Universidade de Brasília

IE - Instituto de Exatas

Departamento de Estatística

Perfil dos estudantes da área de saúde da UnB:
Uma aplicação em Regressão Logística Politômica

Fernanda Luiza Rodrigues de Albuquerque

Relatório Final do Projeto Final

Orientadora: Profª Maria Teresa Leão Costa

Brasília

Junho de 2015

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução e Justificativa

A grande necessidade de pesquisas em diversas áreas do conhecimento fazem da estatística uma importante aliada na hora da tomada de decisões, inferências, previsões, entre outras coisas. Técnicas estatísticas são constantemente aplicadas para ajudar na obtenção e análise de informações para o estudo adequado de fenômenos. Utilizando-se de análises iniciais, como medidas-resumo e gráficos e posteriormente análises mais complexas como, por exemplo, modelagem, é possível obter conclusões do que se quer estudar.

Uma técnica estatística muito usada é a regressão, onde a partir de uma variável dependente(resposta) e uma ou mais variáveis independentes (explicativas) pode-se descrever a relação entre essas variáveis. A regressão pode ser simples(quando há apenas uma variável explicativa), ou múltipla(quando há mais de uma variável explicativa).

Uma ferramenta estatística muito útil é a análise de dados categorizados, na qual trabalha-se com variáveis resposta categóricas, aquelas que são mensuradas por categorias. O tipo de regressão mais adequada para se trabalhar nessa análise é a regressão logística, pois tem-se interesse em estimar probabilidade de um evento ocorrer como função de outros fatores.

A regressão logística mais conhecida talvez seja a dicotômica, ou seja, a

variável resposta tem duas categorias e consequentemente segue uma distribuição binomial. Porém em alguns estudos a variável resposta não terá apenas duas categorias. Para esse caso utiliza-se a regressão logística politômica, também chamada de logística multinomial, na qual a variável resposta tem mais de duas categrias e segue uma distribuição multinomial, sendo portanto, a regressão logística dicotômica um caso especial da politômica, para categoria igual a 2.

Esse tipo de regressão possui aplicação em diversas áreas do conhecimento. Com frequência é observada sua utilização em estudos relacionados à saúde e a educação, os quais lidam muito com variáveis resposta categóricas. A partir da relação entre as variáveis envolvidas é possível obter um grande ganho de informação a respeito do fenômeno que se está estudando, o que a faz de extrema importância.

O trabalho tem como foco a regressão logística politômica, bem como sua aplicação e modelagem.

Capítulo 2

Objetivos

I Objetivo geral

O trabalho tem por objetivo geral o estudo da teoria relacionada à análise de regressão logística politômica e desenvolvimento de uma aplicação em um estudo sobre características dos alunos ingressantes nas faculdades que ofertam cursos na área de saúde na Universidade de Brasília.

II Objetivos específicos

- Estudo da teoria(estimação, intervalos de confiança, testes de hipótese, interpretação do modelo, etc.);
- Realizar análises descritivas e inferência estatística;
- Interpretação dos resultados para traçar um perfil dos alunos que ingressam em cada uma das faculdades que ofertam cursos na área de saúde.

Capítulo 3

Referencial Teórico

3.1 Regressão Logística Dicotômica

Regressão é uma técnica muito utilizada quando se deseja estudar a relação entre uma variável resposta e uma ou mais variáveis explicativas. Quando tem-se um estudo onde define-se uma variável categórica como resposta, a técnica ideal a ser utilizada é a regressão logística. Uma variável categórica é aquela cujos valores representam um conjunto de categorias, podendo a escala de mensuração ser nominal ou ordinal. Primeiramente será abordado o estudo da regressão logística dicotômica para facilitar o entendimento da técnica. Posteriormente será abordado a regressão logística politômica nominal.

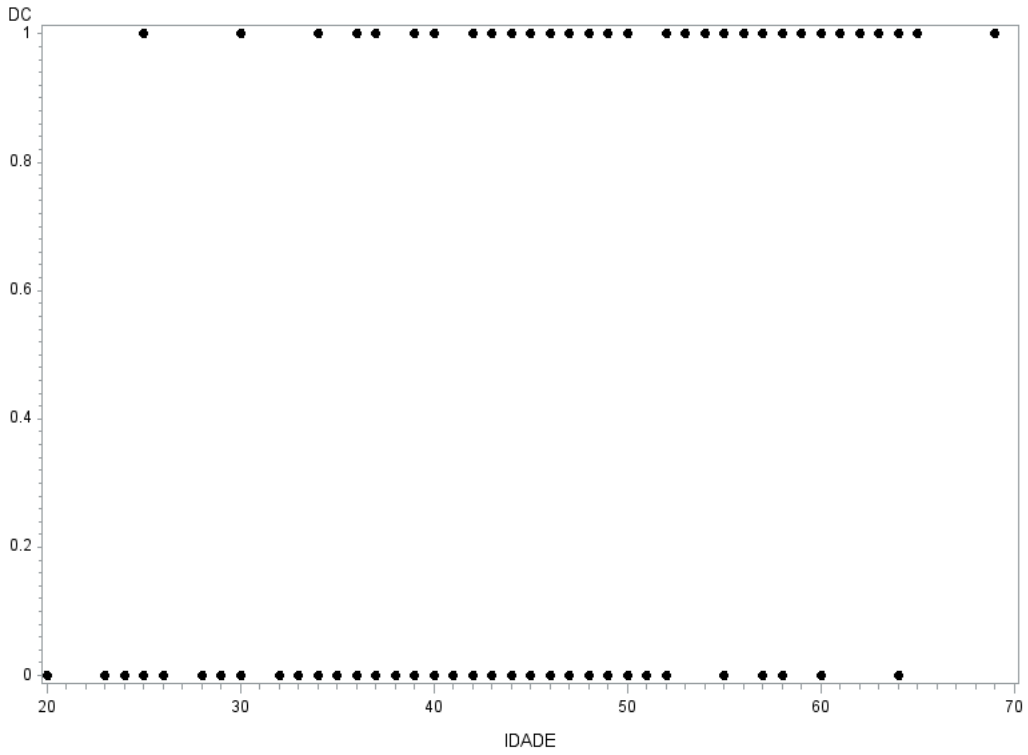
3.1.1 Introdução

O foco desta seção é estudar os modelos com resposta dicotômica(binária), ou seja, aqueles em que os possíveis resultados são "sucesso"e "fracasso". Denotemos uma variável resposta por Y e uma variável explicativa por X. A distribuição de Y é $P(Y = 1) = \pi$ probabilidade de sucesso e $P(Y = 0) = 1 - \pi$ probabilidade de fracasso. O valor de π pode variar de acordo com o valor de X, por essa razão π será representado por $\pi(x) = P(Y = 1|X = x)$ que é a

probabilidade de sucesso quando $X = x$. Para ilustrar considera-se o exemplo do estudo sobre presença de doença cardíaca coronária em função da idade, encontrado em Hosmer e Lemeshow,1989.

EXEMPLO

Um estudo com 100 pessoas foi realizado para explorar a relação entre idade(IDADE) e presença ou ausência de doença coronária(DC). A variável resposta do estudo é DC e tem-se que (y=1) para presença de doença e (y=0) para ausência de doença. A variável IDADE assume valores de 20 à 69 anos. Construindo um gráfico de dispersão teria-se o seguinte resultado:

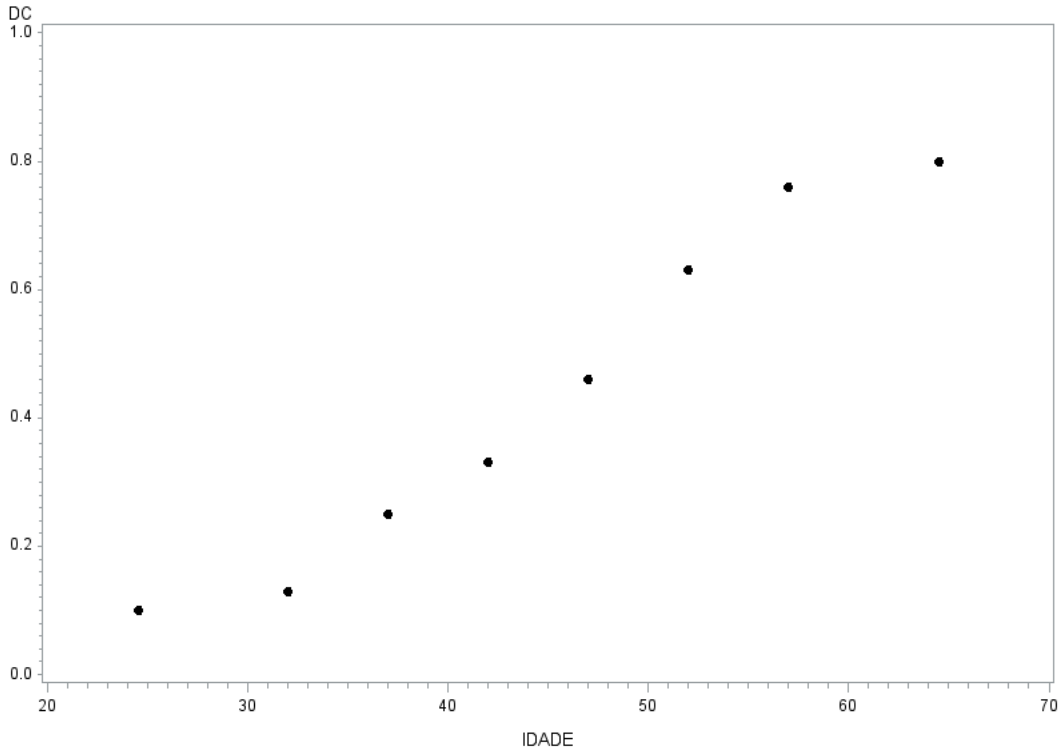


Percebe-se que o gráfico não nos dá uma visão muito boa de como se dá a

relação entre as duas variáveis. Um dos problemas é justamente a alta variabilidade da variável IDADE. Uma forma de manter a estrutura da relação entre as variáveis e remover alguma variação é criar intervalos para a variável explicativa e calcular a proporção de pessoas com doença dentro de cada grupo. Essa proporção será uma estimativa de $\pi(x)$

Tabela 3.1: Proporção de pessoas com a doença

		DC		
Grupo de Idade	n	Não	Sim	Proporção
20-29	10	9	1	0,1
30-34	15	13	2	0,13
35-39	12	9	3	0,25
40-44	15	10	5	0,33
45-49	13	7	6	0,46
50-54	8	3	5	0,63
55-59	17	4	13	0,76
60-69	10	2	8	0,80
Total	100	57	43	0,43



Observando a tabela e o gráfico percebe-se que com o aumento da idade há um aumento na proporção de pessoas com doenças. Mas é necessário descrever essa relação através de uma forma funcional.

Uma primeira possibilidade talvez seja modelar a probabilidade da seguinte forma:

$$\pi(x) = \alpha + \beta x \tag{3.1}$$

onde o parâmetro β representa a mudança na probabilidade por unidade mudada em x. É chamado de modelo de probabilidade linear e tem a forma de um modelo de regressão ordinária. Esse modelo é simples, mas tem um problema que é o fato de funções lineares darem valores na reta real$(-\infty, \infty)$, ou

seja, o modelo prediz $\pi(x) < 0$ e $\pi(x) > 1$, mas $\pi(x)$ é uma probabilidade e não pode ter valores fora do intervalo [0,1]. Portanto precisa-se de uma forma funcional de modelo para os casos de resposta categórica binária que considere este aspecto.

