Universidade de Brasília
IE – Departamento de Estatística
Estágio Supervisionado 2

EVASÃO NA UNIVERSIDADE DE BRASÍLIA:
UM ESTUDO SOBRE OS CURSOS FORMADORES DE DOCENTES PARA MATÉRIAS BÁSICAS DO VESTIBULAR.

Andreza de Oliveira Lima
Brunno Augusto Cardoso Costa

Relatório Final

Orientadora: Profª Maria Teresa Leão Costa

Brasília
Junho de 2012

Andreza de Oliveira Lima
Brunno Augusto Cardoso Costa

EVASÃO NA UNIVERSIDADE DE BRASÍLIA:
UM ESTUDO SOBRE OS CURSOS FORMADORES DE DOCENTES PARA MATÉRIAS BÁSICAS DO VESTIBULAR.

Trabalho de Conclusão do Curso de Bacharelado em Estatística.

Orientadora: Profª Maria Teresa Leão Costa

Brasília
Junho de 2012

Resumo:

Este trabalho teve como objetivo, primeiramente, delinear o perfil dos alunos da Universidade de Brasília que ingressaram no 2º semestre de 2004 nos cursos de Matemática, Física, Química, Ciências Biológicas, Língua Portuguesa e Respectiva Literatura, História, Geografia e Língua Inglesa e Respectiva Literatura, considerados os cursos formadores de docentes das disciplinas de maior peso dos sistemas de ingresso na universidade. Também, sob a ótica da regressão logística, este trabalho visa principalmente detectar alguns fatores que influenciam na evasão dos alunos destes cursos e a diferença existente entre as habilitações de bacharelado e licenciatura nos mesmos. Para tal, foram utilizados os registros dos alunos no Sistema de Informações Acadêmicas de Graduação (SIGRA). As seguintes variáveis estão relacionadas à evasão dos alunos neste estudo: sexo, habilitação, se o aluno foi monitor ou não, curso, número de ingressos, semestres de permanência, posição no fluxo e percentual de disciplinas com reprovação no período em estudo.

Palavras chave: Regressão Logística, Evasão, Perfil dos Alunos, Licenciatura.

SUMÁRIO

LISTA DE FIGURAS.

LISTA DE GRÁFICOS.

LISTA DE TABELAS

1. Introdução

O Ministério da Educação divulgou recentemente um projeto com o intuito de que os estudantes da educação básica passem mais tempo nas escolas, aumentando a quantidade de dias letivos do calendário escolar de 200 para 220 dias ou o tempo diário de aula, que hoje é de 4 horas por dia. O objetivo desta medida é tentar aproximar a carga horária das escolas do Brasil àquela seguida pelos países mais desenvolvidos. De acordo com o ex-ministro da educação Fernando Haddad "Há um consenso de que no Brasil a criança tem pouca exposição ao conhecimento, seja porque a carga horária diária é baixa ou porque o número de dias letivos é inferior ao dos demais países". Apesar de estar em fase inicial de projeto, essa medida de imediato exige uma atenção especial, pois o já conhecido déficit de profissionais da área de ensino ainda é fonte de preocupações. Além disso, em agosto de 2011, a presidenta Dilma anunciou o novo Plano de Expansão do Ensino Superior, onde a proposta é que sejam investidos cerca de 2,4 bilhões de reais entre 2013 e 2017. Surgem então as seguintes questões: as universidades têm cumprido o seu papel na formação de docentes? O sistema de ensino está preparado para tais planos?

De acordo com o Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), cerca de 123 mil alunos de graduação (cerca de 10%) saíram das universidades públicas em 2006 sem concluir um curso, o que nos leva a crer que apenas a expansão do ensino superior pode não ser suficiente para resolver o problema da falta de professores. É necessário que também sejam criadas políticas para manter os alunos nas universidades e, para tal, precisamos entender como se dá a evasão dos alunos. Segundo a Comissão Especial de Estudos sobre a Evasão (SESu/MEC,1996), este tipo de estudo é uma forma de auto conhecimento da instituição, o qual é imprescindível na orientação de tais políticas para a melhoria do ensino de graduação, evitando assim a tomada de decisões ineficazes pelos gestores da instituição.

Embora não exista um consenso para o significado da palavra evasão, esta pode ser interpretada de modo geral como a interrupção de um curso (de graduação), seja esta interrupção temporária ou não; da instituição, do sistema de educação ou do curso. Existem na literatura diversas causas para este tipo de comportamento, que vão desde fatores sócio-econômicos a fatores individuais, como a vocação profissional, a falta de motivação, etc.

Tendo em vista a formação de profissionais das matérias básicas do vestibular, este estudo visa analisar a evasão sob a perspectiva dos alunos de Bacharelado e Licenciatura da Universidade de Brasília

1.1. Objetivos

1.1.1. Objetivo Geral

Estudo da Evasão em cursos (a serem definidos ao longo deste trabalho) da Universidade de Brasília, os quais compõem a formação dos conteúdos básicos e de maior peso nos processos de seleção de ingresso na universidade, com foco na diferença da evasão nas habilitações de Bacharelado e Licenciatura.

1.1.2. Objetivos Específicos

- Estudo descritivo das características dos alunos selecionados para a análise.
- Identificação e mensuração dos fatores associados à evasão através da técnica de Regressão Logística.
- Através do modelo logístico, tentar comparar a evasão entre as habilitações de Bacharelado e Licenciatura.

1.2. Justificativa

A escolha do tema foi impulsionada pelo fato de que o entendimento dos fatores que levam um aluno a evadir é crucial para o total aproveitamento do que o sistema de ensino superior pode oferecer. Embora seja difícil quantificar, é evidente que a evasão universitária causa prejuízos aos cofres públicos e aos cidadãos, seja com o não aproveitamento efetivo de todas as vagas oferecidas pelas universidades, seja pelos investimentos públicos mal direcionados.

Ingressar na universidade não garante que o aluno obterá sucesso na sua formação. Sendo assim, espera-se que este trabalho ajude a entender quais fatores estão ligados à evasão do aluno e, assim, forneça um embasamento para verificar quais medidas podem ser tomadas para auxiliar a universidade na formação de professores.

2. Educação Básica e Evasão

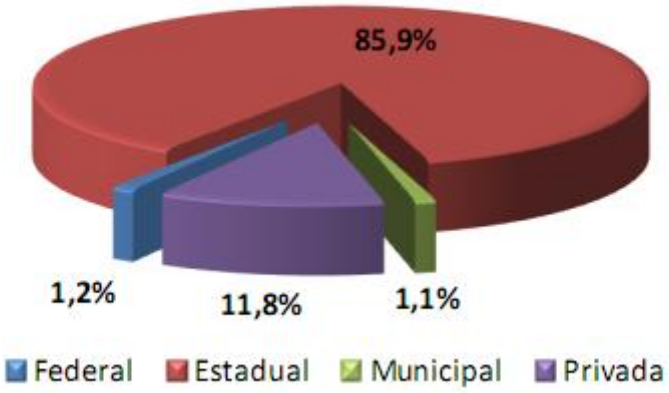
Este capítulo contém informações sobre o panorama da educação e da evasão no Brasil.

2.1. Educação Básica

De acordo com o Censo Escolar da Educação Básica de 2010, realizado pelo Inep, houve um decréscimo de cerca de 2% no número de matrículas em toda a educação básica, que envolve Educação Infantil, Ensino Fundamental, Ensino Médio, Educação de Jovens e Adultos e Educação Profissional. Esse comportamento é devido principalmente ao Ensino Fundamental, onde o histórico de retenção é marcante no Brasil. No Ensino Médio, porém, houve um aumento de 20.515 matrículas de 2009 para 2010, totalizando 8.357.675.

A rede pública estadual é responsável pela maioria dessas matrículas, como pode ser observado no gráfico abaixo:

Gráfico 1 – Matrículas de Educação Básica por Dependência Administrativa, Brasil 2010



Fonte: MEC/Inep. Censo Escolar de 2010.

Em se tratando do número de professores do Ensino Médio, obteve-se a informação, a partir do Censo Escolar da Educação Básica de 2007, que 13% dos 415.555 professores do Ensino Médio não possuem um curso superior com licenciatura.

Observe na tabela 1 abaixo que as áreas com mais profissionais são Letras/Literatura/Língua Portuguesa, Matemática, Pedagogia/Ciências da Educação e Letras/Literatura/Língua Estrangeira e as com menos profissionais são Ciências Sociais/Sociologia, Filosofia e Física. Através desta tabela, é possível observar que as áreas com menos profissionais são as que possuem maior porcentagem de professores sem

licenciatura. Com base nestas informações se obtém uma motivação na escolha dos cursos a serem estudados neste trabalho sobre a evasão.

Tabela 1 – Professores de Nível Superior do Ensino Médio segundo a área de Formação, Brasil 2007

Área de Formação	Total	Porcentagem sem Licenciatura
Letras/Literatura/Língua Portuguesa	67.049	4,11%
Matemática	49.299	5,11%
História	37.999	3,81%
Pedagogia/Ciências da Educação	37.776	6,74%
Letras/Literatura/Língua Estrangeira	32.535	4,74%
Geografia	31.299	4,12%
Ciências Biológicas	28.346	5,11%
Educação Física	27.175	5,29%
Ciências	22.198	4,83%
Química	15.787	7,89%
Belas Artes/Artes Plásticas/Educação Artística	13.793	5,60%
Física	12.355	7,00%
Filosofia	8.535	6,64%
Ciências Sociais/Sociologia	4.896	7,05%
Demais Cursos*	44.976	21,77%

Fonte: MEC/Inep/Deed. Censo da Educação Básica 2007
(1) inclui todos os cursos com proporções inferiores a 1%.

