42,24	38,08	34,56	37,76
Ensino Médio	**6933**	**7223**	**7246**	**7566**	**7391**	**7571**
Somente em escolas públicas	34,03	35,15	38,67	44,78	51,98	49,28
Sobretudo em escolas públicas	2,93	3,78	2,77	2,91	2,21	2,22
Sobretudo em escolas particulares com bolsa	1,83	1,79	2,10	1,64	1,46	1,89
Sobretudo em escolas particulares	3,88	3,97	4,07	3,56	2,69	2,79
Somente em escolas particulares com bolsa	6,26	6,41	6,67	6,13	5,43	5,94
Somente em escolas particulares	51,07	48,90	45,72	40,99	36,23	37,88
Tipo de Ensino	**6927**	**7222**	**7244**	**7566**	**7391**	**7571**
Ensino Médio regular	88,02	84,80	86,06	89,28	91,03	90,79
Supletivo	5,11	6,95	6,34	2,34	1,80	1,65
Educação de Jovens e Adultos	3,36	4,04	3,41	2,84	2,41	2,80
Técnico/Profissionalizante	2,50	2,98	3,05	4,15	3,65	3,98
Magistério	0,45	0,47	0,41	0,40	-	-
Exame de massa/menção	0,32	0,44	0,47	0,75	0,99	0,75
Telecurso	0,25	0,32	0,26	0,24	0,12	0,03
Realizou Curso Preparatório	**7278**	**7223**	**7653**	**7848**	**7470**	**7571**
Sim	34,87	51,09	38,6	39,5	41,98	36,26
Não	56,72	48,84	58,89	56,94	56,95	63,74
Ignorado	8,41	0,07	2,51	3,56	1,07	-
Tipo de Curso Preparatório[1]	**3188**	**3054**	**2775**	**3008**	**3060**	**2685**
Público ou gratuito	1,69	1,57	3,24	4,19	8,66	16,09
Particular com bolsa integral	3,7	3,14	3,96	4,12	3,1	4,06
Particular com bolsa parcial	23,12	25,44	25,69	22,47	22,22	21,38
Particular	71,49	69,84	67,1	69,22	66,01	58,47
Número de Tentativas[1]	**3197**	**7202**	**7432**	**7550**	**7374**	**7571**
Nenhuma, esta é a primeira	17,92	24,88	37,19	28,75	39,54	49,47
Uma	10,54	16,33	18,03	17,96	14,42	13,01
Duas	15,64	23,44	19,52	21,4	18,71	14,79
Três	13,01	18,18	11,93	12,78	10,89	8,32
Mais de três	19,05	12,8	7,19	9,62	8,58	6,43
Já realizei, s/c, outro(s) curso(s) na UnB	14,86	4,35	4,83	7,25	5,87	5,46
Já conclui outro(s) curso(s) na UnB	8,98	0,03	1,31	2,24	1,98	2,52
É o curso desejado	**7278**	**7223**	**7653**	**7848**	**7470**	**7571**
Sim	48,86	58,16	62,69	56,35	57,31	56,32
Não	40,19	41,52	34,42	39,86	41,38	43,68
Ignorado	10,95	0,32	2,89	3,8	1,31	-
Trocaria de curso	**6624**	**7223**	**7187**	**7548**	**7370**	**7571**
Sim	32,73	31,75	33,6	34,09	40,83	41,25
Não	67,27	68,25	66,4	65,91	59,17	58,75

Tabela 32: Carreira Profissional e Expectativas.