No gráfico 1.2 nota-se que a mudança na probabilidade para cada alteração de uma unidade em x se torna menor quando a probabilidade se aproxima de 0 ou 1. Isso dá um formato de S na curva, portanto a relação entre $\pi(x)$ e x é geralmente não linear. Várias possibilidades poderiam ser consideradas, mas a logística foi escolhida por ser uma função fácil e flexível de ser usada do ponto de vista matemático e também dada a interpretação de seus parâmetros.

A forma específica do modelo de regressão logística a ser usada é

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{(\beta_0+\beta_1 x)}} \tag{3.2}$$

Uma transformação de $\pi(x)$ que é importante para o estudo de regressão logística é a **transformação logito**, a qual é definida em termos de $\pi(x)$ como

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{3.3}$$

Essa transformação é importante pelo fato de $g(x)$ ter propriedades desejáveis do modelo de regressão linear, sendo $g(x)$ linear nos parâmetros e podendo ser um número real, assim não tem-se o problema do intervalo [0,1].

Portanto, na regressão logística dicotômica tem-se que a probabilidade de sucesso $\pi(x)$ é delimitada pelo intervalo [0,1]. A distribuição dos erros é binomial, e não normal como acontece no caso da regressão linear, a qual será a base das análises e os princípios que orientam a análise de regressão logística.

3.1.2 O modelo de regressão logística múltiplo

Na seção anterior, o modelo de regressão logística foi introduzido utilizando apenas uma variável explicativa para facilitar a apresentação das ideias. Porém, na prática, quando se quer realizar um estudo, não se terá apenas uma variável explicativa e sim muitas. Cada uma dessas variáveis podem ter naturezas diferentes (quantitativa ou qualitativa), o que exige uma maneira diferente de lidar com cada uma delas e que será introduzido mais adiante.

Considere que tenhamos p variáveis explicativas denotadas por $X_1, X_2, ..., X_p$. Agora tem-se que $\pi(\mathbf{x}) = P(Y = 1|x_1, x_2, ..., x_p)$ é a probabilidade da resposta estar presente dado o vetor $\mathbf{x} = (x_1, x_2, ..., x_p)$, então o logito do modelo de regressão logístico múltiplo é

$$g(\mathbf{x}) = ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{3.4}$$

e $\pi(\mathbf{x})$ é dado por

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{3.5}$$

Quando tem-se variáveis explicativas que são nominais, como por exemplo, sexo, estado civil e raça utiliza-se variáveis denominadas "dummy"para adicioná-las ao modelo. Para ilustrar, pega-se a variável estado civil como exemplo. Suponha que ela tenha sido categorizada como "solteiro", "casado"e "outros". Assim uma maneira de utilizar a variável dummy(r) seria o seguinte

$$r_1 = \ 1 \text{ para solteiro}, 0 \text{ caso contrário}$$

$$r_2 = \ 1 \text{ para casado}, 0 \text{ caso contrário}$$

Não há necessidade de colocar a terceira categoria(outros), pois esta será quando $r_1 = r_2 = 0$. Portanto um modelo contendo como variáveis expliativas o estado civil e a idade seria

$$g(\mathbf{x}) = ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 r_1 + \beta_2 r_2 + \beta_3 idade \tag{3.6}$$

Considerando uma variável explicativa nominal com k categorias, então, no geral serão necessárias apenas $k-1$ categorias.

Suponha que a j-ésima variável explicativa x_j tem k_j categorias. As $k_j - 1$ variáveis dummy serão denotadas por r_{ju} e os coeficientes dessas variáveis serão denotados por β_{ju} com $u = 1, 2, ..., k_j - 1$. Portanto uma forma geral de representar variáveis dummy em um modelo de regressão logística seria

$$g(\mathbf{x}) = ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \sum_{u=1}^{k_j-1} \beta_{ju} r_{ju} + \beta_p x_p \tag{3.7}$$

3.1.3 Estimação dos Parâmetros do Modelo

Na Regrassão Logística o método de estimação utilizado para estimar os parâmetros β_j do modelo é o método de máxima verossimilhança. Esse método utiliza o princípio da máxima verossimilhança que consiste em escolher aqueles valores dos parâmetros que maximizam a probabilidade de obter a amostra observada expressa pela função de verossimilhança.

Assumindo a independência das observações, a função de verossimilhança é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} (\pi(\mathbf{x}_i))^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i} \tag{3.8}$$

Porém matematicamente é mais fácil trabalhar com o log dessa expressão. Portanto a log-verossimilhança é definida como

$$l(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n}[y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]] \tag{3.9}$$

Para encontrar o valor de β que maximiza $L(\beta)$ calcula-se a derivada de $L(\beta)$ em relação a cada um dos parâmetros do modelo e iguala o resultado a zero

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)]$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)]$$

e portanto

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0$$

são chamadas de equações de verossimilhança. Essas equações são não lineares nos parâmetros e os estimadores não tem uma expressão com forma fechada, portanto necessitam de métodos numéricos de solução que são encontrados em muitos softwares. Utiliza-se o método de Newton-Raphson para obter a solução deste sistema.

A solução dessas equações será a estimativa de máxima verossimilhança $\hat{\beta}$, e para achar os valores ajustados(preditos) $\hat{\pi}(\mathbf{x}_i)$, basta substituir $\hat{\beta}$ na equação 3.5. Esta quantidade nos dá uma estimativa da probabilidade condicional de $Y = 1$ dado $x = x_i$. Pode-se também achar o logito estimado $\hat{g}(\mathbf{x})$ fazendo a mesma substituição em 3.4.

O método de estimação das variâncias e covariâncias dos estimadores dos

parâmetros segue da teoria de estimação de máxima verossimilhança. Essa teoria diz que os estimadores são obtidos da matriz de derivadas parciais de segunda ordem da função de log verossimilhança. Essas derivadas são

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} \mathbf{x}_{ij}^2 \pi(\mathbf{x}_i)(1 - \pi(\mathrm{x}_i)) \tag{3.10}$$

e

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} \mathbf{x}_{ij} \mathbf{x}_{iu} \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i)) \tag{3.11}$$

para j,u=0,1,2,...p

A matriz (p+1) x (p+1) que contém os valores negativos das expressões acima citadas é chamada de matriz de informação e é denotada por $I(\boldsymbol{\beta})$. A variância e covariância dos coeficientes estimados são obtidos do inverso dessa matriz que será denotada como $\Sigma(\beta) = I^{-1}(\beta)$. Assim, o j^{th} elemento da diagonal principal da matriz será denotado por $\sigma^2(\beta_j)$ que é a variância de $\hat{\beta}_j$ e os demais elementos da matriz será denotado por $\sigma(\beta_j, \beta_u)$ que é a covariância de $\hat{\beta}_j$ e $\hat{\beta}_u$. Os estimadores das variâncias e covariâncias, denotado por $\hat{\Sigma}(\boldsymbol{\beta})$ é obtido calculando $\Sigma(\beta)$ com os valores de $\hat{\beta}$. Para se referir aos valores dessa matriz usa-se $\hat{\sigma^2}(\hat{\beta}_j)$ e $\hat{\sigma}(\hat{\beta}_j, \hat{\beta}_u)$. Portanto o erro padrão estimado dos coeficientes estimados será

$$\hat{EP}(\hat{\beta}_j) = \sqrt{[\hat{\sigma^2}(\hat{\beta}_j)]} \tag{3.12}$$

3.1.4 Inferência para Regressão Logística

Após estimar os parâmetros do modelo, é necessário realizar inferências à respeito do mesmo. Nessa seção o foco será mostrar técnicas utilizadas para essa finalidade e que ajudam a julgar o efeito das variáveis do modelo.

3.1.4.1 Significância do modelo

Inicialmente é necessário verificar a significância das variáveis do modelo. Para isso são necessários testes estatísticos de hipóteses que ajudam a identificar se as variáveis explicativas do modelo são significativamente relacionadas à variável resposta. Um motivo para se realizar esses testes é o fato de se querer saber se uma variável no modelo pode dizer mais sobre a variável resposta do que se ela não estivesse no modelo. Para verificar essa hipótese pode-se comparar os valores observados da variável resposta com aqueles preditos em um modelo que contenha a variável em questão e em outro que não contenha a variável. Em regressão logística, a comparação entre os valores observados e preditos é baseada na função de log-verossimilhança, já mostrada em 3.9. Para comparar o valor esperado e predito usando a função de verossimilhança tem-se a seguinte estatística

$$G = -2\ln\left[\frac{\text{verossimilhança sem a variável}}{\text{verossimilhança com a variável}}\right] = -2\ln(\frac{L_0}{L_1}) = -2(l_0 - l_1) \quad (3.13)$$

Esse teste é chamado de Teste da Razão de Verossimilhança e a estatística conhecida também por **deviance**, onde l_0 é a verossimilhança do modelo reduzido e l_1 a verossimilhança do modelo completo. Em regressão logística pode-se utilizar essa estatística para testar várias hipóteses à respeitos dos parâmetros do modelo.

Se o interesse é verificar se os p parâmetros do modelo são iguais à zero, a distribuição de G será qui-quadrado com p graus de liberdade, sendo as hipóteses

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : \text{Pelo menos um } \beta_j \neq 0$$

e considerando um nível de significância $\alpha = 5\%$, se o p-valor obtido na análise for menor que esse valor, ocorrerá a rejeição de H_0 e portanto pelo menos um β_j do modelo será diferente de zero.

Uma análise também importante seria testar os parâmetros do modelo separadamente. Para isso pode-se também utilizar a deviance com distribuição qui-quadrado com 1 grau de liberdade sob as hipóteses

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

assim faria-se um teste individualmente para cada parâmetro β_j do modelo. Portanto, utilizando esse teste tem-se uma noção de quais variáveis do modelo são ou não são significativas.

Um teste semelhante também conhecido para testar essas hipóteses é o **teste univariado de Wald**, cuja estatística é dada por

$$W_j = \frac{\hat{\beta}_j}{\hat{EP}(\hat{\beta}_j)} \sim N(0,1) \tag{3.14}$$

que é usado para se verificar as mesmas conclusões do teste relatado acima.

Um dos interesses em regressão é achar o melhor modelo que contenha o mínimo de parâmteros necessários. Portanto seria interessante comparar o modelo que contenha todas as variáveis explicativas(completo) com o modelo que contém menos variáveis (reduzido), sendo um modelo caso particular do outro. Para realizar essa comparação usamos a deviance com distribuição qui-quadrado e grau de liberdade igual a diferença do número de parâmetros entre os modelos. As hipóteses são

$$H_0 : \text{O modelo reduzido se ajusta tão bem quanto o completo}$$

$$H_1 : \text{O modelo completo se ajusta melhor que o reduzido}$$

Caso a hipótese nula não seja rejeitada, o modelo reduzido seria tão bom quanto o completo, portanto parece ser razoável utilizar o modelo reduzido no lugar do completo. Porém não deve-se basear a escolha do modelo apenas em testes de significância. Mais adiante no trabalho serão abordadas outras considerações que influenciam a decisão de se retirar ou não variáveis do modelo.