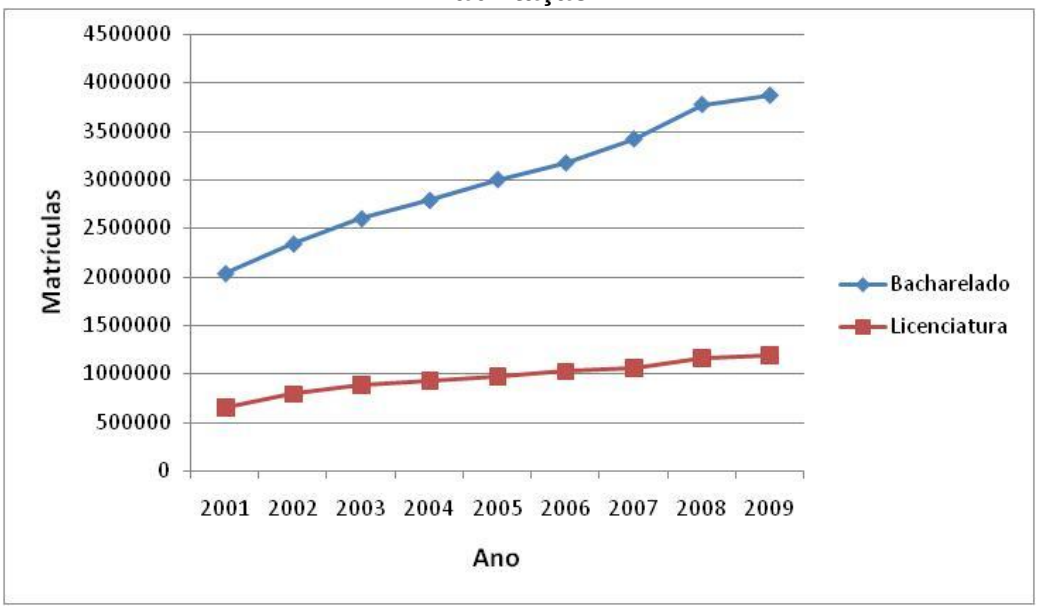
2.2. A Evasão

A evasão é um fenômeno que vem sendo muito estudado nos últimos tempos devido a sua importância para as Instituições de Ensino Superior (IES), sejam elas públicas ou privadas. Apesar disso, por se tratar de um fenômeno muito complexo, torna-se difícil a mensuração e a generalização dos resultados destes estudos. Para começar, o conceito de evasão não está bem definido, o que pode causar resultados completamente diferentes para o mesmo estudo de caso. Em seguida, estudos como o de Almeida e Veloso (2002) apontam que o comportamento da evasão pode variar dependendo do curso, da área do conhecimento e de fatores sócio-econômicos entre outros motivos. Portanto, faz-se necessário para qualquer estudo sobre a evasão, definir bem o conceito a ser utilizado.

No Brasil, estão disponíveis os resultados do Censo da Educação Superior de 2009, fornecidos pelo Inep. Segundo o resumo técnico deste Censo, 71% das matrículas de graduação são de bacharelado, apenas 15% são de Licenciatura e 4% cursam as duas habilitações. No gráfico 2 a seguir, pode-se acompanhar a evolução do número de matrículas nas duas habilitações a partir de 2001 até 2009. É possível observar que o número de matrículas do Bacharelado cresce mais rápido que o número de matrículas da Licenciatura, e essa é uma das motivações deste estudo. Deseja-se descobrir com este

trabalho como é a relação de matrículas entre Bacharelado e Licenciatura na Universidade de Brasília e se a habilitação escolhida pelo estudante influencia na evasão.

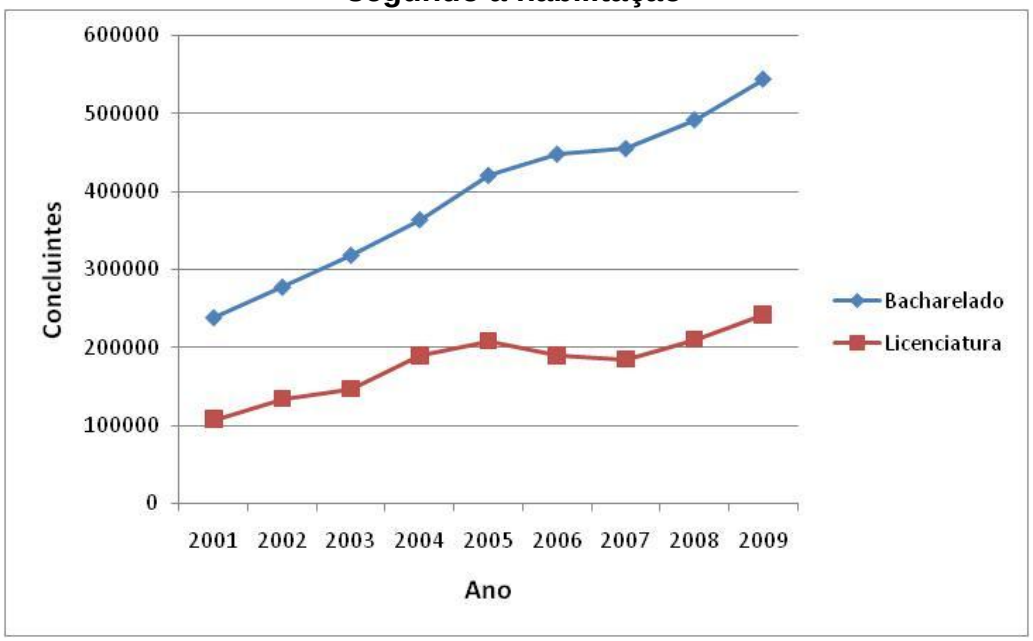
Gráfico 2 – Evolução do número de matrículas no Brasil de 2001 a 2009 segundo a habilitação.



Fonte: MEC/Inep/DEED. Censo da Educação Superior 2009

Já no próximo gráfico, pode-se acompanhar o número de estudantes que concluíram algum curso em cada uma das habilitações, onde se observa novamente que o número de estudantes de bacharelado é maior que o de licenciatura.

Gráfico 3 – Evolução do número de concluintes no Brasil de 2001 a 2009 segundo a habilitação



Fonte:MEC/Inep/DEED. Censo da Educação Superior 2009

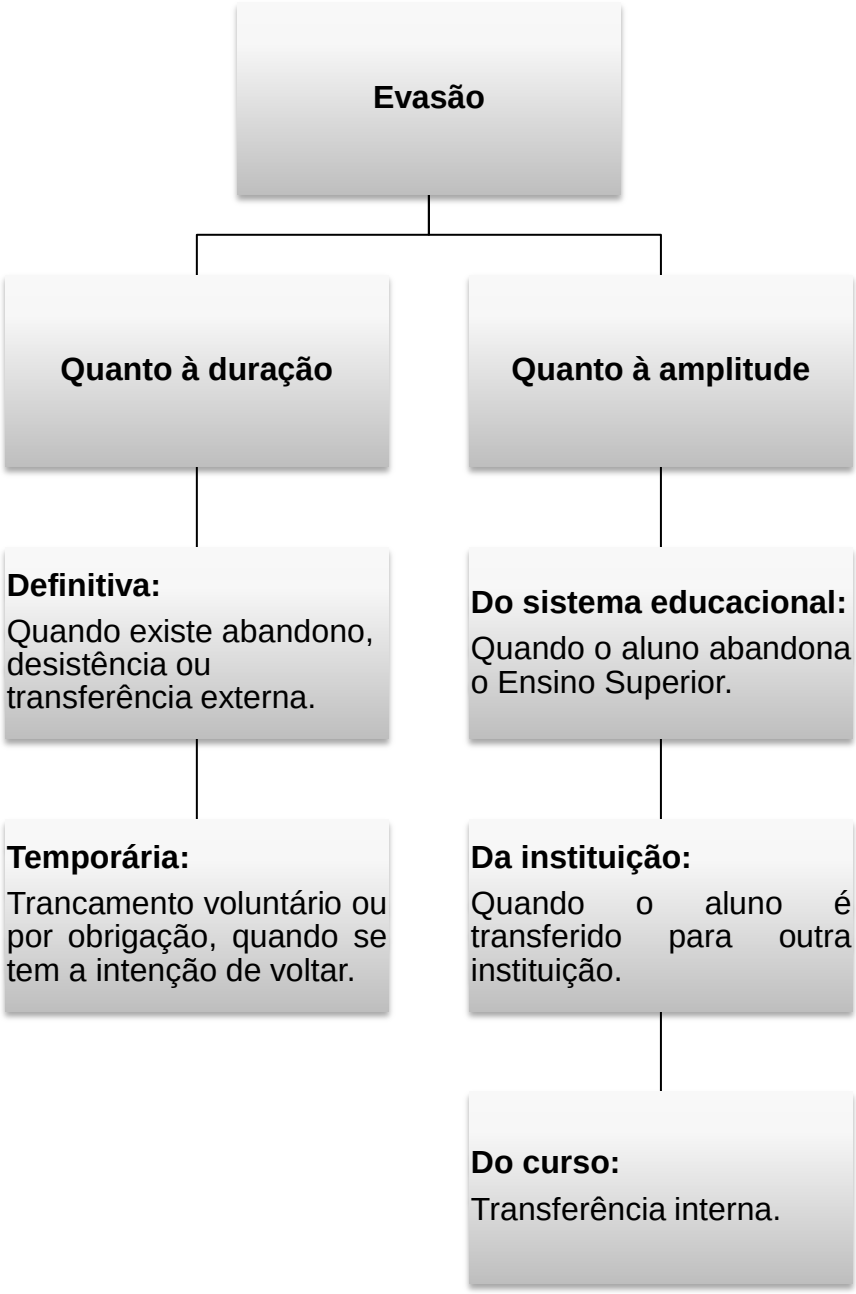
Definidas as coordenadas do estudo, o qual visa diferenciar a evasão nos cursos de bacharelado e licenciatura, é preciso definir o que será tratado por evasão neste trabalho. Como já foi dito, muitos conceitos foram encontrados na literatura sobre o termo evasão. De acordo com Miranda (2006)[1], a evasão pode ser dividida em relação à duração da evasão, se é definitiva ou temporária, e em relação ao que a autora chama de amplitude, que envolve evasão do sistema de ensino superior, da instituição e do curso. Ainda, para a autora estes casos estão interligados. Pode-se, por exemplo, existir evasão de curso temporária ou da instituição definitiva.

Outros estudiosos pregam a divisão entre evasão e exclusão ou evasão e mobilidade. Estes conceitos apresentam diferenças mais sutis, os quais de alguma forma estão englobados na separação de Miranda. Nestes casos, a evasão é considerada como uma decisão do aluno por se desligar da universidade, a exclusão é uma decisão da universidade e a mobilidade se confunde com os conceitos de transferência.

O esquema a seguir define e relaciona os conceitos de evasão:

[1] Referências de Miranda: COSTA, A. L. **Evasão dos cursos de graduação da UFGRS em 1885, 1986 e 1987**.Porto Alegre: UFGRS, 1991.
MEC. Ministério da Educação e da Cultura. **Programa de avaliação institucional das universidades brasileiras**. Brasília: MEC, 1994.