Características	2012	2013	2014	2015	2016	2017
Grau Máximo de Estudo Pretendido	**6785**	**7223**	**6862**	**7544**	**7366**	**7571**
Completar o ensino superior (graduação)	6,59	8,42	4,10	8,56	9,22	10,62
Completar uma pós-graduação (doutorado)	63,08	60,58	63,00	60,35	59,23	57,07
Completar uma pós-graduação (especialização)	13,68	13,61	14,82	13,11	14,12	14,04
Completar uma pós-graduação (mestrado)	15,42	16,03	16,55	16,38	15,15	15,93
Indefinido	0,88	0,94	0,76	1,03	0,00	0,00
Outros	0,35	0,42	0,77	0,56	2,28	2,34
Atividade Econômica	**6778**	**7223**	**7181**	**7527**	**0**	**0**
Nunca trabalhei	54,79	51,09	53,56	51,68	-	-
Não trabalho no momento	24,73	22,39	24,76	26,42	-	-
Exerço atividades remuneradas eventualmente	2,67	5,26	4,37	4,57	-	-
Faço estágio técnico	2,20	2,42	1,62	1,41	-	-
Trabalho sem carteira assinada	3,45	3,30	2,70	2,39	-	-
Trabalho com carteira assinada	8,96	8,75	7,26	7,69	-	-
Sou servidor(a) público(a)	3,20	6,80	5,74	5,83	-	-
Renda Mensal	**6546**	**6556**	**6653**	**6666**	**6172**	**6474**
Até 3 SM	13,52	14,43	12,13	12,27	-	-
De 3 a 10 SM	3,39	2,97	2,13	2,61	-	-
De 10 a 20 SM	0,69	0,56	0,53	0,72	-	-
Mais de 20 SM	0,41	0,29	0,08	0,18	-	-
Não possui renda mensal	81,56	80,95	84,53	83,45	99,74	99,60
Não sei	0,43	0,79	0,60	0,77	0,26	0,40
Horas de Trabalho	**1078**	**1250**	**1029**	**1102**	**1199**	**1123**
Menos de 10 horas	17,90	21,20	23,52	21,78	11,18	12,64
Entre 10 e 20 horas	14,75	15,36	15,26	14,79	13,18	12,38
Entre 20 e 30 horas	14,47	13,76	14,19	13,34	15,01	13,54
Entre 30 e 40 horas	33,40	31,76	28,67	29,58	37,20	39,36
Mais de 40 horas	19,48	17,92	18,37	20,51	23,44	22,08
Perspectiva Profissional	**6868**	**7223**	**6994**	**7526**	**7354**	**7571**
Ainda não me decidi	15,48	15,69	16,61	17,69	17,76	18,87
Atividade na minha área de graduação	42,59	42,27	42,71	42,59	43,21	42,66
Lecionar para ensino fundamental ou médio	2,50	2,49	2,83	2,68	3,20	3,18
Lecionar para ensino superior	3,26	3,52	3,75	4,07	3,66	4,13
Pretendo trabalhar em empresa privada	4,34	4,49	2,02	3,52	3,48	3,50
Vou prestar concurso	27,27	25,34	26,18	23,04	22,95	21,15
Já tenho trabalho e pretendo continuar nele*	0,79	1,37	1,32	1,26	0,92	1,15
Já tenho trabalho e pretendo continuar nele**	3,23	4,37	3,93	4,12	3,60	4,17
Outros	0,54	0,47	0,66	1,04	1,21	1,18

[1]*Fora da área de graduação; **Dentro da área de graduação.

Tabela 33: Atividades Extracurriculares.

Características	2012	2013	2014	2015	2016	2017
Atividades Extracurriculares	**22290**	**34062**	**33924**	**33894**	**32604**	**45426**
Idiomas	54,10	62,99	62,65	58,59	58,37	59,54
Música	16,50	17,00	17,26	14,07	13,16	12,24
Esportes	42,61	43,19	42,84	39,58	37,65	38,17
Artes Visuais	3,55	3,77	3,24	3,35	2,96	3,05
Dança	8,34	9,62	8,91	7,95	7,05	7,36
Teatro	7,16	6,80	6,74	6,25	5,89	6,05

Tabela 34: Motivos para a Escolha do Curso pelos Ingressantes.

Características	2012	2013	2014	2015	2016	2017
Motivo da Escolha do Curso	**70585**	**107863**	**107426**	**107331**	**103246**	**143849**
Vocação ou aptidões pessoais	48,75	57,37	54,09	54,82	55,65	58,54
Gosto pessoal	67,21	75,87	70,53	72,47	74,07	76,28
Pela exigência intelectual	10,71	15,45	14,26	15,76	15,75	15,78
Disponibilidade de vagas no mercado	22,45	24,45	22,85	22,38	21,94	22,51
Possibilidades salariais	24,95	27,23	25,19	26,43	23,76	21,89
Baixa concorrência no vestibular	8,43	10,48	8,86	9,19	9,44	9,38
Possibilidade de realização pessoal	38,63	46,47	44,66	44,82	46,34	46,72
Possibilidade de contribuir	26,57	32,32	31,34	34,11	38,72	37,85
Pela responsabilidade ética	6,65	10,99	10,17	10,55	12,53	12,52
Tenho facilidade de acesso ao mercado	5,22	7,66	6,76	7,12	6,55	7,08
Proporciona um emprego seguro	7,89	10,27	9,32	9,82	9,61	9,31
A profissão permite auton. na gestão do tempo	3,07	5,60	5,43	5,77	5,91	6,38
Exclusão, os outros cursos não me agradavam	4,47	4,67	4,92	4,90	5,70	6,39
Exclusão, o curso de preferência não era oferecido	2,02	1,62	1,26	1,65	1,53	1,72
Por indicação de teste vocacional	3,39	4,83	4,60	4,55	4,53	4,70
Tradição familiar	1,67	2,41	1,98	1,61	1,99	2,06
Desejo da família	2,45	2,85	2,78	3,36	3,26	3,87
Influência de amigos ou terceiros	6,16	8,31	8,21	7,79	8,30	7,95
Complementar formação prof. que já exerce	2,31	4,79	3,63	4,14	4,36	4,74

Tabela 35: Motivos do não Ingresso no Curso Desejado.

Características	2012	2013	2014	2015	2016	2017
Motivos	**14860**	**22708**	**21736**	**22593**	**21736**	**30284**
Concorrência e falta de preparo	12,17	15,31	13,49	14,21	19,69	23,35
Tentativas anteriores sem sucesso	8,75	13,85	8,07	9,97	7,91	8,56
Já cursei outros cursos sem concluir	8,21	6,69	4,40	6,27	4,91	5,44
Já concluiu outros cursos	8,08	0,19	1,54	2,80	2,80	3,39

[1]Fonte elaboração própria, dados Observatório da Vida Estudantil.