Se um modelo não for caso particular do outro, a comparação pode ser dada através de medidas conhecidas como AIC(Critério de Informação de Akaike) e BIC(Critério de Informação Bayesiano), as quais são definidas abaixo

$$\text{AIC} = \text{-2(log-verossimilhança - n}^{\underline{o}} \text{ de parâmetros no modelo)}$$

$$\text{BIC} = \text{-2(log-verossimilhança - n}^{\underline{o}} \text{ de parâmetros no modelo ln(n))}$$

onde p é o número de parâmetros e n é o tamanho da amostra.

3.1.4.2 Intervalos de Confiança

Um intervalo de confiança de $100(1-\alpha)\%$ para os parâmetros β_j do modelo é

$$\hat{\beta}_j \pm Z_{\alpha/2}\hat{EP}(\hat{\beta}_j)$$

exponenciando os extremos do intervalo encontra-se um intervalo para e^{β_j}.

O logito é dado pela expressão em 3.4 e portanto o logito estimado é dado por

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

e sua variância é dada por

$$\hat{\sigma^2}(\hat{g}(\mathbf{x})) = \sum_{j=0}^{p} x_j^2 \hat{\sigma^2}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{u=j+1}^{p} 2x_j x_u \hat{cov}(\hat{\beta}_j, \hat{\beta}_u)$$

.

Assim um intervalo de $100(1-\alpha)\%$ para o logito é

$$\hat{g}(\mathbf{x}) \pm Z_{\alpha/2} \hat{EP}(\hat{g}(\mathbf{x}))$$

Pode-se achar a partir do intervalo do logito um intervalo para a probabilidade $\pi(\mathbf{x})$. Basta pegar os pontos extremos do intervalo do logito e substituir na expressão dada pela probabilidade $\pi(\mathbf{x}) = \dfrac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}$.

3.1.4.3 Interpretação dos Parâmetros do Modelo

Nas seções anteriores foi mostrado o modelo de regressão logística dicotômica, estimação e alguns testes utilizados para a avaliação dos parâmetros. Agora o foco é a interpretação desses parâmetros. Os coeficientes estimados para as variáveis explicativas simbolizam a inclinação ou a taxa de variação da variável resposta por unidade mudada na variável explicativa.

Em regressão logística $\beta_1 = g(x+1) - g(x)$, ou seja o coeficiente de inclinação representa a variação no logito para uma unidade mudada em x. Portanto β_1 representa a taxa de crescimento ou decrescimento na curva "S"($\pi(x)$), seu sinal indica se a curva cresce ou decresce conforme x cresce ou decresce e sua magnitude determina o quão rápido a curva cresce ou decresce. Considerando um $\beta_1 = 0$, a curva se transforma em uma linha reta horizontal, assim o valor de $\pi(x)$ seria o mesmo para todos os valores de x e portanto a resposta dicotômica Y é independente de X.

3.1.4.4 Razão de chances

A chance de "sucesso", isto é, de $Y = 1$ é dada pela razão entre a probabilidade de sucesso e a probabilidade de fracasso

$$\frac{\pi(x)}{1 - \pi(x)}$$

Lembrando-se da transformação logito dada em 3.3, tem-se que

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x} = e^{\beta_0}(e^{\beta_1})^x$$

uma interpretação que se pode retirar é que a cada unidade acrescida em x a chance é multiplicada por e^{β_1}, ou seja, a chance no nível $x + 1$ é igual ao nível x multiplicado por e^{β_1}. Caso $\beta_1 = 0$, então $e^{\beta_1} = 1$ e a chance não mudaria conforme x mude.

A razão de chances é a chance de sucesso em determinado grupo(categoria) em relação a chance de sucesso em outro grupo(categoria). Por exemplo, considerando π_1 e π_2 a probabilidade de sucesso no grupo 1 e 2 respectivamente, então a razão de chances é dada por

$$\psi = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \tag{3.15}$$

A razão de chances é sempre positiva. Quando $\psi = 1$ tem-se que X e Y são independentes, pois π_1=π_2 e portanto o "sucesso"(evento de interesse) é igualmente provável de ocorrer nos dois grupos. Assim $\psi = 1$ serve como uma referência para comparação, sendo valores de ψ no intervalo (1,∞) um indicativo de que a chance de sucesso é maior no grupo 1 do que no grupo 2($\pi_1 > \pi_2$) e em contra partida, se ψ estiver em um intervalo entre (0,1) a chance de sucesso no grupo 1 é menor que no grupo 2($\pi_2 > \pi_1$). Então considerando, por

exemplo, que $\psi = 6$, a chance de sucesso no grupo 1 é 6 vezes a chance de sucesso no grupo 2.

A razão de chances é uma medida de associação que tem uso muito extenso, principalmente em áreas da saúde, como é o caso da epidemiologia. O fato da técnica ajudar a verificar o quão mais provável ou não a resposta esteja presente para os valores $x = 1$ que para valores $x = 0$ é de extrema importância no que se refere à medidas preventivas e tratamentos mais elaborados. Um exemplo seria a conscientização da população em relação ao fumo após pesquisas revelarem que pessoas que fumam tem maior chance de desenvolver câncer de pulmão.

3.2 Regressão Logística Politômica

Na **regressão logística politômica**, a variável resposta é multicategorizada, ou seja, apresenta mais de duas categorias. Portanto a resposta agora tem distribuição multinomial e não mais binomial como no caso da dicotômica, mas pode-se verificar que a regressão logística dicotômica é um caso especial da regressão logística politômica para quando se tem 2 categorias na variável resposta.

3.2.1 O modelo

Supondo que a resposta Y é uma variável nominal com J categorias, a ordem das categorias é irrelevante por não apresentar nenhuma ordenação natural e denotando $\pi_1, \ldots, \pi_J$ como sendo a probabilidade de resposta onde $\sum_j \pi_j = 1$. A distribuição de probabilidade para o número de respostas que ocorrem para cada uma das J categorias é a multinomial. O modelo logito politômico se refere à todos os pares de categorias e relatam a chance de resposta em uma categoria no lugar de outra. O modelo compara cada categoria

da resposta com uma categoria de referência, a qual pode ser escolhida de forma arbritária. Quando a última categoria J é escolhida como referência, os logitos serão

$$ln\left(\frac{\pi_j}{\pi_J}\right) \qquad para \quad j = 1, ..., J-1$$

Dado que a resposta caia na categoria j ou J este é o logaritmo da chance que a resposta é j. O modelo logito tem a forma

$$g_j(\mathbf{x}) = ln\left(\frac{\pi_j}{\pi_J}\right) = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p \tag{3.16}$$

com $j = 1, ..., J-1$

O modelo consiste em $J-1$ logitos, com parâmteros separados para cada. Por exemplo, considerando $J = 3$ categorias, sendo a categoria 3 a de referência e codificando a variável resposta como 0,1 e 2, onde $Y = 0$ é a referência. Portanto teria-se $ln(\pi_1/\pi_3)$ e $ln(\pi_2/\pi_3)$, ou seja, em termos de modelo

$$g_1(\mathbf{x}) = ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p$$

$$g_2(\mathbf{x}) = ln\left[\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p$$

assim para $J = 3$ temos $J - 1 = 2$ logitos. Quando denotamos $J = 2$ o modelo seria $ln(\pi_1/\pi_2) = ln(\pi_1/(1-\pi_1))$, que é a regressão dicotômica já estudada.

Um exemplo prático para regressão logística politômica seria o estudo para saber a relação entre a área de formação(exatas, humanas e biológicas) e algumas características, como idade, sexo e raça de estudantes universitários. Portanto a idéia é usar um modelo de regressão logístico politômico, por causa

do caráter categórico da variável resposta e o fato de ela ter mais de duas categorias. Considerando que as categorias foram codificadas como sendo 0 = biológicas, 1 = exatas, 2 = humanas e utilizamos a categoria "biológicas"como sendo a referência o modelo para esse estudo seria da forma

$$g_1(\mathbf{x}) = ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{10} + \beta_{11}\text{idade} + \beta_{12}\text{sexo} + \beta_{13}\text{raça}$$

$$g_2(\mathbf{x}) = ln\left[\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{20} + \beta_{21}\text{idade} + \beta_{22}\text{sexo} + \beta_{23}\text{raça}$$

Assim o interesse é comparar a categoria "exatas"com a categoria de referência(biológicas) e também a categoria "humanas"com a categoria de referência.

Pode-se também encontrar logitos para os outros pares das categorias da variável resposta. Considere, por exemplo, um par de categorias arbitrárias a e b,

$$\begin{aligned} ln\left(\frac{\pi_a}{\pi_b}\right) = ln\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) &= ln\left(\frac{\pi_a}{\pi_J}\right) - ln\left(\frac{\pi_b}{\pi_J}\right) \\ &= (\beta_{a0} + \beta_{a1}x) - (\beta_{b0} + \beta_{b1}x) \\ &= (\beta_{a0} - \beta_{b0}) + (\beta_{a1} - \beta_{b1})x \end{aligned}$$

Assim, o logito para as categorias a e b tem intercepto $(\beta_{a0} - \beta_{b0})$ e inclinação $(\beta_{a1} - \beta_{b1})$.

As probabilidades estimadas da resposta no caso politômico são dadas pela expressão abaixo

$$\pi_j = P(Y=j|x) = \frac{e^{g_j(\mathbf{x})}}{\sum_{k=0}^{j} e^{g_k(\mathbf{x})}}$$

sendo $g_0(\mathbf{x}) = 0$ para a categoria de referência.

Para o exemplo acima as probabilidades são

$$\pi_0 = \frac{1}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}}$$

$$\pi_1 = \frac{e^{g_1(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}}$$

$$\pi_2 = \frac{e^{g_2(\mathbf{x})}}{1 + e^{g_1(\mathbf{x})} + e^{g_2(\mathbf{x})}}$$

para $j = 0, 1, 2$.