Figura 1 – Tipos de Evasão Universitária



Em relação às possíveis causas da evasão, Tigrinho (2008) cita a repetência, principalmente em disciplinas dos primeiros semestres; a orientação vocacional/profissional, já que muitos estudantes não sabem o que cursar e acabam escolhendo um curso de forma aleatória, o que leva a mudanças de curso; o desprestígio da profissão; a conciliação entre trabalho e faculdade e a desmotivação, além de fatores sócio-econômicos interligados aos anteriores.

O banco de dados disponível para a análise contém informações dos alunos matriculados na Universidade de Brasília no período de 2004 a 2010. Inicialmente possuíam-se informações até o período de 2º/2008. Mais tarde obtiveram-se as informações dos anos de 2009 e 2010. Sendo assim, pela falta de tempo hábil para outras pesquisas, decidiu-se por estudar neste trabalho apenas a evasão dos cursos escolhidos no período já mencionado, sem levar em consideração a duração do fenômeno. Também, a partir das variáveis disponíveis no banco de dados, será possível associar os resultados do estudo com as causas relacionadas à repetência e às características acadêmico-demográficas dos alunos.

3. Regressão Logística

A principal técnica a ser utilizada neste projeto final é a Regressão Logística, que se trata de um caso especial dos Modelos Lineares Generalizados *(*GLM*,* do inglês *Generalized Linear Models*), onde a variável resposta é categórica[2], ou seja, a variável que se deseja predizer está dividida em categorias, e as variáveis explicativas, aquelas usadas para tentar explicar a variável resposta, podem ser tanto categóricas como quantitativas (assumem valores numéricos em determinada escala). Detalha-se inicialmente o modelo de regressão logística simples, onde existe apenas uma variável explicativa. A partir disso, apresenta-se a extensão para o caso em que existem mais de uma variável explicativa.

3.1. Regressão Logística Simples

Neste projeto, cada Y_i será tratado como o número de "sucessos" de uma amostra de tamanho fixo*,* onde:

➔A probabilidade de "sucesso" $= \mathrm{P}(\mathrm{Y} = 1) = \pi$.

➔A probabilidade de "fracasso" $= \mathrm{P}(\mathrm{Y} = 0) = (1 - \pi)$.

➔Y_i's são independentes entre si.

Alguns pontos importantes que devem ser levados em consideração:

- Os valores de π podem variar conforme os valores da variável explicativa X. Sendo assim, a notação a ser utilizada será $\pi(\mathbf{x})$.

- As relações entre $\pi(\mathbf{x})$ e x geralmente são mais não lineares do que lineares, ou seja, mudanças em x terão menos impacto quando π está próximo de 0 ou de 1 do que quando π está no meio do intervalo.

- As relações entre $\pi(\mathbf{x})$ e x são monotônicas.

- A transformação $\log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})})$ de π é chamada é chamada logito, representada por *logit($\pi(\mathbf{x})$).*

O modelo de regressão logística simples (com uma variável explicativa) é, portanto:

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

[2] Variáveis categóricas podem ser variáveis qualitativas nominais, qualitativas ordinais ou até mesmo variáveis quantitativas que foram alocadas em grupos de intervalo.

Através da transformação logito, pode-se reescrever o modelo de forma linear, o que equivale à ideia da regressão linear:

$$logit[\pi(x)] = \log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}) = \ \beta_0 + \beta_1 x$$

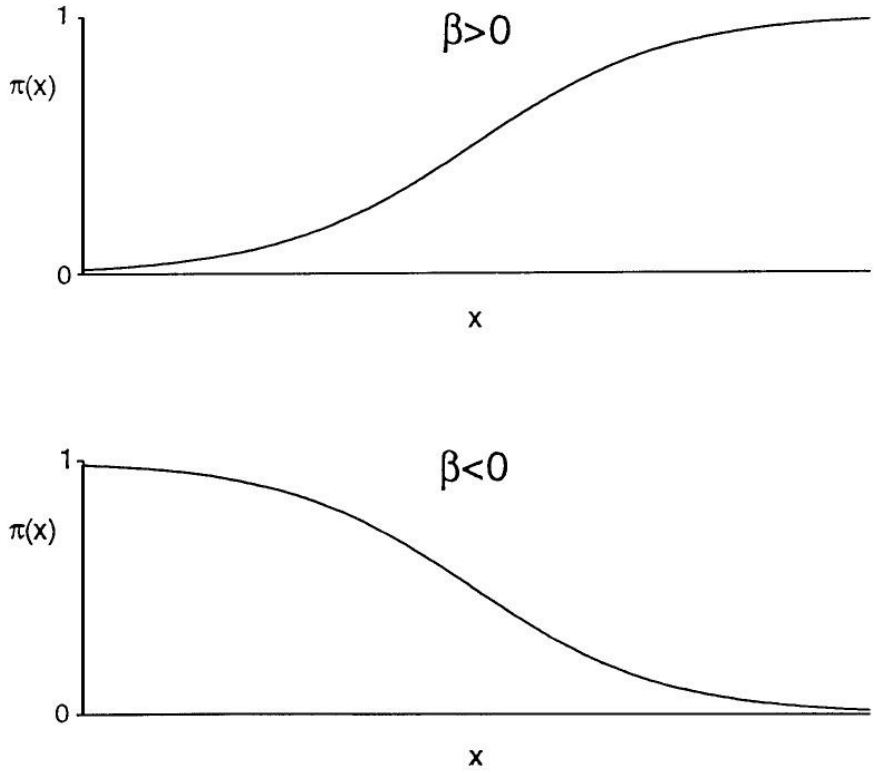
Uma das vantagens do modelo de regressão logística é que como $\pi(\mathbf{x})$ é restrito ao intervalo (0,1), a função *logit* pode assumir qualquer número real, e, portanto, é possível obter um maior o alcance para as predições.

3.1.1. Interpretação do modelo

Para a interpretação dos parâmetros é importante ter ideia de como cada parâmetro contribui para o modelo. A representação de $\pi(\mathbf{x})$ através dos níveis de x é vista através da curva S, como na figura 2.

Através da curva S, é possível observar uma das características mais interessantes do modelo de regressão logística que é a possibilidade de mudança nas taxas conforme os níveis de x variam, ao contrário dos modelos de probabilidade linear. A curva S é representada na figura a seguir:

Figura 2 – Representação da curva em forma de S para $\pi(\mathbf{x})$



O parâmetro β_1 determina a taxa de aumento ou diminuição da curva S. O sinal de β_1 indica se a curva cresce ou decresce para os níveis de x e a magnitude de crescimento/decrescimento é dada pelo valor $|\beta_1|$. Se β_1=0 a curva S se torna uma reta horizontal onde os valores de $\pi(\mathbf{x})$ se tornam constantes para os níveis de x.

Como a aparência da figura é uma curva, então as taxas de mudança de $\pi(\mathbf{x})$ variam conforme os níveis de x. Uma reta tangente à curva em um determinado valor de x nos fornece a taxa de variação de $\pi(\mathbf{x})$ neste ponto. Para a regressão logística o parâmetro β_1 nesta linha terá inclinação igual a:

$$\beta_1\pi(\mathbf{x})\big(1-\pi(\mathbf{x})\big)$$

Tem-se ainda que:

➔A inclinação se aproxima de 0 quando $\pi(\mathbf{x})$ se aproxima de 0 ou de 1, ou seja, a taxa de variação se torna muito pequena quanto mais próxima estiver de 0 ou de 1.

➔A mais íngreme inclinação ocorre quando $\pi(\mathbf{x})$=0,5 e o valor de x em que isso ocorre é igual a $x = -\frac{\beta_0}{\beta_1}$. Esse valor de x é chamado de nível efetivo da mediana e representa o nível em que cada resultado tem 50% de chance de ocorrer.

O efeito da variável explicativa pode ser interpretado através da chance (*odds*) e da razão de chances (*odds ratio*). Considerando o modelo, observa-se que:

$$odds1 = \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = \exp(\beta_0 + \beta_1 x)$$

Ao adicionarmos uma unidade de x:

$$odds2 = \frac{\pi(\mathbf{x}+\mathbf{1})}{1-\pi(\mathbf{x}+\mathbf{1})} = \exp(\beta_0 + \beta_1 (x+1))$$

Através da diferença dos logaritmos da *odds1* e da *odds2* obtém-se a razão de chances (*odds ratio*) igual a exp(β_1). Se $\beta_1 = 0$, então a razão de chances não se altera para mudanças nos níveis de x e a variável resposta é independente da variável explicativa, ou seja, não existe relação.

O logaritmo da chance (*odds*), que é a função *logit*, possui a relação linear $logit[\pi(x)] = \beta_0 + \beta_1 x$, que nos mostra que para cada aumento de uma unidade de x a *logit* irá mudar em β_1 unidades, porém, apesar da facilidade de compreensão, a interpretação através da função *logit* não é tão simples e possui uso limitado.

Em modelos logísticos, as variáveis explicativas qualitativas são trabalhadas através do uso de variáveis *dummy*. Existem diversas formas de se escolher a composição da variável *dummy*, mas por simplicidade a forma a ser adotada será:

$$X = \begin{matrix} 1\text{ , nível de interesse} \\ 0\text{, caso contrário} \end{matrix}$$

Supondo o modelo:

$$logit[\pi(x)] = \beta_0 + \beta_1 x$$

Onde a variável X é uma variável qualitativa com dois níveis (0 e 1), a própria variável está no formato de uma variável *dummy*, então basta estimar os valores para os parâmetros, os quais possuem simples interpretação.

Sejam $\hat{\beta}_0$ *e* $\hat{\beta}_1$ os valores estimados dos parâmetros. Então:

$$logit[\hat{\pi}(x)] = \hat{\beta}_0 + \hat{\beta}_1 x$$

Para o nível x=1:

$$logit[\hat{\pi}(1)] = \hat{\beta}_0 + \hat{\beta}_1 1 = \hat{\beta}_0 + \hat{\beta}_1$$

Para o nível x=0:

$$logit[\hat{\pi}(0)] = \hat{\beta}_0 + \hat{\beta}_1 0 = \hat{\beta}_0$$

Subtraindo os dois resultados, obtém-se o resultado igual a $\hat{\beta}_1$. Sendo assim, exp($\hat{\beta}_1$) é então a *odds ratio*, cuja interpretação para o modelo é semelhante ao visto para a variável explicativa quantitativa, porém, não se trata mais da chance de sucesso no aumento de uma unidade de x e sim a chance de sucesso quando se muda a categoria de referência da variável x.

Quando a variável explicativa qualitativa possui mais de 2 categorias, variáveis *dummys* adicionais serão necessárias. Em geral, quando a variável qualitativa possuir k níveis, serão necessárias k-1 variáveis *dummys* para o problema, onde cada uma assumirá a seguinte forma:

$$X_1 = \begin{matrix} 1\,, nível\ 1 \\ 0, \text{caso contrário} \end{matrix}$$

$$X_2 = \begin{matrix} 1\,, nível\ 2 \\ 0, \text{caso contrário} \end{matrix}$$

$$\vdots$$

$$X_{k-1} = \begin{matrix} 1\,, nível\ (k-1) \\ 0, \text{caso contrário} \end{matrix}$$

O nível k acontecerá quando todos os X_k's forem iguais a 0. A seguinte tabela resume esta informação:

Tabela 2 – Caracterização das variáveis *Dummy* de uma variável categórica com k níveis

Níveis da variável qualitativa	Variáveis *Dummy*				
	X_1	X_2	X_3	...	X_{k-1}
1	1	0	0	..	0
2	0	1	0	...	0
3	0	0	1	...	0
...	0	0	0	...	0
k-1	0	0	0	0	1
k	0	0	0	0	0

A interpretação para este caso também é simples, pois, de forma análoga ao caso em que x possui apenas dois níveis, pode-se comparar as chances de sucesso entre os k níveis de x dois a dois através dos valores de $\exp(\hat{\beta}_k)$ considerando o nível de referência quando todos as variáveis *Dummy* são iguais a zero.

3.1.2. Inferências para o modelo de Regressão Logística

3.1.2.1. Estimação dos parâmetros

Em regressão linear, a estimação dos parâmetros geralmente é realizada através do método dos mínimos quadrados, o qual tem por objetivo minimizar a soma dos desvios ao quadrado dos valores observados em relação aos valores estimados pelo modelo. Através deste método, os estimadores possuem algumas propriedades estatísticas desejáveis. Porém, quando a variável em estudo é dicotômica, os estimadores de mínimos quadrados perdem essas propriedades. O método a ser utilizado então é o Método de Máxima Verossimilhança, o qual estima os parâmetros em estudo que maximizam a probabilidade de se obter o conjunto de dados observados.

Primeiramente, constrói-se a função de verossimilhança, que expressa a probabilidade de se observar o conjunto de dados em função dos parâmetros desconhecidos. Os estimadores de máxima verossimilhança destes parâmetros são escolhidos para os valores que maximizam a função.

Na regressão logística, Y pode assumir os valores 0 ou 1 com probabilidade $\pi(\mathbf{x})$, então a função de verossimilhança do valor observado Y_i para cada X_i é:

$$\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

Como os $Y_i's$ são independentes, a função de verossimilhança de (Y_1,..., Y_n) é dada por:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

Por facilidade matemática, trabalha-se com o logaritmo natural da função de verossimilhança $l(\beta_0, \beta_1)$, pois é mais fácil para encontrar o máximo dessa função:

$$L(\beta_0, \beta_1) = \ln l(\beta_0, \beta_1) = \sum_{i=1}^{n} \{yi \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

Diferenciando a função $L(\beta_0, \beta_1)$ em relação a β_0 e β_1 e igualando o resultado a zero, obtêm-se as equações de verossimilhança:

$$\sum [y_i - \pi(x_i)] = 0$$

$$\sum x_i [y_i - \pi(x_i)] = 0$$

Onde:

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Ao contrário da regressão linear, as equações de verossimilhança em regressão logística não são lineares, o que torna a sua solução extremamente complicada. Apesar disso, a estimação é possível através de métodos computacionais como o método de iteração de Newton-Raphson.

O estimador de máxima de verossimilhança de $\pi(\mathbf{x})$ para um valor de x será:

$$\hat{\pi}(\mathbf{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

3.1.2.2. Intervalos de Confiança

Para grandes amostras, o intervalo de confiança para o parâmetro β_1 no modelo de regressão logística será:

$$\beta_1 \in \widehat{\beta_1} \pm z_{\alpha/2} ASE$$

Onde ASE, do inglês *Asymptotic Standard Error*, é o erro padrão assintótico de β_1.

Aplicando a função exponencial aos valores obtidos no intervalo, tem-se o intervalo para $\exp(\beta_1)$ e, assim, pode-se obter um intervalo de confiança também para a *odds ratio*.

Pode-se também aplicar os valores do intervalo de confiança na taxa de variação de $\pi(\mathbf{x})$ em um determinado valor de x: $\beta_1 \pi(\mathbf{x})\big(1 - \pi(\mathbf{x})\big)$, o que irá fornecer um intervalo de confiança da taxa de variação.

Para a construção do intervalo de confiança para o real valor de $\pi(\mathbf{x})$, pode-se utilizar da matriz de covariâncias das estimativas dos parâmetros, onde primeiramente calcula-se a variância para $\hat{\beta}_0 + \hat{\beta}_1 x$, que, para grandes amostras, possui o erro padrão assintótico igual a:

$$\text{ASE} = \sqrt{\hat{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)} = \sqrt{\hat{Var}(\hat{\beta}_0) + \hat{Var}(\hat{\beta}_1 x) + 2\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1)}$$

Sendo assim, o intervalo de confiança sob um nível de significância α para $\beta_0 + \beta_1 x$ em um determinado ponto x será igual a:

$$\big(\hat{\beta}_0 + \hat{\beta}_1 x\big) \pm z_{\alpha/2} ASE$$

Aplicando os valores obtidos no intervalo acima na fórmula:

$$\hat{\pi}(\mathbf{x}) = \frac{\exp(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \boldsymbol{x})}{1 + \exp(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \boldsymbol{x})}$$

Obtém-se então um intervalo de confiança para o real valor de $\pi(\mathbf{x})$ sob um nível significância α para um determinado valor de x.

3.1.2.3. Testes de Significância

Existem alguns métodos para se testar a significância da hipótese $H_0: \beta_1 = 0$. Para o modelo de regressão logística, a hipótese nula $H_0: \beta_1 = 0$ equivale a afirmar que a probabilidade de sucesso independe de x.

O primeiro teste a ser levado em consideração é o teste de Wald, que utiliza para grandes amostras a normalidade no método de máxima verossimilhança.

A estatística do teste de Wald é:

$$z = \frac{\widehat{\beta_1}}{ASE}$$

A qual, sob H_0, segue aproximadamente uma distribuição normal padrão. Pode-se conduzir o teste calculando-se o p-valor unilateral ou bilateral para a normal.

O segundo teste é o da Razão de Verossimilhança que utiliza a razão entre:

$$G^2 = \frac{l_0}{l_1} = \frac{máximo\ da\ função\ de\ verossimilhança\ restrito\ aos\ valores\ de\ H_0}{máximo\ da\ função\ de\ verossimilhança\ irrestrita}$$

A estatística do teste de Razão de verossimilhança é:

$$G^2 = -2\log\left(\frac{l_0}{l_1}\right) = -2\ (L_0 - L_1)$$

Que, sob $H_0: \beta_1 = 0$ e para um tamanho grande de amostra, segue aproximadamente uma distribuição qui-quadrado com um grau de liberdade.

Para tamanhos grandes de amostra, ambos os testes fornecem resultados similares, mas o teste de Razão de Verossimilhança possuiu maior poder e é mais confiável, sendo assim o mais utilizado na prática. Um problema que pode surgir é que se os resultados dos testes divergirem, então a distribuição de $\widehat{\beta_1}$ pode não ser Normal. Neste caso, testes de significância para pequenas amostras são mais adequados.

3.1.3. Verificação do Modelo

Se o modelo de regressão logística representa bem a relação entre $\pi(\mathbf{x})$ e x, pontos e intervalos de confiança fornecem informações importantes sobre o comportamento dos dados na população. Porém, é necessário investigar se realmente o modelo de regressão logística se ajusta bem aos dados amostrais.

Através do modelo logístico, que fornecerá os valores preditos de probabilidade para Y=1, o objetivo é comparar os valores observados com os ajustados utilizando a estatística X^2 de Pearson ou a razão de verossimilhança G^2:

$$X^2 = \sum \frac{(observado - ajustado)^2}{ajustado} = \sum \frac{(\mathrm{y}_i - \hat{\pi}_i n_i)^2}{\hat{\pi}_i n_i}$$

$$G^2 = 2\sum (observado)\log(\frac{observado}{ajustado}) = 2\sum (\mathrm{y}_i)\log(\frac{\mathrm{y}_i}{\hat{\pi}_i n_i})$$

Para um número fixo de grupos para cada variável explicativa e quando a maioria das frequências (75%) são maiores ou iguais a 5, X^2 e G^2 têm distribuição aproximadamente qui-quadrado com g-p graus de liberdade, onde g é o número de grupos da variável explicativa e p é o número de parâmetros do modelo. As hipóteses a serem testadas são:

$$H_0\colon\ modelo\ se\ ajusta\ bem\ aos\ dados$$
$$H_a\colon falta\ de\ ajustamento\ dos\ dados\ com\ o\ modelo$$

Valores grandes de X^2 e G^2 fornecem evidências de falta de ajustamento. Quando o ajuste não se mostra adequado, outras medidas e o estudo dos resíduos fornecem informações sobre as observações individuais de falta de ajuste no modelo.

Quando a variável explicativa é quantitativa, a criação de grupos se torna necessária para a análise de ajustamento dos dados. O problema que surge são os critérios para a construção dos grupos. Muitos grupos podem gerar problemas, pois as frequências em cada grupo poderão ser pequenas, o que inviabiliza a aproximação das estatísticas X^2 e G^2 para a distribuição qui-quadrado. Para evitar esse tipo de problema e para a criação de um padrão na construção de grupos, a técnica a ser adotada neste caso será a de Hosmer e Lemeshow, que consiste na criação de 10 grupos onde pares de observações e valores

ajustados ordenados alocados ao primeiro grupo serão os referentes ao primeiro decil, os próximos pares do segundo grupo serão referentes ao segundo decil e assim por diante. Após a criação dos grupos, o procedimento de análise de ajuste dos dados se dá de forma igual ao X^2 de Pearson e, da mesma forma, valores grandes de X^2 indicam falta de ajustamento.

3.1.3.1. Resíduos para o Modelo Logístico

Medidas de ajuste dos dados como X^2 e G^2 resumem de maneira geral a qualidade dos ajustes, porém, para uma análise mais completa, diagnósticos adicionais ajudam a descrever a natureza da falta de ajustamento. Os resíduos gerados na comparação entre as frequências dos valores observados e os valores ajustados são úteis neste propósito.