3.2.2 Interpretação do modelo

Após introduzir a regressão logística politômica, o próximo passo agora é a interpretação dos coeficientes, o qual constitui uma das fases mais importantes da análise estatística, pois é a partir dela que conclusões serão feitas para o estudo em questão. A razão de chances, já mencionada anteriormente, é uma ferramenta de extrema importância nesse passo, visto que ela traz a informação necessária para a análise das variáveis. Para mostrar de forma simples a interpretação do modelo será utilizado o exemplo descrito no início dessa seção, onde tem-se uma variável resposta com 3 categorias(exatas, humanas e biológicas) e as variáveis explicativas idade, sexo e raça. Algumas informações são descritas abaixo

Tabela 3.2: Informações das variáveis

Variável	Código
Área de formação	0=Biológicas 1=Exatas 2=Humanas
Sexo	0=Masculino 1=Feminino
Idade	17-40
Raça	0=Outras 1=Negro 2=Branco

Pode-se observar que há uma variável explicativa dicotômica(sexo), uma politômica(raça) e uma quantitativa(idade). Usando r para representar a variável dummy referente a variável raça, na qual a categoria escohida como referência foi "outras", o modelo para esse estudo seria da forma

$$g_1(\mathbf{x}) = ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right] = \beta_{10} + \beta_{11}\text{idade} + \beta_{12}\text{sexo} + \beta_{13}r_1 + \beta_{14}r_2$$

$$g_2(\mathbf{x}) = ln\left[\frac{P(Y=2|x)}{P(Y=0|x)}\right] = \beta_{20} + \beta_{21}\text{idade} + \beta_{22}\text{sexo} + \beta_{23}r_1 + \beta_{24}r_2$$

Suponha que ao analisar esse banco de dados em um software de estatística utilizando regressão logística, os resultados obtidos sejam

Tabela 3.3: Resultados

Logito	Variável	Coeficiente estimado	Razão de chances
1	Idade	0,123	1,13
	Sexo	1.234	3,4349
	$r(1)$	0,017	1,017
	$r(2)$	0.904	2,469
2	Idade	0.370	1,48
	Sexo	2.237	9,365
	$r(1)$	0.555	1,742
	$r(2)$	0.780	2,18

A tabela acima possiu apenas as informações das variáveis, dos coeficientes e da razão de chance, pois o objetivo é apenas mostrar a interpretação dos coeficientes do modelo e não realizar um estudo completo, como o qual será realizado mais adiante neste trabalho. Em um estudo real, as informações dos resultados seriam mais completos, contendo também estimativas de erro padrão e testes de hipóteses que irão ajudar na estratégia de , por exemplo, seleção do modelo que será aboradada na próxima seção. Voltando aos resultados encontrados, para o primeiro logito observa-se que para pessoas do sexo feminino a chance estimada de escolha de Exatas comparada à Biológicas é 1,13 vezes a chance do sexo masculino. Considerando agora a variável raça, as conclusões seriam que para negros a chance estimada de escolher Exatas comparada à Biológicas é 1,017 vezes a chance estimada para a categoria outras. Para brancos a chance de escolher Exatas comparada a Biológicas é 2,469 vezes a chance da categoria outras. Para a variável idade, percebe-se que o coeficiente estimado tem sinal positivo, indicando então que para pessoas mais velhas há um aumento da chance de escolha da categoria Exatas e também Humanas em relação a Biológicas.

As interpretações para o segundo logito são similares à do primeiro, mas agora tomando o cuidado de observar que no segundo logito estamos comparando Humanas com Biológicas e não mais Exatas com Biológicas.

3.3 Seleção do modelo

Em análises estatísticas envolvendo regressão, em geral,não é utilizado todo o conjunto de variáveis explicativas do estudo, pois além de ser mais complexo trabalhar com muitas informações, algumas delas não trazem uma contribuição importante para a análise. Nesta seção serão apresentadas algumas estratégias que ajudam a escolher as variáveis que mais trazem contribuição para o modelo e a eliminar aquelas que são irrelevantes do ponto de vista estatístico, ou seja, o objetivo é encontrar o "melhor"modelo que explique os dados usando técnicas que ajudem a minimizar a quantidade de variáveis sem perda de informações importantes. A seleção de variáveis é um passo necessário , pois a falta dessa etapa pode produzir altas estimativas de erro padrão e uma certa dependência do modelo em relação aos dados utilizados.

Em praticamente todo estudo estatístico o passo inicial das análises consiste em um estudo do comportamento das variáveis, ou seja, uma análise exploratória dos dados, que pode ser feita no âmbito univariado e posteriormente bivariado, para se verificar a relação de cada uma com a variável resposta.

Uma forma de selecionar as variáveis que serão candidatas ao modelo de regressão é realizar uma regressão logística simples para cada variável do modelo e observar o p-valor do teste de hipótese realizado. Aquelas que apresentarem um p-valor menor que 0.25 podem ser "candidatas"a entrar no modelo que inclua as variáveis selecionadas. O uso de um nível de significância de 0.25 e não o usual 0.05 se dá pelo fato de que ao usar o nível tradicional poderia-se não escolher adequadamente as variáveis, pois uma variável pode ser não significativa sozinha, mas quando incorporada ao modelo junto com outra se torna um importante preditor para o modelo. Após essa primeira triagem pega-se todas as variáveis candidatas que foram escolhidas na etapa anterior e executa-se o modelo de regressão logística politômica múltipla. O próximo passo é verificar qual das variáveis nesse modelo múltiplo não é significativa, agora já

considerando um nível de significância de 5%. Executa-se o modelo sem a variável não significativa, o qual será chamado de modelo reduzido. Utiliza-se o teste da razão de verossimilhança, já mencionado anteriormente, para comparar o modelo completo(com todas as variáveis candidatas) com o modelo reduzido(não possui a variável não significativa). Assim, se o p-valor for maior que 0.05, não rejeitamos a hipótese nula do teste e portanto o modelo reduzido se ajusta tão bem quanto o completo, sendo então o modelo reduzido preferível ao completo. Então realiza-se o mesmo procedimento para o modelo que foi escolhido verificando qual variável não é significativa, retirando-a do modelo e comparando novamente os modelos até se achar o modelo que melhor explica os dados.

Existem também meios de se fazer essa seleção de variáveis de forma automática, utilizando métodos iterativos que estão implementados nos softwares de estatística. Os métodos ***backward***, ***forward*** e ***stepwise*** são utilizados para esse fim. O método *backward* começa com todas as variáveis explicativas no modelo, calcula-se a deviance parcial e o respectivo p-valor para cada variável no modelo, se algum p-valor for maior que o nível de significância, a variável com maior p-valor é removida do modelo e volta-se novamente para o passo do cálculo da deviance, se o p-valor for menor que o nível se significância , então tem-se o modelo final. O *forward* começa de forma contrária ao método anterior, sem nenhuma variável explicativa no modelo, calcula-se a deviance e p-valor para cada variável não incluída no modelo, se o p-valor for menor que o nível de significância a variável com menor p-valor entra no modelo e volta-se para o passo do cálculo da deviance, se o p-valor for maior que o nível de siginificância então chegou-se no modelo final. Para esses dois métodos mostrados a seleção se dá praticamente em uma única direção, adicionando variáveis ou excluindo variáveis durante todo o processo. O último método a ser abordado é o *stepwise*, que consiste em uma combinação dos outros dois métodos, ou seja, pode-se adicionar ou excluir variáveis em um mesmo processo. Começa

sem nenhuma variável no modelo, calcula-se a deviance e o p-valor para cada variável explicativa, se houver algum p-valor maior que o nível de significância, a variável com maior p-valor é retirada do modelo e volta-se para o passo anterior, se o p-valor for menor que o nível de segnificância, calcula-se a deviance e o p-valor para cada variável não incluida no modelo, verifica-se se há algum p-valor menor que o nível se significância, se sim a variável explicativa com menor p-valor entra no modelo, se não tem-se o modelo final.

Os critérios AIC e BIC, já mencionados anteriormente também podem ajudar na seleção de modelo, sendo que valores menores para esses critérios são preferíveis.

3.4 Qualidade do ajuste

Após a seleção do modelo, a próxima etapa é verificar a "qualidade"do modelo escolhido. Algumas medidas que ajudam nesse processo são o teste qui-quadrado de Pearson, teste deviance e teste de Hosmer-Lemeshow, sendo os dois primeiros usados para dados onde há valores repetidos de observações e o último é usado geralmente quando há poucos valores repetidos, como por exemplo, no caso de variáveis contínuas.

Pearson X^2

O teste de qui-quadrado de Pearson para qualidade de ajuste do modelo tem como hipóteses

$$h_0 : \text{O modelo ajusta bem}$$

$$h_1 : \text{O modelo não ajusta bem}$$

e sua estatística é dada por

$$X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

onde c é número de grupos, O_{jk} são os valores observados e E_{jk} são os valores esperados. Se o modelo for apropriado X^2 se aproxima da distribuição χ^2 com $c - p$ graus de liberdade.

Teste Deviance

A Deviance pode ser utilizada também como medida de qualidade de ajuste, utilizando também a comparação entre valores observados e esperados

$$G^2 = \sum \sum O_{jk} ln \frac{O_{jk}}{E_{jk}}$$

sob as mesmas hipóteses do teste anterior.

Teste de Hosmer-Lemeshow

No caso de variáveis explicativas contínuas os teste relatados acima não são adequados, visto que não se teria uma distribuição aproximadamente qui-quadrado. Assim Hosmer e Lemeshow propuseram o agrupamento de dados em classes com valores parecidos de $\pi(\hat{x}_i)$ com aproximadamente o mesmo número de casos em cada classe. Depois do agrupamento utiliza-se o teste de qui-quadrado de Pearson com $c - 2$ graus de liberdade.