O i-ésimo resíduo de Pearson para o ajuste no grupo i é:

$$e_i = \frac{y_i - \hat{\pi}_i n_i}{\sqrt{\hat{\pi}_i n_i (1 - \hat{\pi}_i)}}$$

Onde y_i é o número de sucessos em n_i tentativas do i-ésimo grupo da variável explicativa, $\hat{\pi}_i$ será a probabilidade predita de sucesso para o modelo ajustado e $\hat{\pi}_i n_i$ será o número de sucessos ajustado.

Quando os n_i são grandes, os resíduos de Pearson seguem aproximadamente uma distribuição Normal. Se o número de observações é grande comparado com o número de parâmetros estimados, então a aproximação será para uma distribuição Normal padrão, o que implica que se $|e_i| > 2$ então nesta célula, levando em consideração a tabela bidimensional dos dados, existe indicação de falta de ajustamento dos dados. Quando as frequências relativas são pequenas para cada grupo, não se pode aproximar os resíduos para uma distribuição normal e o uso dos resíduos para esse tipo de estudo não terá significado.

3.1.3.2. Medidas de Diagnóstico de Influência

Em alguns casos pode ocorrer de um ou mais valores dentro das variáveis explicativas estarem influenciando fortemente na estimação dos parâmetros do modelo, principalmente se estes valores forem valores atípicos *(outliers)*. Um modo interessante de analisar estes casos é observar o comportamento do novo modelo ajustado sem essas observações.

Algumas medidas indicam estes sinais de influência através de novos cálculos retirando-se essas observações e realizando uma comparação entre os dois modelos. Essas medidas estão relacionadas algebricamente aos valores da diagonal da matriz **H** (*Hat*), que são chamados de observações influentes. Quanto maior o valor das observações influentes, maior o potencial de influência dessas observações na estimação dos parâmetros e, por consequência direta, podem estar contribuindo para a falta de ajuste do modelo logístico com os dados observados.

Algumas das principais medidas estão listadas abaixo:

➔Dfbetas: mostram para cada parâmetro β a mudança na estimação do parâmetro quando a observação é excluída e dividida pelo seu respectivo erro padrão.

➔*c*: é uma medida que analisa as mudanças ocorridas na estimação conjunta dos intervalos de confiança dos parâmetros $\beta's$

➔ X^2diff e G^2diff: mostram as mudanças ocorridas nas estatísticas de ajustamento quando as observações são excluídas.

Outra aplicação das observações influentes é na construção dos resíduos de Pearson ajustados que possuem valores absolutos um pouco maiores que os resíduos de Pearson, mas que seguem aproximadamente e sem problemas a distribuição Normal padrão, tendo a mesma interpretação dos resíduos de Pearson:

$$\frac{e_i}{\sqrt{1-h_{ii}}} = \frac{(y_i - \hat{\pi}_i n_i)}{\sqrt{\hat{\pi}_i n_i (1-\hat{\pi}_i)(1-h_{ii})}}$$

Uma medida alternativa para o resíduo ajustado de Pearson para a análise de ajustamento dos dados é dado pela *Deviance* residual que segue aproximadamente uma distribuição normal com uma variância um pouco menor que um. O conceito de *Deviance* será explicado mais adiante.

Para os modelos de Regressão Logística Múltipla, que serão explicados a seguir, a existência de muitas variáveis preditoras pode implicar em problemas de multicolinearidade,

que é resultante de uma forte correlação entre estas variáveis. Uma de suas formas de manifestação é quando o modelo de regressão parece ser significante, mas quando se analisa a significância dos parâmetros individualmente todos não parecem ser significantes. Um estudo gráfico e a matriz de correlações podem ajudar na identificação da multicolinearidade. Quando problemas de multicolinearidade existem, algumas medidas devem ser tomadas antes de prosseguir com a modelagem, de modo a reduzir os efeitos desta na análise.

3.2. Regressão Logística Múltipla

A base para este estudo consiste em uma generalização do modelo com apenas uma variável, fazendo uma extensão dos resultados para o caso em que possuímos mais de uma variável preditora no modelo para explicar a variável resposta binária Y_i. Sejam as p variáveis preditoras independentes, $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$, as quais podem ser tanto qualitativas quanto quantitativas. Nos casos em que existem variáveis qualitativas, uma transformação da variável em variáveis *dummys* se faz necessária. O modelo *logit* para regressão logística múltipla é dado por:

$$logit[\pi(\boldsymbol{X_i})] = \log\left(\frac{\pi(\mathbf{X_i})}{1-\pi(\mathbf{X_i})}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Em função de $\pi(\mathbf{X})$ tem-se:

$$\pi(\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}$$

Onde o parâmetro β_i se refere ao efeito de X_i dado que Y=1 (sucesso) e mantendo sob controle as demais variáveis X_i's. O valor exp(β_i) é o efeito multiplicativo na *odds* quando se aumenta uma unidade na variável X_i, mantendo-se constantes os outros níveis das demais variáveis preditoras.

3.2.1. Inferências para o modelo de Regressão Logística Múltipla

3.2.1.1. Estimação dos Parâmetros

Para a estimação dos parâmetros, assim como no caso univariado, utiliza-se o método da máxima verossimilhança. As equações normais são semelhantes ao do caso univariado, mas agora existirão p+1 equações de verossimilhança obtidas pela diferenciação do logaritmo da função de verossimilhança em relação a cada um dos p+1 parâmetros. As equações resultantes são expressas por:

$$\sum_{i=1}^{n}[y_i - \pi(\boldsymbol{x_i})] = 0$$

$$\sum_{i=1}^{n}\boldsymbol{x_{ij}}[y_i - \pi(\boldsymbol{x_i})] = 0 \quad para \quad j = 1,2,\dots,p$$

Onde:

$$\boldsymbol{x_i} = (x_{i1},\dots,x_{ip})$$

E

$$\pi(\mathbf{x_i}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})}$$

Assim como no modelo de Regressão Logística Simples, as soluções para essas equações não possuem fórmulas fechadas e a estimação dos parâmetros é obtida através de métodos computacionais.

O estimador de máxima de verossimilhança de $\pi(\mathbf{x})$ para um dado valor de **x** será:

$$\hat{\pi}(\mathbf{x_i}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip})}$$

Pode-se considerar agora o caso da estimação dos erros padrões. As estimativas dos erros padrões podem ser obtidas através da matriz de derivadas parciais de segunda

ordem obtidas do logaritmo da função de verossimilhança. As formas gerais dessas derivadas são:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(\boldsymbol{x_i})[1-\pi(\boldsymbol{x_i})] \quad j = 1,2,\dots,p$$

E

$$\frac{\partial^2 L(\beta)}{\partial \beta_l \partial \beta_j} = -\sum_{i=1}^{n} x_{ij}\, x_{il}\, \pi(\boldsymbol{x_i})[1-\pi(\boldsymbol{x_i})] \quad j = 1,2,\dots,p$$

Seja a matriz $\boldsymbol{I}(\boldsymbol{B})_{(p+1)x(p+1)}$ a matriz de informação de Fisher observada que contém os valores esperados destas derivadas. Exceto em alguns casos especiais em que não é possível escrever os elementos da matriz, as variâncias dos coeficientes estimados são obtidas através da inversa da matriz de informação.

$$Var(\beta) = \boldsymbol{I^{-1}}(\boldsymbol{\beta})$$

Então a $Var(\beta_j)$ corresponderá ao j-ésimo elemento da diagonal da matriz $\boldsymbol{I^{-1}}(\boldsymbol{\beta})$ e as covariâncias $Cov(\beta_j, \beta_l)$ serão os elementos fora da diagonal da matriz. Os estimadores $\hat{Var}(\hat{\beta}_j)$ e $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_l)$ são obtidos calculando-se $Var(\boldsymbol{\beta})$ em $\widehat{\boldsymbol{\beta}}$**.**

O erro padrão então será:

$$ASE(\widehat{\beta_j}) = \sqrt{\hat{Var}(\hat{\beta}_j)}$$

A matriz de informação estimada será:

$$\hat{I}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X'VX}$$

Onde **X** é a matriz de dados, em que as colunas correspondem às variáveis e as linhas a cada uma das n observações:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{nxp+1}$$

E **V** será a matriz de variâncias:

$$\begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{nxn}$$

3.2.1.2. Testes de Significância

Para testar a significância do modelo usa-se a seguinte hipótese:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a: \beta_j \neq 0 \; para \; algum \; j$$

Para testar esta hipótese, existe uma extensão dos testes de Wald e de Razão de Verossimilhança para a regressão logística múltipla.

Para a o teste de Razão de Verossimilhança, tem-se:

$$\frac{l_0}{l_1} = \frac{máximo\ da\ função\ de\ verossimilhança\ restrito\ aos\ valores\ de\ H_0}{máximo\ da\ função\ de\ verossimilhança\ irrestrita}$$

A estatística do teste então será:

$$G^2 = -2\log\left(\frac{l_0}{l_1}\right) = -2\,(L_0 - L_1)$$

Que, sob H_0 e para um tamanho grande de amostra, segue uma distribuição qui-quadrado com p graus de liberdade.

Para o teste de Wald, a estatística do teste será:

$$W = \widehat{\boldsymbol{\beta}}'\left[\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\left(\boldsymbol{X'VX}\right)\widehat{\boldsymbol{\beta}}$$

Que segue sob H_0 uma distribuição qui-quadrado com p+1 graus de liberdade, onde p é o número de variáveis explicativas no modelo.

Pode-se testar ainda a significância dos parâmetros individualmente através do teste de Wald:

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

A estatística do teste é:

$$z = \frac{\widehat{\beta_J}}{ASE(\hat{\beta}_j)}$$

Que, sob H_0, segue aproximadamente uma distribuição Normal padrão. Calcula-se então o p-valor unilateral ou bilateral.

3.2.1.3. Intervalos de Confiança

Para grandes amostras, os intervalos de confiança, sob um nível de significância α, para os parâmetros β_j no modelo de regressão logística múltiplo serão:

$$\beta_j \in \widehat{\beta_J} \pm z_{\alpha/2} ASE(\hat{\beta}_j)$$

Aplicando-se a função exponencial nos limites do intervalo, obtém-se o intervalo de confiança para exp(β_j) que é a *odds* do aumento de uma unidade da variável X_j mantendo-se constante as demais variáveis.

Para a construção do intervalo de confiança para o real valor de $\pi(\mathbf{x})$, utiliza-se a matriz de covariâncias das estimativas dos parâmetros, onde se calcula primeiramente o intervalo para g(**x**), tal que:

$$ASE\big(\hat{g}(\mathbf{x})\big) = \sqrt{\sum_{j=0}^{p} x_j^2 v\hat{a}r\big(\widehat{\beta_j}\big) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k c\hat{o}v(\widehat{\beta_j}, \widehat{\beta_k})}$$

E:

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}$$

Em forma matricial, $\hat{g}(\mathbf{x})$ pode ser escrito como $\hat{g}(\mathbf{x}) = \boldsymbol{x'\hat{\beta}}$, sendo os vetores $\boldsymbol{x'} = (1, x_1, \ldots, x_p)$ e $\boldsymbol{\hat{\beta}'} = (\beta_0, \ldots, \beta_p)$. Então a $\hat{Var}\big(\hat{g}(\mathbf{x})\big)$ pode ser calculada como:

$$\hat{Var}\big(\hat{g}(\mathbf{x})\big) = \hat{Var}\big(\boldsymbol{x'\hat{\beta}}\big) = \boldsymbol{x'}\hat{Var}\big(\boldsymbol{\hat{\beta}}\big)\boldsymbol{x} = \boldsymbol{x'}\big(\boldsymbol{X'VX}\big)^{-1}\boldsymbol{x}$$

Sendo assim, o intervalo de confiança sob um nível de significância α para g(**x**) em um determinado vetor de pontos **x** será igual a:

$$\hat{g}(\mathbf{x}) \pm z_{\alpha/2} ASE\big(\hat{g}(\mathbf{x})\big)$$

Por fim, para obter o intervalo de confiança para $\pi(\mathbf{x})$ aplicam-se os valores superiores e inferiores obtidos nos intervalos de confiança para $\hat{g}(\mathbf{x})$ em:

$$\frac{\exp(\hat{g}(\mathbf{x}))}{1 + \exp(\hat{g}(\mathbf{x}))}$$

3.2.2. Seleção de Variáveis para o Modelo

No modelo de regressão logística múltipla existem diversos possíveis modelos que podem ser obtidos pela combinação das variáveis preditoras do problema. Em modelos de Regressão Linear Múltipla a seleção de variáveis ocorre através de uma análise da decomposição da soma de quadrados (R^2) e soma de quadrados ajustados (R^2 *adjusted*) que fornecem noções sobre a variabilidade do conjunto de dados. Em regressão logística, uma ideia similar é aplicada através das *Deviances.* A seleção do modelo também pode ser feita de uma forma prática, mas com os devidos cuidados, através de critérios de seleção automática como *Backward, Forward* ou *Stepwise.*

3.2.2.1. Deviance

Seja:

L_s: O valor maximizado do logaritmo da função de verossimilhança do modelo saturado, ou seja, o modelo mais complexo possível (com o máximo de variáveis preditoras e suas interações).

L_m: O valor maximizado do logaritmo da função de verossimilhança de um modelo em que se está interessado.

A *Deviance* é calculada como sendo:

$$Deviance = -2\,[L_m - L_s]$$

Observando atentamente, a *Deviance* é a estatística do teste de Razão de Verossimilhança entre o modelo saturado e o modelo de interesse. Então, a *Deviance* segue uma distribuição qui-quadrado com v graus de liberdade, onde v é o número de parâmetros do modelo saturado menos o número de parâmetros do modelo de interesse.

Suponha que se deseja comparar dois modelos M_0 e M_1, sendo o modelo M_1 um modelo válido mais complexo que M_0 e M_0 um submodelo de M_1. Para testar se o modelo M_0 é também um modelo válido, realiza-se um teste de significância entre a diferença das *Deviances* dos dois modelos:

$$G^2 = -2\,[L_0 - L_1] = -2[L_0 - L_S] - (-2[L_1 - L_s]) = Deviance_0 - Deviance_1$$

Assim G^2 segue uma distribuição qui-quadrado com q graus de liberdade, onde q é igual ao número de parâmetros adicionais não redundantes existentes no modelo M_1, mas não em M_0.

3.2.2.2. Métodos Automáticos de Seleção de Variáveis

Para a seleção do modelo, todos esses métodos são métodos iterativos que adicionam e retiram as variáveis do modelo analisando a sua significância através de uma determinada regra fixa.

O método *Backward* inicia com o modelo mais complexo (saturado) e sucessivamente vai retirando os componentes do modelo. Em cada etapa elimina-se o

componente que possuir o maior p-valor obtido quando se testam as significâncias dos componentes. A seleção termina quando não existir nenhum componente com p-valor maior que um nível de significância α fixado inicialmente.

O método *Forward* inicia com o modelo mais simples (contendo somente o intercepto) e sucessivamente vai incluindo os componentes no modelo. Em cada etapa inclui-se o componente que possuir o menor p-valor obtido quando se testam as significâncias dos componentes. A seleção termina quando não existir nenhum componente com p-valor menor que um nível de significância α fixado inicialmente.

O método *Stepwise* inicia também com o modelo mais simples (contendo somente o intercepto), mas sucessivamente por etapas ele inclui e exclui componentes do modelo, sendo o método mais iterativo. Na primeira etapa calcula-se o p-valor de cada componente fora do modelo e inclui-se no modelo o que possuir o menor p-valor obtido. A partir da segunda etapa ele realiza o mesmo procedimento para analisar a inclusão de um novo componente, mas também realiza o teste para a exclusão de componentes da mesma forma que o método *Backward*. A seleção de variáveis termina quando não existr mais nenhum componente para se adicionar ou remover do modelo.

Cuidados devem ser tomados ao se adotar os métodos automáticos pois:

- Nem todos os possíveis modelos são testados nos procedimentos.
- Variáveis que são imprescindíveis no modelo podem ter sido excluídas.
- Dependendo do nível de significância α o modelo final pode ser completamente diferente.
- Quando componentes possuem um p-valor próximo ao nível de significância α é recomendado um estudo separado deste componente.

Por fim, são realizadas as etapas de estudo de resíduos e de medidas de diagnóstico para valores influentes conduzidos de forma semelhante à realizada para o modelo de regressão simples.

4.3. Modelagem

Considerando então as tabelas bidimensionais das prováveis variáveis explicativas com a variável resposta evasão (se o aluno evadiu ou não) e tomando como base as tabelas 46 e 47, que contêm um resumo das associações destas variáveis com a evasão, iniciou-se o processo de modelagem. Os resultados da modelagem foram gerados por uma rotina do PROC LOGISTIC do *software* SAS. Vale ressaltar ainda que 12 observações foram desconsideradas na estimação dos parâmetros dos modelos, já que continham dados faltantes para alguma das variáveis. Sendo assim, a amostra para a modelagem contém 544 observações. Também, a variável proporção de Disciplinas Obrigatórias com Reprovação foi multiplicada por 100, para facilitar os cálculos. Importante se atentar ao fato de que a variável Fluxo é ordinal e o respectivo parâmetro estimado foi calculado considerando-se a variável como quantitativa, utilizando escores igualmente espaçados de uma unidade. De acordo com Agresti (1996), como as categorias não são muito desbalanceadas, a escolha de diferentes escores terá pouca influência nos resultados.

Outro ponto importante de se salientar é que o nível de significância utilizado para decidir se uma possível variável explicativa seria considerada na fase de modelagem foi de 20%. Após a primeira triagem, o nível de significância utilizado na modelagem foi de 10%. Já que o objetivo deste trabalho é entender quais fatores influenciam um aluno a evadir, e não fazer predições sobre novos casos, considera-se que 10% é um risco razoável.

Na primeira triagem realizada através dos testes de associação, a variável cota de ingresso não foi considerada como significante. Sendo assim, testou-se um modelo (M_1) com as seguintes variáveis: sexo, habilitação, turno, monitoria, ingresso, cursos, semestres de permanência, fluxo, idade de ingresso e reprovação em disciplinas obrigatórias. Através do estudo descritivo surgiu a suspeita de que os efeitos dos Cursos e Habilitações não são separados, existindo assim um efeito de associação conjunto entre eles. Assim, testou-se também um segundo modelo (M_2) incluindo ao modelo anterior a interação entre habilitação e curso. Ambos os modelos mostraram-se significantes tanto no teste de razão de verossimilhança quanto no teste de Wald. As seguintes medidas foram então obtidas:

Modelo	Deviance	Gl
M2	397,5867	23
M1	377,2687	16
Diferença	20,3180	7

Segue então que com 7 graus de liberdade, o teste da razão de verossimilhança para a diferença das *deviances* apresenta um p-valor de 0,005, ou seja, a interação é

significativa no modelo. A partir então do modelo com a interação, testou-se a significância das variáveis e as seguintes foram retiradas: turno e idade de ingresso. O modelo final ficou da seguinte forma:

$$\pi(\boldsymbol{x}) = \frac{\exp(g(\boldsymbol{x}))}{1 + \exp(g(\boldsymbol{x}))}$$

Onde:

$$\begin{aligned} g(\boldsymbol{x}) = {} & 2{,}3129 - 0{,}5809 * sexo - 2{,}1787 * habilitação + 1{,}0706 * monitoria \\ & -1{,}8915 * Ciências\_Biológicas - 0{,}00082 * Física - 0{,}4388 * Geografia \\ & - 2{,}4759 * História + 1{,}0178 * Língua\_Portuguesa \\ & - 0{,}0665 * Língua\_Inglesa - 1{,}4815 * Matemática \\ & - 0{,}8257 * ingresso - 0{,}3007 * Semestres\_de\_Permanência \\ & - 0{,}1251 * fluxo + 0{,}1038 * reprovações \\ & + 1{,}1321 * habilitação * Ciências\_Biológicas \\ & + 1{,}3877 * habilitação * Física + 1{,}0041 * habilitação * Geografia \\ & + 3{,}9992 * habilitação * História \\ & + 1{,}7625 * habilitação * Língua\_Portuguesa \\ & + 4{,}0318 * habilitação * Língua\_Inglesa \\ & + 3{,}2407 * habilitação * Matemática \end{aligned}$$

Neste modelo, *reprovações* são as porcentagens de disciplinas obrigatórias com reprovação.

O teste da razão de verossimilhança com a hipótese inicial de que todos os parâmetros do modelo são iguais à zero foi rejeitado com um p-valor menor que 0,0001, ou seja, pelo menos um dos parâmetros é diferente de zero. Porém, como se pode observar na tabela abaixo, alguns níveis dessas variáveis não foram significativos na tentativa de explicar a evasão e, por isso, seus efeitos não foram estimados. Os valores das estatísticas dos testes de Wald podem ser encontrados na tabela abaixo:

Tabela 48 – Estimativas e Estatísticas de Wald do Modelo Logístico.