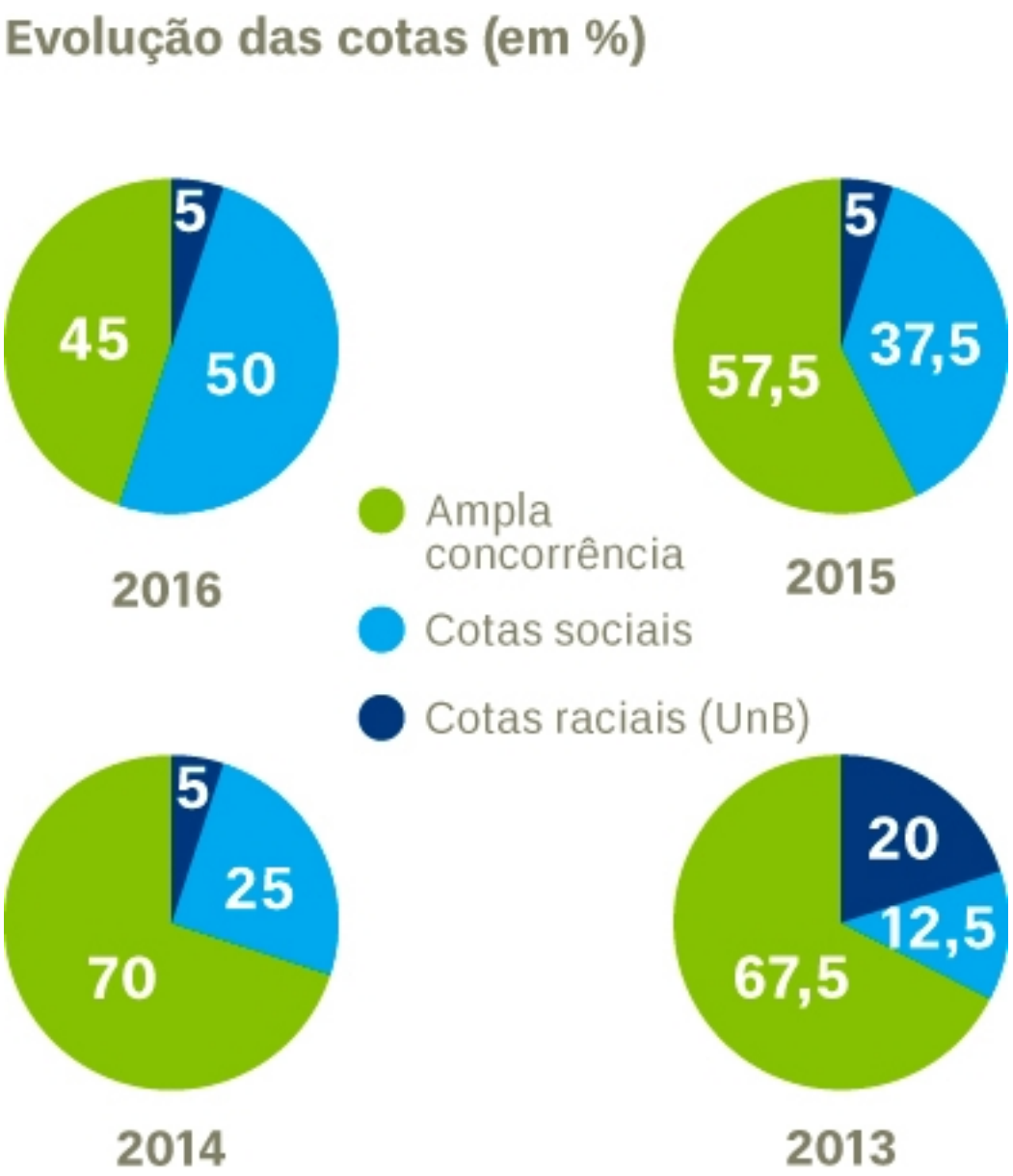
Capítulo 4

Aplicação

4.1 Introdução

A Universidade de Brasília(UnB) está distribuída em 4 campus localizados nas regiões de Ceilândia, Planaltina, Gama e Asa Norte. Oferece por volta de 70 cursos de graduação, entre licenciaturas ou bacharelados, sendo a maioria no turno diurno e alguns também no turno noturno. O sistema de entrada de alunos na UnB se dá em duas oportunidades, sendo uma no primeiro semestre do ano e a outra no segundo semestre do ano através de sistemas de seleção. Essa seleção pode ser feita das seguintes maneiras: vestibular, ENEM(Sisu), seleção para vagas remanescentes, admissão para portador de diploma de curso superior, transferência facultativa, transferência obrigatória e ainda o PAS(somente para ingressantes no primeiro semestre do ano). O ENEM só passou a ser utilizado de fato como porta de entrada principal a partir do primeiro semestre de 2014, antes era utilizado apenas para vagas remanescentes.

A Universidade de Brasília possui ainda alguns sistemas de concorrência baseados em cotas raciais e sociais. Os gráficos abaixo, retirados do site da UnB mostra como se dá essa evolução



Ou seja, no ano de 2013 a grande maioria das vagas eram para o sistema universal e a expectativa é que em 2016 metade das vagas sejam destinadas para alunos da rede pública de ensino. Sendo que dessa metade, 50% seriam destinadas a alunos com renda inferior a 1,5 salário mínimo e a outra metade a alunos com rendas superior a 1,5 sálario mínimo levando em consideração também a raça. O quadro abaixo explica melhor essa composição

Por causa dessa diversidade de alunos e campus na UnB seria interessante traçar um perfil dos alunos que entram e verificar se determinadas características estão atribuídas aos Campus e assim realizar uma classificação.

O Observatório da Vida Estudantil, com o intuito de estudar informações socioeconômicas e demográficas dos alunos ingressantes na UnB, promoveu um questionário no qual os alunos responderam perguntas relativas a ingresso, perfil socioeconômico e demográfico, trajetória pré-universitária e inserção universitária. As respostas foram entregues no momento do registro na UnB, sendo a base de dados gerenciada pelo LimeSurve. As bases de dados do observatório são constituídas por 7 períodos, sendo do primeiro semestre de 2012 ao primeiro semestre de 2015. As bases possuem informações de todos os alunos ingressantes, ou seja, por volta de 3500 a 4000 alunos por período registrados em todos os cursos. Por esse motivo os dados se tornam muito heterogêneos, então, nesse trabalho será considerado apenas os alunos da área de saúde distribuídos em três faculdades

- **Faculdade de Ceilândia**

Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Gestão em Saúde e Terapia Ocupacional.

- **Faculdade de Saúde(Campus Darcy Ribeiro)**

  Ciências Farmacêuticas, Enfermagem, Gestão em Saúde Coletiva, Nutrição e Odontologia.

- **Faculdade de Medicina(Campus Darcy Ribeiro)**

  Medicina

Para o estudo foram considerados 7 bases de dados referentes aos semestres de 1/2012, 2/2012, 1/2013, 2/2013, 1/2014, 2/2014 e 1/2015 dos alunos ingressantes. A ideia inicial é realizar uma análise descritiva para uma comparação entre esses períodos e posteriormente utilizar a modelagem para traçar o perfil dos alunos nessas três faculdades.

4.2 Análise Descritiva

4.2.1 Análise Geral

Com o intuito de apresentar as características gerais dos alunos da área de saúde, foi realizada uma análise descritiva inicial. Para uma melhor compreensão, os resultados foram separados em tabelas, as quais mostram as características de ingresso, perfil socioeconômico e demográfico, trajetória pré-universitária e inserção universitária.

Tabela 4.1: Ingresso

Variáveis		Período						
		1/2012	**2/2012**	**1/2013**	**2/2013**	**1/2014**	**2/2014**	**1/2015**
Sistema de Ingresso	Universal	94,9%	84,3%	80,9%	79,4%	61,0%	73,1%	49,7%
	Cotas Raciais	5,1%	15,7%	19,1%	11,6%	4,5%	6,0%	6,8%
	Cotas Sociais ou Raciais	-	-	-	4,4%	16,4%	2,7%	25,2%
	Cotas para escolas públicas	-	-	-	4,6%	18,1%	18,1%	18,2%
Modalidade de Ingresso	Vestibular	37,2%	58,3%	48,1%	93,1%	0,6%	94,2%	-
	ENEM(Sisu)	6,8%	18,8%	7,4%	0,8%	30,8%	1,0%	48,1%
	PAS	48,5%	-	30,8%	-	48,5%	-	49,7%
	Outros	7,4%	22,9%	13,7%	6,1%	20,1%	4,8%	2,2%
Campus e Turno	Darcy Ribeiro (Diurno)	38,5%	37,8%	37,9%	37,0%	34,9%	34,8%	36,8%
	Darcy Ribeiro (Noturno)	13,0%	13,3%	14,0%	11,2%	9,2%	12,3%	10,8%
	Ceilândia	48,5%	48,9%	48,1%	52,8%	55,9%	52,9%	52,4%
Curso	**Ceilândia**							
	Gestão de Saúde	11,2%	12,4%	10,9%	7,1%	10,1%	8,4%	10,8%
	Enfermagem	9,0%	7,4%	9,2%	10,8%	10,1%	9,5%	9,6%
	Fisioterapia	10,1%	10,0%	12,2%	11,2%	10,9%	8,9%	9,0%
	Farmácia	8,8%	8,9%	7,9%	10,0%	10,3%	10,1%	16,2%
	Terapia Ocupacional	9,2%	10,0%	7,9%	6,5%	8,6%	9,3%	7,4%
	Fonoaudiologia	-	-	-	6,2%	5,8%	6,6%	6,3%
	Saúde							
	Saúde Coletiva	7,3%	7,9%	8,9%	4,8%	5,5%	6,4%	6,1%
	Enfermagem	9,5%	8,5%	8,1%	7,7%	6,2%	8,0%	7,4%
	Odontologia	6,8%	4,8%	6,4%	6,7%	5,6%	5,6%	6,8%
	Farmácia	13,2%	12,7%	13,9%	13,7%	11,9%	13,0%	4,9%
	Nutrição	6,4%	7,2%	5,8%	6,5%	6,4%	5,4%	6,3%
	Medicina							
	Medicina	8,4%	10,0%	8,6%	8,7%	8,4%	8,7%	9,2%

Para a variável sistema de ingresso, nota-se que há predominância de entrada pelo sistema universal, com exceção de 1/2015. Com o passar dos períodos o percentual de ingressantes por cotas aumenta, provavelmente reflexo das novas políticas de cotas que aumentaram o número de vagas para esse sistema. Considerando a modalidade de ingresso, observa-se que no período de 1/2012 à 2/2013 o vestibular tinha grande expressão na entrada dos alunos e o ENEM pouca, mas em 1/2014 e 1/2015 o quadro se reverte. Em 2/2014 94,2% dos alunos entraram pelo vestibular, mostrando que para o segundo semestre do ano ele ainda está sendo utilizado. O percentual dos cursos nas três faculdades parece não mudar no decorrer dos períodos, sendo o curso de farmácia da faculdade de saúde o que tem a porcentagem um pouco mais elevada, fato que ocorre provavelmente por esse curso ser ofertado no turno

diurno e noturno.

Tabela 4.2: Perfil Socioeconômico e Demográfico

Variáveis		Período						
		1/2012	2/2012	1/2013	2/2013	1/2014	2/2014	1/2015
Sexo	Feminino	76,6%	74,8%	71,5%	76,9%	69,2%	72,5%	67,3%
	Masculino	23,4%	25,2%	28,5%	23,1%	30,8%	27,5%	32,7%
Nacionalidade	Brasileiro(a)	100%	99,8%	99,5%	99,6%	99,8%	99,8%	99,6%
	Estrangeiro(a)	-	0,2%	0,5%	0,4%	0,2%	0,2%	0,4%
Estado Civil	Solteiro(a)	96,5%	94,0%	94,1%	95,2%	93,1%	94,8%	94,7%
	Outros	3,5%	6,0%	5,9%	4,8%	6,9%	5,2%	5,3%
Cor/Raça	Branca	46,6%	37,9%	45,9%	44,3%	43,4%	45,0%	37,4%
	Preta ou Parda	49,4%	55,8%	50,3%	51,3%	52,8%	51,9%	58,7%
	Amarela ou Indígena	4,0%	6,31%	3,8%	4,4%	3,8%	3,2%	3,8%
UF de Residência	DF	89,1%	85,2%	93,1%	91,4%	90,4%	92,2%	89,5%
	Outros	10,9%	17,8%	6,9%	8,6%	9,6%	7,8%	10,5%
RA de Residência(1)	DF Alta Renda	15,7%	16,2%	17,8%	16,8%	15,7%	17,3%	12,2%
	DF Média Renda	46,1%	43,3%	46,1%	45,2%	43,9%	45,9%	42,4%
	DF Baixa Renda	25,9%	23,2%	24,9%	28,2%	28,6%	26,0%	33,7%
	GO Entorno	6,3%	6,0%	4,6%	4,3%	5,5%	4,3%	5,8%
	Outros	6,0%	11,1%	6,6%	5,6%	6,4%	6,5%	6,0%
Com quem Reside	Família	89,8%	86,8%	92,4%	87,8%	86,4%	87,8%	89,9%
	Outros	10,2%	13,2%	7,6%	12,2%	13,6%	12,2%	10,1%
Transporte Próprio	Sim	17,2%	18,1%	26,2%	19,8%	18,2%	23,0%	22,0%
	Não	82,8%	81,9%	73,8%	80,2%	81,8%	77,0%	78,0%
Transporte Público	Sim	78,2%	69,0%	75,3%	78,0%	77,0%	76,8%	59,9%
	Não	21,8%	31,0%	24,7%	22,0%	23,0%	23,2%	40,1%
Outros Transportes	Sim	18,3%	15,7%	21,9%	19,5%	23,2%	17,8%	26,1%
	Não	81,7%	84,3%	78,1%	80,5%	76,8%	82,2%	73,9%
Renda Familiar	Até 3 s.m.	25,7%	36,1%	28,8%	28,1%	27,6%	28,6%	28,0%
	3 até 10 s.m.	47,5%	41,2%	40,3%	42,2%	38,3%	37,2%	40,9%
	10 até 20 s.m.	16,5%	11,9%	20,0%	19,5%	19,5%	20,1%	19,8%
	mais de 20 s.m.	10,3%	10,8%	11,0%	10,2%	14,6%	14,1%	11,2%
Recebe Benefício Social	Sim	5,9%	5,8%	5,3%	5,8%	6,3%	4,2%	6,1%
	Não	94,1%	94,2%	94,7%	94,2%	93,7%	95,8%	93,9%
Vivem da renda do domicílio	Até 2 pessoas	12,2%	17,4%	15,0%	14,8%	14,0%	17,5%	16,0%
	3 ou 4 pessoas	56,2%	53,5%	59,0%	58,4%	56,3%	52,7%	57,0%
	Mais de 4 pessoas	31,6%	29,1%	26,0%	26,8%	29,7%	29,8%	27,0%
Escolaridade do Pai	Ens. Fundamental	18,1%	25,3%	23,3%	22,4%	23,6%	20,3%	23,4%
	Ens. Médio	37,9%	35,4%	29,3%	34,8%	31,9%	37,9%	33,8%
	Ens. Superior ou Pós	44,0%	39,3%	47,4%	42,9%	44,5%	41,8%	42,8%
Escolaridade da Mãe	Ensino Fundamental	14,8%	19,5%	14,1%	5,3%	15,4%	14,8%	15,1%
	Ensino Médio	34,8%	39,4%	35,1%	40,8%	35,1%	39,9%	35,3%
	Ensino Superior ou Pós	50,3%	41,1%	49,8%	53,9%	49,6%	45,3%	49,6%
Convênio ou Plano de Saúde	Sim	57,2%	53,6%	59,8%	55,0%	58,5%	61,0%	53,6%
	Não	42,8%	46,4%	40,2%	45,0%	41,5%	39,0%	46,4%