Parâmetro		Estimativa	G.L.	Estatística de Wald	P-valor
Intercepto		2,3129	1	4,971	0,0258
Sexo	(Feminino vs. Masculino)	-0,5809	1	3,721	0,0537
Habilitação	(Bacharelado vs. Licenciatura)	-2,1787	1	7,2054	0,0073
Monitoria	(Não vs. Sim)	1,0706	1	10,3901	0,0013
Ciências Biológicas	(C1 vs. Química)	-1,8915	1	7,426	0,0064
Física	(C2 vs. Química)	-0,00082	1	0	0,999
Geografia	(C3 vs. Química)	-0,4388	1	0,2794	0,5971
História	(C4 vs. Química)	-2,4759	1	3,8564	0,0496
Língua Portuguesa	(C5 vs. Química)	1,0178	1	1,7613	0,1845
Língua Inglesa	(C6 vs. Química)	-0,0665	1	0,0117	0,9137
Matemática	(C7 vs. Química)	-1,4815	1	4,058	0,044
Ingresso		-0,8257	1	6,2906	0,0121
Semestres de Permanência		-0,3007	1	28,9407	<0,0001
Fluxo		-0,1251	1	3,3878	0,0657
Porcentagem de Reprovações		0,1038	1	68,3894	<0,0001
Bacharelado* Ciências Biológicas		1,1321	1	1,0175	0,3131
Bacharelado* Física		1,3877	1	1,3848	0,2393
Bacharelado* Geografia		1,0041	1	0,6614	0,4161
Bacharelado*História		3,9992	1	6,6388	0,01
Bacharelado* Língua Portuguesa		1,7625	1	1,476	0,2244
Bacharelado* Língua Inglesa		4,0318	1	12,835	0,0003
Bacharelado* Matemática		3,2407	1	7,3061	0,0069

Para testar se o modelo representa bem a relação entre π(**x**) e **x**, usou-se a técnica de Hosmer e Lemeshow para testar a hipótese de que os dados se ajustam bem ao modelo. Obteve-se então uma estatística X^2=6,2057 com 8 graus de liberdade, o que resulta em um p-valor de 0,6242, ou seja, a um nível de significância de 10% não existem evidências para rejeitar a hipótese nula do teste e portanto os dados parecem se ajustar bem ao modelo. É importante citar também que o modelo escolhido foi o que apresentou o menor critério de Akaike (AIC) dentre os que foram testados.

Antes de continuar com as análises, devem-se analisar os diagnósticos deste modelo. Os gráficos a seguir contêm as informações necessárias para averiguar se existem pontos com falta de ajustamento ou pontos discrepantes que estejam influenciando muito na estimação dos parâmetros de forma a prejudicar os resultados.

Observe que para os resíduos de Pearson, as observações 204, 206 e 325 são as mais discrepantes e indicam que estas não foram bem estimadas pelo modelo. Como uma alternativa, o gráfico da *Deviance* residual também fornece informações sobre faltas de ajustamento e as mesmas observações se sobressaem, porém, ainda dentro do limite

desejado para este resíduo, que segue aproximadamente uma distribuição normal padrão. Então, em relação aos resíduos, não existem muitos casos alarmantes.

Gráfico 39 – Resíduos de Pearson do Modelo Logístico Ajustado.

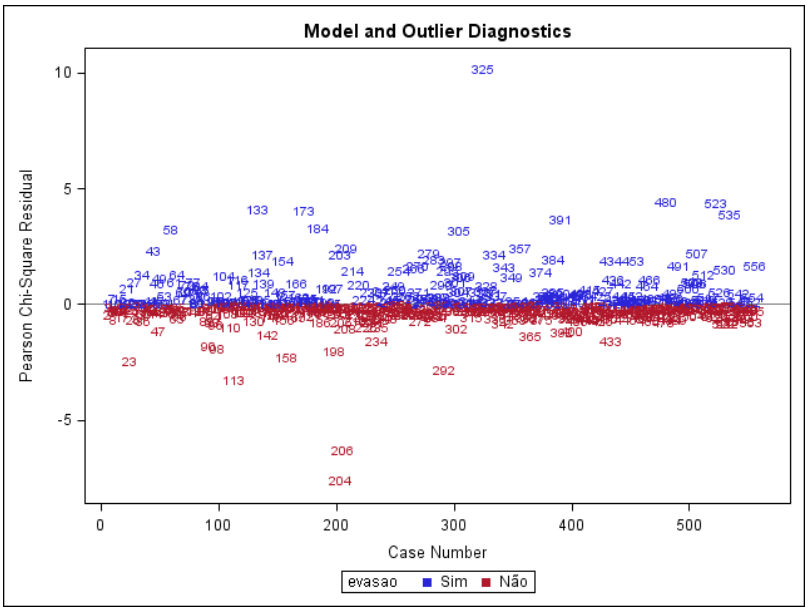
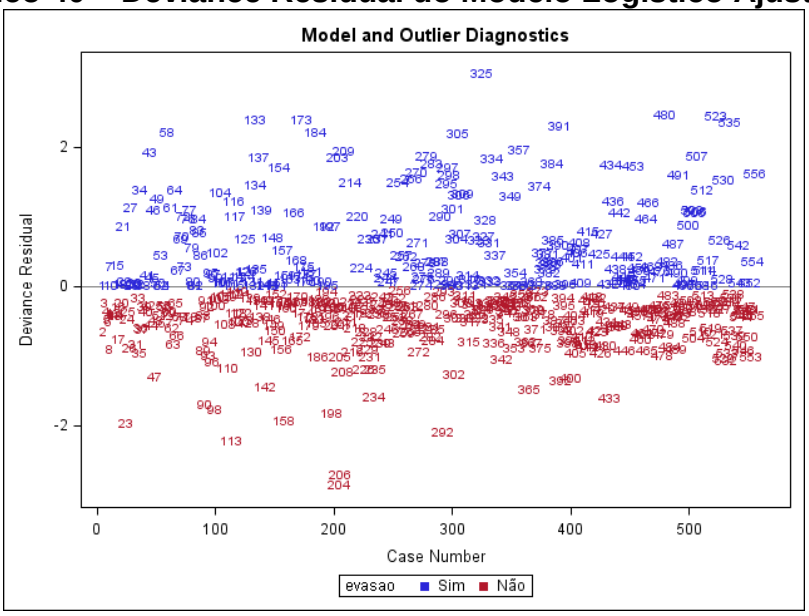


Gráfico 40 – Deviance Residual do Modelo Logístico Ajustado.



Já em relação aos DfBetas, somente a observação 58 representa grande influência na estimação dos parâmetros relacionados ao curso de história. Para todos os outros parâmetros os DfBetas são menores que 0,5.

Gráfico 41 – Diagnóstico do Modelo Logístico Ajustado para o Curso de História: DfBetas.

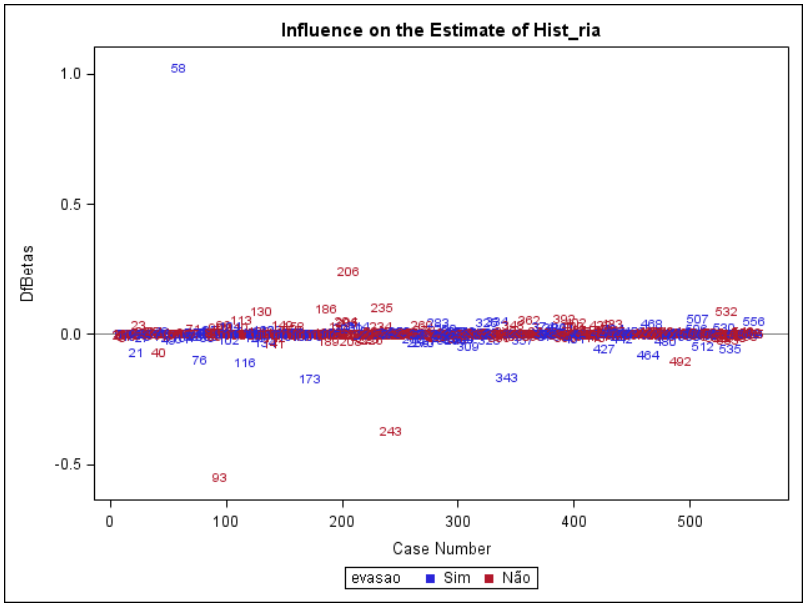
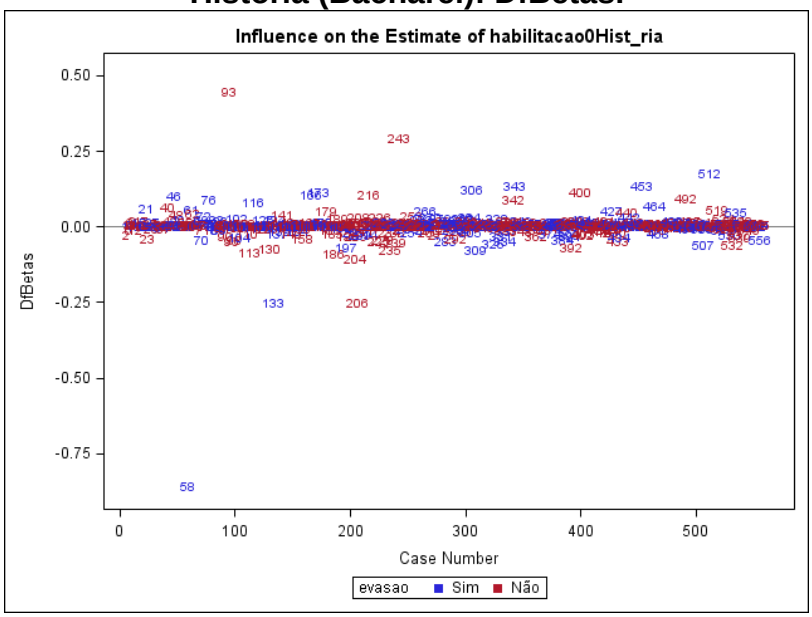
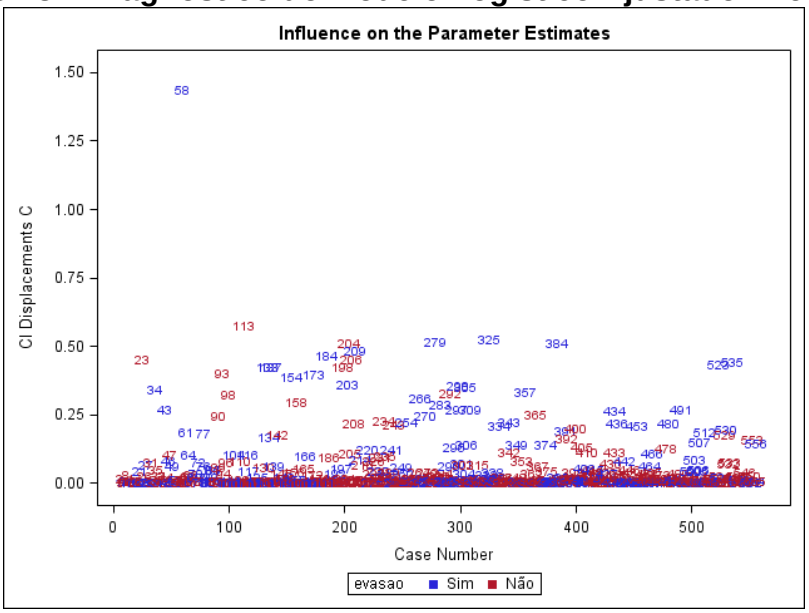


Gráfico 42 – Diagnóstico do Modelo Logístico Ajustado para o Curso de História (Bacharel): DfBetas.