Notas: (1) RA de Residência: Foram consideradas como de Alta Renda as RA´s Park Way, Lago Sul e Norte, Sudoeste, Octogonal e Brasília. Como Média Renda foram consideradas Sobradinho, Taguatinga, Águas Claras, Cruzeiro, Guará, Gama, Candangolândia, Núcleo Bandeirante, São Sebastião, Riacho Fundo I e Vicente Pires. Ceilândia, Recanto das Emas, Samambaia, Riacho Fundo II, Paranoá, Planaltina, Brazlândia, Santa Maria, Valparaíso I e II foram considerados Baixa Renda. O Entorno foi caracterizado por Cocalzinho de Goiás, Alexânia, Águas Lindas, Cidade Ocidental, Cristalina, Formosa, Luziânia, Novo Gama, Planaltina de Goiás, Padre Bernardo, Santo Antônio do Descoberto e Valparaíso de Goiás. Localidades em outros estados foram considerados na categoria "outros".

A maioria dos alunos que ingressaram em cursos da área de saúde são do sexo feminino, solteiros, praticamente metade são brancos e a outra metade pardos ou pretos, residentes no DF com a família, sendo aproximadamente metade residentes em regiões administrativas de média renda. Analisando o transporte, grande parcela utiliza o público(ônibus ou metrô) e a minoria utiliza transportes próprios(carro ou moto) e outros transportes(bicicleta, carona e a pé).

A renda familiar dos alunos é mais expressiva até 10 salários mínimos, a maioria não recebe benefício social e vivem da renda do domicílio de 3 a 4 pessoas. O grau de escolaridade do pai e da mãe se constituem principalmente em nível médio e nível superior e pós, sendo que as categorias nível médio e nível fundamental agregam os níveis completo e incompleto. Em relação a plano de saúde, um pouco mais da metade declarou possuí-lo.

É possível verificar que não há muita diferença entre os percentuais das variáveis em relação aos períodos, mostrando que o perfil socioeconômico e demográfico dos alunos se manteve o mesmo com o passar dos anos.

Tabela 4.3: Trajetória Pré-Universitária

Variáveis		**Período**						
		1/2012	**2/2012**	**1/2013**	**2/2013**	**1/2014**	**2/2014**	**1/2015**
Ensino Médio	Pública	$38,9\%$	$47,3\%$	$39,9\%$	$43,0\%$	$46,0\%$	$42,3\%$	$54,5\%$
	Particular	$61,1\%$	$52,7\%$	$60,1\%$	$57,0\%$	$54,0\%$	$57,7\%$	$45,5\%$
Tipo de Ensino Médio	Regular	$95,0\%$	$85,0\%$	$91,1\%$	$80,0\%$	$90,1\%$	$81,3\%$	$91,1\%$
	Outros	$5,0\%$	$15,0\%$	$8,9\%$	$20,0\%$	$9,9\%$	$18,7\%$	$8,9\%$
Curso Preparatório	Sim	$47,7\%$	$54,2\%$	$42,2\%$	$49,1\%$	$42,6\%$	$47,8\%$	$35,0\%$
	Não	$52,3\%$	$45,8\%$	$57,8\%$	$50,9\%$	$57,4\%$	$52,2\%$	$65,0\%$

Mais da metade dos alunos declararam ter feito ensino médio em escola particular nos períodos de 1/2012 a 2/2014, porém em 1/2015 mais da metade declarou ter feito em escola pública. Como já relatado acima provavelmente esse fato decorre do aumento de vagas para cotas raciais e sociais. A grande maioria realizou o ensino médio regular, enquanto a minoria declarou ter rea-

lizado educação de jovens e adultos(EJA), supletivo, telecurso, etc, as quais foram agrupadas na categoria outros. Nos períodos a porcentagem de alunos que realizaram curso preparatório e os que não realizaram são bem semelhantes, sendo no 1/2015 a diferença um pouco maior, 65% para aqueles que não realizaram.

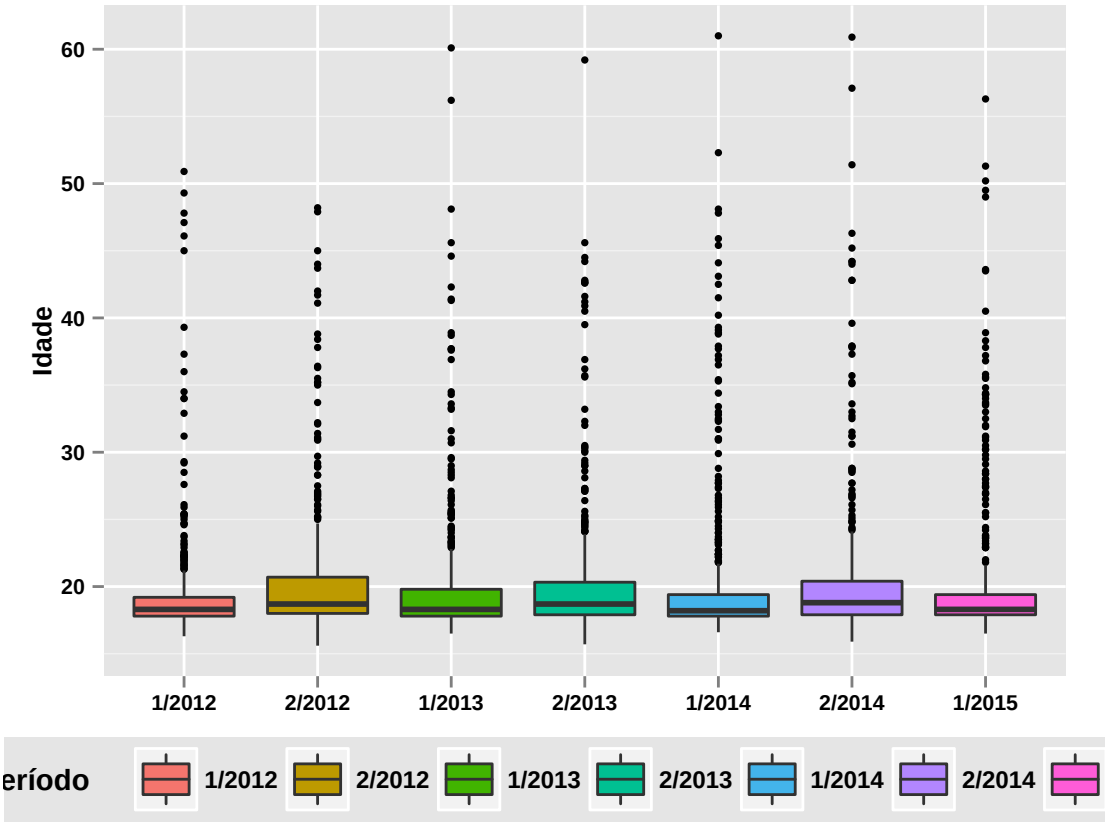
Tabela 4.4: Inserção Universitária

Variáveis		Período						
		1/2012	**2/2012**	**1/2013**	**2/2013**	**1/2014**	**2/2014**	**1/2015**
Curso atual é a 1ª escolha	Sim	57,9%	44,8%	54,5%	53,0%	47,7%	57,3%	50,4%
	Não	42,1%	55,2%	45,5%	47,0%	52,3%	42,7%	49,6%
Se pudesse trocaria de curso	Sim	47,8%	46,2%	47,3%	40,5%	53,0%	39,3%	54,7%
	Não	52,2%	53,8%	52,7%	59,5%	47,0%	60,7%	45,3%
Tentativas de Ingresso na UnB	Esta é a primeira	-	-	31,3%	28,4%	39,0%	28,9%	40,0%
	Uma	-	-	19,1%	15,5%	21,4%	14,0%	24,6%
	Duas	-	-	21,6%	25,0%	20,5%	24,5%	18,8%
	Mais de duas	-	-	28,0%	31,0%	19,2%	32,6%	16,6%

Há uma alternância entre os períodos no que diz respeito a 1ª escolha de curso. Em 1/2012, 1/2013, 2/2013 e 2/2014 a maior parte declarou fazer o curso de primeira escolha, já em 2/2012 e 1/2014 foi o contrário. Os percentuais são bem parecidos, podendo-se então dizer, que praticamente metade dos alunos não fazem o curso de 1ª escolha. Isso também é observado na segunda variável da tabela, pois alunos que não cursam a 1ª opção trocariam de curso se pudessem. As tentativas de ingresso na UnB não foram computadas no banco de dados para os períodos de 1/2012 e 2/2012. Percebe-se que os valores são muito parecidas em 2/2013 e 2/2014 e que há uma porcentagem alta de alunos que já tentaram mais de uma vez entrar na UnB.

Para uma análise da variável idade foi utizado o gráfico boxplot por período.

Figura 4.1: *Boxplot* da Idade X Periodo



Há presença de muitos outliers, sendo as maiores idades em torno de 60 anos. A mediana é próxima em todos os períodos e o comportamento no geral parece não mudar, mas pode-se observar que para os primeiros semestres a variabilidade é menor, podendo esse efeito ser reflexo do PAS, onde os alunos tendem a entrar mais novos, em torno de 18 anos.

4.2.2 Análise Bivariada

Após um conhecimento geral das variáveis, a ideia agora é verificar se elas têm algum tipo de influência em relação as Faculdades de Ceilândia, Saúde e Medicina. Para isso foi realizada uma análise bivariada das variáveis expli-

cativas com a variável resposta(Faculdade). O teste qui quadrado foi utilizado para verificar se há associação entre essas variáveis. De acordo com Hosmer e Lemeshow um nível de significância bom de ser usado é o de 0,25, pois assim não haverá tanto rigor na hora de decidir se a variável será considerada na modelagem, já que uma variável pode não ser significativa sozinha quando considera-se α=0,05, mas ser significativa quando adicionada a outra variável. Na tabela abaixo foram descritas apenas as variáveis que foram consideradas possíveis de explicar a variável Faculdade.

Tabela 4.5: Análise Bivariada: Ingresso

Variáveis		Ceilândia	Medicina	Saúde	Estatística do Teste	P-valor
Sistema de ingresso	Universal	$52,3\%$	$7,8\%$	$39,9\%$	15,5223	0,0004
	Cotas	$53,0\%$	$11,8\%$	$35,2\%$		
Modalidade de Ingresso	ENEM(Sisu)	$61,9\%$	$5,1\%$	$33,0\%$	76,5837	<0,0001
	PAS	$48,9\%$	$8,5\%$	$42,6\%$		
	Vestibular	$46,5\%$	$10,7\%$	$42,7\%$		
	Outros	$67,8\%$	$3,3\%$	$28,9\%$		
Período	Antes do SISU	$49,4\%$	$8,9\%$	$41,6\%$	15,7173	0,0004
	Após SISU	$56,0\%$	$8,7\%$	$35,2\%$		

A variável Período foi reagrupada em apenas duas categorias, sendo a categoria Antes do SISU contendo os períodos de 1/2012 a 2/2013 e Após SISU os períodos de 1/2014 a 1/2015, já que a UnB só adotou o SISU no primeiro semestre de 2014. Para a tabela acima todos os testes foram significativos, havendo então associação entre as variáveis, portanto elas serão incluidas na modelagem.

Tabela 4.6: Análise Bivariada: Perfil Socioeconômico e Demográfico

Variáveis		Ceilândia	Medicina	Saúde	Estatística do Teste	P-valor
Sexo	Feminino	56,1%	5,3%	38,6%	141,0160	<0,0001
	Masculino	42,8%	18,1%	39,1%		
Estado Civil	Solteiro	52,2%	9,1%	38,7%	5,3817	0,0678
	Outros	56,7%	4,0%	39,3%		
Cor/Raça	Branca	47,3%	9,7%	43,0%	25,7941	<0,0001
	Parda ou Preta	56,2%	8,2%	35,6%		
	Amarela ou Indígena	57,5%	8,2%	34,3%		
RA de residência	DF Alta Renda	14,9%	23,5%	61,6%	660,7354	<0,0001
	DF Média Renda	51,7%	5,2%	43,1%		
	DF Baixa Renda	79,6%	1,9%	18,4%		
	GO Entorno	51,2%	5,9%	42,9%		
	Outros	39,4%	27,1%	33,5%		
Com quem reside	Família	53,0%	8,9%	38,2%	1,7162	0,4240
	Outros	49,4%	8,9%	41,6%		
Transporte Próprio	Sim	27,4%	19,4%	53,2%	256,3430	<0,0001
	Não	58,7%	6,1%	35,2%		
Transporte Público	Sim	60,3%	5,2%	34,5%	285,6488	<0,0001
	Não	31,0%	18,8%	50,2%		
Outros Transportes	Sim	39,7%	14,4%	45,8%	66,8396	<0,0001
	Não	55,5%	7,4%	37,1%		
Renda Familiar	Até 3 s.m.	67,3%	4,4%	28,3%	349,5252	<0,0001
	3 até 10 s.m.	57,2%	5,1%	37,7%		
	10 até 20 s.m.	42,2%	11,8%	46,0%		
	mais de 20 s.m.	20,7%	25,5%	53,8%		
Recebe Benefício Social	Sim	60,4%	5,5%	34,1%	5,8818	0,0528
	Não	52,0%	9,1%	39,0%		
Vivem da renda do domicílio	Até 2 pessoas	52,0%	5,9%	42,1%	11,7843	0,0190
	3 ou 4 pessoas	52,1%	8,7%	39,2%		
	Mais de 4 pessoas	53,6%	10,5%	35,8%		
Escolaridade do Pai	Ens. Fundamental	64,0%	4,1%	31,9%	173,3659	<0,0001
	Ens. Médio	59,8%	5,5%	34,6%		
	Ens. Superior e Pós-Graduação	39,7%	14,5%	45,8%		
Escolaridade da Mãe	Ens. Fundamental	62,3%	2,7%	34,9%	166,0659	<0,0001
	Ens. Médio	62,2%	4,8%	33,0%		
	Ens. Superior e Pós-Graduação	42,0%	13,9%	44,1%		
Convênio ou Plano de Saúde	Sim	44,4%	12,0%	43,6%	113,6628	<0,0001
	Não	62,4%	4,7%	32,9%		

As variáveis Estado Civil e Recebe Benefício Social tiveram p-valor respectivamente 0,0678 e 0,0528. Como será considerado um nível de significância de 25% elas, a princípio, entrarão no modelo, assim como todas as demais que foram significativas. A variável Com Quem Reside teve um p-valor igual a 0,4240, então não será considerada no modelo.

Tabela 4.7: Análise Bivariada: Trajetória Pré-Universitária

Variáveis		Ceilândia	Medicina	Saúde	Estatística do Teste	P-valor
Ensino Médio	Particular	$45,8\%$	$11,30\%$	$42,9\%$	78,6366	<0,0001
	Pública	$60,6\%$	$5,7\%$	$33,7\%$		
Tipo de Ensino Médio	Regular	$52,0\%$	$9,3\%$	$38,7\%$	6,0183	0,0493
	Outros	$55,7\%$	$5,7\%$	$38,6\%$		
Curso Preparatório	Sim	$45,1\%$	$14,1\%$	$40,8\%$	116,7931	<0,0001
	Não	$58,7\%$	$4,4\%$	$36,9\%$		

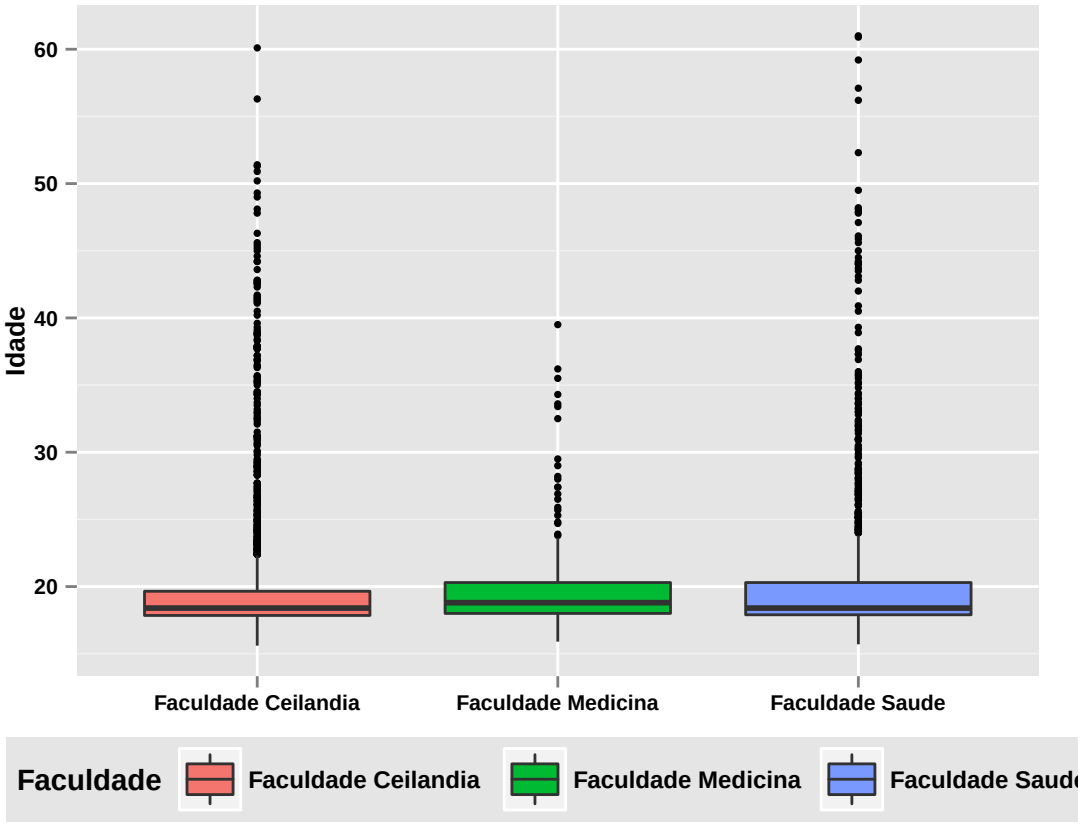
Tabela 4.8: Análise Bivariada: Inserção Universitária

Variáveis		Ceilândia	Medicina	Saúde	Estatística do Teste	P-valor
Curso atual é a 1ª escolha de curso	Sim	$48,3\%$	$14,6\%$	$37,1\%$	144,7143	<0,0001
	Não	$56,9\%$	$2,5\%$	$40,6\%$		
Se pudesse trocaria de curso	Sim	$59,6\%$	$0,4\%$	$40,0\%$	252,3963	<0,0001
	Não	$46,3\%$	$16,2\%$	$37,4\%$		
Tentativas de ingrsso na UnB	Esta é a primeira	$57,4\%$	$7,1\%$	$35,5\%$	37,0122	<0,0001
	Uma	$55,7\%$	$7,6\%$	$36,7\%$		
	Duas	$59,5\%$	$5,1\%$	$35,4\%$		
	Mais de duas	$46,7\%$	$13,4\%$	$39,9\%$		

Todas as variáveis das tabela acimas serão consideradas no modelo inicial, visto que todas foram significativas para um α=0,25.

Para explorar a distribuição da variável idade em relação a variável Faculdade foi realizado o boxplot abaixo

Figura 4.2: *Boxplot* da idade X Faculdade



Os três boxplots são bem parecidos, parece não haver diferença entre as Faculdades com relação a idade. Observa-se que a Faculdade de Medicina tem uma mediana um pouco mais alta em relação as outras duas e possui menos outliers.

Para verificar a associação da variável quantitativa com a variável resposta utilizou-se uma regressão logística politômica simples. O teste realizado para verificar a significância do parâmetro foi o de Wald

Tabela 4.9: Resultado

Parâmetro	Logito	Estimativa do Parâmetro	Erro Padrão	Wald	p-valor
Intercepto	1	-1,5539	0,2844	29,8585	<0,0001
Idade		-0,0119	0,0139	0,7258	0,3943
Intercepto	2	-0,5275	0,1464	12,9856	0,0003
Idade		0,0107	0,00698	2,3453	0,1257

Como o boxplot não indicou nenhuma grande diferença e os p-valores da regressão logística foram altos, essa variável não será considerada no modelo inicial.

4.2.3 Modelagem

Identificada as possíveis variáveis explicativas ajustou-se um modelo de Regressão Logística Politômica devido ao caráter categórico da variável resposta e o fato de ela possuir 3 categorias. O banco utilizado na modelagem foi uma junção dos 7 bancos referidos anteriormente, contituindo 3362 observações, sendo que Período foi usado como uma variável explicativa. Apenas a variável Estado Civil não foi incluída no modelo inicial. A variável Tentativas de Ingresso na UnB acabou sendo retirada do modelo pela alta quantidade de missings(valores faltantes) que ela apresentava. O modelo foi realizado no software de estatítica SAS(Statistical Data Analysis) com a utilização do comando PROC LOGISTIC e usando a opção LINK=GLOGIT para a resposta politômica.

Para a verificação da validação do modelo, dividiu-se o banco de dados de 3362 observações em duas partes de 1681 observações, sendo o primeiro banco chamado de "construção"e o segundo de "validação". A ideia é usar o mesmo modelo para o banco validação e observar se os valores das estimativas são semelhantes.

Em regressão a ideia é sempre tentar encontrar o modelo mais simples que melhor explique os dados, por isso para ajudar na seleção do modelo utilizou-se o método STEPWISE de seleção, assim como também medidas AIC. As

variáveis selecionadas foram Sexo, Transporte Público, 1ª Opção de Curso e Período, então o modelo encontrado é da forma

$$g_1(\mathbf{x}) = \beta_{10} + \beta_{11}\text{Sexo} + \beta_{12}\text{TranspPublico} + \beta_{13}1^{\underline{a}}\text{ opção} + \beta_{14}\text{Período}$$

$$g_2(\mathbf{x}) = \beta_{20} + \beta_{21}\text{Sexo} + \beta_{22}\text{TranspPublico} + \beta_{23}1^{\underline{a}}\text{ opção} + \beta_{24}\text{Período}$$

onde $g_1(\mathbf{x})$ corresponde ao logito 1 e $g_2(\mathbf{x})$ ao logito 2. A categoria de referência escolhida para a variável resposta foi Faculdade de Ceilândia, para Sexo foi a categoria feminino, para as variáveis Transporte Público e O Curso Corresponde a 1ª Opção foi a categoria não. Os testes de qui-quadrado de Pearson e Deviance tiveram p-valores de 0,3413 e 0,3161, respectivamente, indicando que o modelo se ajusta bem.

Os resultados dos parâmteros estimados estão apresentados na tabela abaixo, considerando o banco construção, validação e até mesmo o banco geral.

Tabela 4.10: Resultado da validação

Variável	Logito	Estimativa Construção	Estimativa Validação	Estimativa Modelo Geral
Intercepto	1	-2,32(0,31)	-2,48(0,30)	-2,40(0,21)
Sexo		1,44(0,20)	1,69(0,20)	1,56(0,14)
Transporte Público		-2,29(0,21)	-2,10(0,21)	-2,17(0,15)
Curso é a 1ªopção		2,07(0,27)	1,80(0,26)	1,93 (0,18)
Período		0,29(0,20)	0,35(0,20)	0,32(0,14)
Intercepto	2	0,38(0,15)	0,36(0,14)	0,38(0,10)
Sexo		0,39(0,13)	0,24(0,13)	0,31(0,08)
Transporte Público		-1,29(0,14)	-1,17(0,13)	-1,23(0,09)
Curso é a 1ªopção(*)		0,10(0,11)	0,03(0,10)	0,06(0,07)
Período		0,35(0,11)	0,30(0,11)	0,32(0,08)

As estimativas dos parâmetros são bem parecidos entre os bancos de construção e validação e também para o banco geral. Assim pode-se concluir que o modelo é válido. O * na terceira variável indica que a estimativa não foi

significativa. Os valores em parênteses são os respectivos erros padrão das estimativas.

Em regressão logística o maior interesse é em interpretar a razão de chances estimada pelo modelo. Os valores encontrados para essa medida são mostrados na tabela abaixo

Tabela 4.11: Estimativa da razão de chance

Variável	Logito	Estimativa	Intervalo de Confiança
Sexo(Masc. vs Femin.)	1	4,802	(3,604 ; 6,397)
Transporte Público(Sim vs Não)		0,114	(0,084 ; 0,153)
Curso é a 1ªopção(Sim vs Não)		6,927	(4,785 ; 10,027)
Período(Antes SISU vs Pós SISU)		1,384	(1,042 ; 1,837)
Sexo(Masc. vs Femin.)	2	1,375	(1,153 ; 1,640)
Transporte Público(Sim vs Não)		0,291	(0,241 ; 0,352)
Curso é a 1ªopção(*)(Sim vs Não)		1,072	(0,921 ; 1,249)
Período(Antes SISU vs Pós SISU)		1,390	(1,192 ; 1,622)

Na 3ª variável do logito 2 a presença do * significa que o parâmetro não foi significativo a um nível de 5%, isso pode ser observado no intervalo de confiança, que contém o valor 1. Ná análise bivariada também nota-se isso, visto que aparentemente não há diferença entre as categorias sim e não quando compara-se a Faculdade de Saúde com Ceilândia. As demais variáveis foram significativas. Para facilitar a interpretação será usada as estimativas pontuais da razão de chance.

Para o logito 1 a variável sexo tem uma razão de chances de 4,802, ou seja, para alunos do sexo masculino a chance de ter ingressado na Faculdade de Medicina é 4,802 vezes a chance para alunos do sexo feminino comparada a Faculdade de Ceilândia. Para alunos que usam transporte público a chance de ter ingressado na Faculdade de Medicina é 0,114 vezes a chance para alunos que não utilizam transporte público comparado a Faculdade de Ceilândia. Para os que declararam cursar a primeira opção de curso a chance de pertencer a Medicina é 6,927 vezes a chance para alunos que declararam não cursar a primeira opção de curso comparado a Ceilândia. Para alunos que entraram antes do SISU a chance de pertencer a Medicina é 38% maior que a chance

de alunos que entraram depois do SISU comparado a Ceilândia. Na análise bivariada aparentemente após SISU há mais chance de pertencer a Faculdade de Ceilândia.

Considerando agora o logito 2, para alunos do sexo masculino a chance de ter ingressado na Faculdade de Saúde é 37% maior que a chance de alunos do sexo feminino comparado a Faculdade de Ceilândia, o que também pode ser observado na análise bivariada. Para os que usam transporte público a chance de pertencer a Saúde é 0,291 vezes a chance para alunos que não usam transporte público comparado a Ceilândia. Para os que cursam a primeira opção de curso a chance de pertencer a Saúde é a mesma que a chance para os que não cursam a 1ª opção comparado a Ceilândia. Para os que ingressaram antes do SISU a chance de pertencer a Saúde é 39% maior que a chance para aqueles que ingressaram após o SISU comparado a Ceilândia. Toda análise de cada uma dessas variáveis em cada logito é realizada mantendo as demais constantes.

Os resultados acima foram obtidos através da interpretação dos dois logitos apresentados na saída do SAS, ou seja, as Faculdades de Medicina e Saúde comparadas a Faculdade de Ceilândia, mas também é possível comparar a Faculdade de Medicina com a Faculdade de Saúde, subtraindo a estimativa do parâmetro do primeiro em relação ao segundo. Assim, para alunos do sexo masculino a chance de ter ingressado na Faculdade de Medicina é 3,49 vezes a chance para alunos do sexo feminino comparado a Faculdade de Saúde. Considerando o transporte público, para alunos que usam transporte público a chance de pertencer a Faculdade de Medicina é 0,39 vezes a chance para os que não usam transporte público comparado a Faculdade de Saúde. Considerando agora os que cursam a 1ª opção, a chance de pertencer a Medicina é 6,49 vezes para os que não cursam a 1ª opção comparado a Saúde. Para a variável Período a razão de chances é próxima de 1, indicando que para alunos ingressantes antes do SISU a chance de ter ingressado na Faculdade de Medi-

cina é a mesma de alunos ingressantes após o SISU comparado a Faculdade de Saúde.

Capítulo 5

Conclusão

Os diferentes aspectos que cercam os alunos da UnB fazem com que algumas questões sejam levantadas, como a hipótese de que as características dos alunos se diferem entre as Faculdades de Ceilândia, Medicina e Saúde. Como forma de investigação utilizou-se um modelo de regressão logística politômica, no qual a partir da interpretação de seus parâmetros foi possível enxergar alguns fatores relacionados à variável resposta. No início da modelagem havia um modelo com muitas variáveis explicativas e consequentemente muitos parâmetros, o que nem sempre é bom do ponto de vista estatístico. Após uma seleção chegou-se naquele que melhor explicaria a variável resposta e conclusões puderam ser retiradas dele.

O modelo permitiu mostrar que na faculdade de Medicina o perfil dos alunos, em geral, gira em retorno de pessoas do sexo masculino, que cursam sua primeira opção de curso e não utilizam tanto o transporte público quando comparados a aqueles de Ceilândia. Os alunos da área de saúde tem menor chance de utilizar transporte público, assim como menor chance de sexo feminino quando comparado a Ceilândia, mas parecem ter um perfil um pouco mais parecido com ela no que diz respeito a cursar o curso de 1ª opção, provavelmente pelo fato que alguns alunos que tem a ideia inicial de cursar medicina

acabam optanto por cursar um curso na área de saúde, seja por afinidade com a área ou para aproveitar créditos e que seja menos concorrido do que medicina. Para Ceilândia há uma chance maior de sexo feminino do que em relação às outras faculdades, assim como também a utilização de transporte público, provavelmente em decorrência do fato de a faculdade de Ceilândia estar situada em uma RA considerada de baixa renda.

Dentro da Regressão Logística, a mais encontrada é o caso da resposta com duas categorias, a qual foi mostrada no início do referecial teórico do trabalho. Em geral, os livros abordam com abrangência essa teoria, principalmente no que diz respeito a análise de resíduos. Já no caso da Politômica não há tanta abrangência, se tornando um pouco complicado a realização de alguns procedimentos, principalmente no que diz respeito a programação em softwares de estatística.

Referências Bibliográficas

[1] AGRESTI, Alan. **An Introduction to Categorical Data Analysis**. Second Edition. Hoboken, New Jersey: John Wiley & Sons, 2007.

[2] BITTENCOURT,Hélio. Regressão Logística Politômica: revisão teórica e aplicações.**Acta Scientiae**,Canoas,v.5, n.1,2012.

[3] CODY,Ron.**Learning SAS by Example:A Programmer's Guide**,Cary,NC:SAS Institute Inc.,2007.

[4] HOSMER,David.LEMESHOW,Stanley**Applied Logistic Regression**,United States:John Wiley & Sons,1989.

[5] McCullagh,P.NELDER,J.**Generalized Linear Models**.Second Edition.London:Chapman & Hall,1989.

[6] NETER,John; et al. **Applied Linear Statistical Models**,Fifth Edition.

[7] STOKES,Maura; et al. **Categorical Data Analysis Using SAS**,Third Edition,2012.