Em se tratando da estimação conjunta dos intervalos de confiança para os parâmetros, novamente a observação 58 se sobressai como uma medida influente (medida C).

Gráfico 43 – Diagnóstico do Modelo Logístico Ajustado: medida C.



Para as estatísticas de ajustamento, é fácil notar que novamente as observações 204, 206 e 325 se destacam. Essas observações correspondem a estudantes do sexo masculino, de Licenciatura em Matemática, Licenciatura em Química e Bacharelado em Ciências Biológicas respectivamente e não há muitas características em comum entre eles. Porém, como uma análise mais profunda percebe-se que as probabilidades estimadas para estes casos são de 0,983, 0,975 e 0,009, que são os valores extremos que o modelo pode assumir.

Gráfico 44 – Diagnóstico do Modelo Logístico Ajustado: X^2 $Diff$.

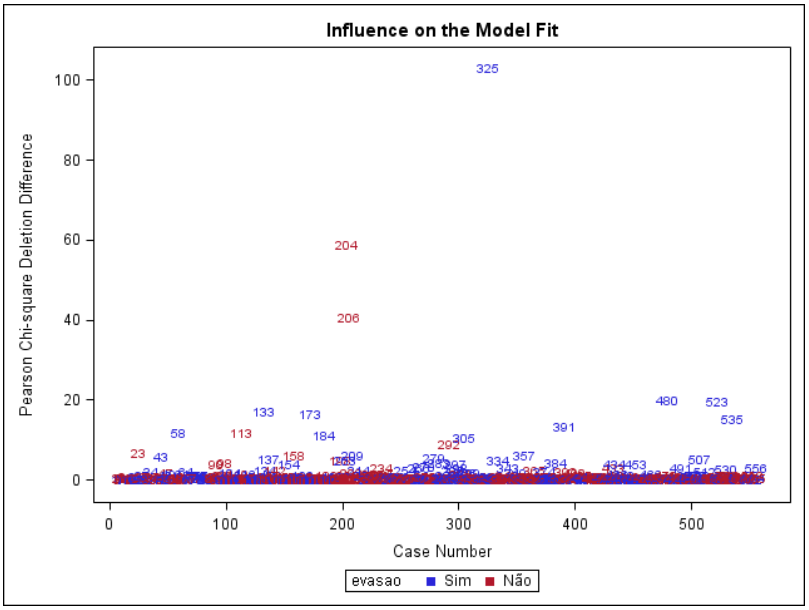
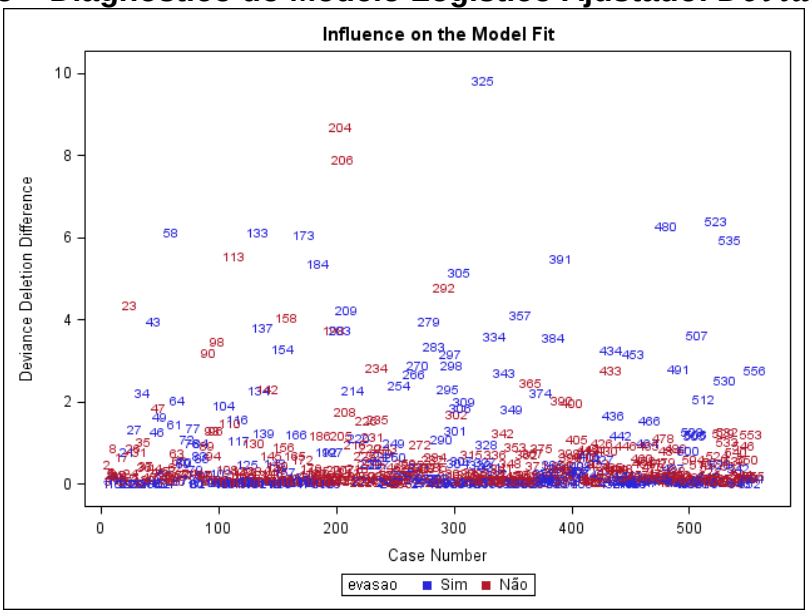


Gráfico 45 – Diagnóstico do Modelo Logístico Ajustado: $Deviance\ Diff$.



De um modo geral, pode-se dizer que o modelo está adequado, embora algumas poucas informações apresentem resultados indesejáveis nos estudos de diagnóstico. Sendo assim, não se fazem necessários estudos mais aprofundados sobre reponderar as informações influentes no modelo ou excluí-las da análise. Com isso, segue a análise do modelo com a interpretação das razões de chances.

Tabela 49 – Estimativas das Razões de Chances para as Variáveis Significativas do Modelo Logístico.

Variável	Efeito	Razão de Chances	Intervalo de Confiança (10%)	
			Inferior	Superior
Sexo	Fem. vs. Masc. (Masc. vs. Fem.)	0,5589 (1,789)	0,341	0,918
Monitoria	Não vs. Sim	2,917	1,689	5,038
Ingresso		0,438	0,255	0,753
Semestres de Permanência		0,74	0,675	0,812
Fluxo		0,882	0,789	0,987
Reprovação em Disciplinas Obrigatórias		1,109	1,087	1,133

Observe que a chance de um estudante do sexo masculino evadir é 78,9% maior que a chance de um estudante do sexo feminino evadir, mantidas constantes as outras variáveis, ou seja, estudantes do sexo masculino estão mais sujeitos à evasão neste estudo. Adicionalmente, um estudante que não foi monitor tem 191,7% mais chances de evadir que um estudante que já foi monitor, mantidas constantes as outras variáveis.

Já para a variável número de ingressos na universidade, tem-se que a chance de evasão fica 56,2% menor quando se aumenta de uma unidade o número de ingressos, mantidas constantes as outras variáveis, ou seja, pessoas com mais ingressos na universidade têm menos chance de evadir. Em relação aos semestres de permanência e ao fluxo, tem-se respectivamente que, mantidas constantes as demais variáveis, as chances de evasão ficam 26% e 11,8% menor no aumento de uma unidade do semestre de permanência e no aumento de uma categoria ordinal do fluxo respectivamente. Também, quanto maior é a taxa de reprovação em disciplinas obrigatórias, maior é a chance de evadir. Ao se aumentar um ponto percentual na taxa de reprovação tem-se que a chance de evadir aumenta em 10,9%.

Como no modelo a interação entre o curso e a habilitação foi significativa, não faz sentido analisar estas variáveis separadamente. Para as interações destas variáveis têm-se:

Tabela 50 – Estimativa das Razões de Chances para as Interações entre Curso e Habilitação.

Variável	Bacharelado			Licenciatura		
Curso	Razão de Chances	Intervalos de Confiança (90%)		Razão de Chances	Intervalos de Confiança (90%)	
		Inferior	Superior		Inferior	Superior
Ciências Biológicas vs Química	0,468	0,108	2,032	0,151	0,048	0,472
Física vs Química	4,003	0,791	20,261	0,999	0,327	3,051
Geografia vs Química	1,76	0,392	7,903	0,645	0,165	2,526
História vs Química	4,588	1,141	18,454	0,084	0,011	0,669
Língua Portuguesa vs Química	16,124	2,027	128,255	2,767	0,784	9,77
Língua Inglesa vs Química	52,734	10,718	259,459	0,936	0,341	2,569
Matemática vs Química	5,808	1,283	26,285	0,227	0,068	0,762

Nesta tabela constam as razões de chances para todos os cursos em comparação à referência, que é o curso de Química, para cada habilitação. Segue então que os intervalos de confiança que contêm o número 1 indicam que a combinação das variáveis parece não influenciar na evasão. Observa-se que os limites de confiança para o Bacharelado são maiores que para a Licenciatura, onde os valores são mais estáveis. Esperava-se que os resultados desta tabela fossem iguais àqueles listados na tabela 48, principalmente para os cursos de Bacharelado, porém, observou-se que naquela tabela a interação entre habilitação e o curso de Língua Portuguesa para o Bacharelado não foi significativa para explicar a evasão e na tabela 50 ocorre o contrário, mas com um intervalo de confiança para a razão de chances muito amplo. Para o bacharelado, alunos do curso de História têm 4,588 vezes mais chances de evadir que um aluno de Química; de Inglês, 16,124 vezes mais

chances; de Língua Portuguesa, 52,734 vezes mais de chances e de Matemática 5,808 vezes mais chances de evadir em relação a um aluno do curso de Química.

Já para as licenciaturas, vê-se que um aluno do curso de Química tem 6,62 vezes mais chances de evadir que um aluno do curso de Ciências Biológicas, 11,9 vezes mais chances de evadir que um aluno do curso de História e 4,4 vezes mais chances de evadir que um aluno do curso de Matemática. Os demais cursos de licenciatura não apresentaram diferenças em relação à Química. Percebe-se então que nos bacharelados Química é um curso com menos chances de evasão, mas na licenciatura, é um curso com mais chances de evadir em comparação com os demais cursos significativos.

5. Conclusão

Após diversas etapas do estudo foi possível traçar um perfil dos alunos ingressantes em 2º/2004. Com o auxílio da técnica de Regressão Logística foi possível chegar a um modelo final que fornece informações imprescindíveis ao estudo da evasão.

São muitas as variáveis relacionadas aos alunos da Universidade de Brasília, porém, algumas delas foram selecionadas para que pudesse ser feito um delineamento do perfil destes estudantes. Sendo assim, as tabelas abaixo trazem a caracterização desse perfil primeiramente em relação às diferentes habilitações e em seguida à forma de saída da universidade. Para as variáveis qualitativas foram considerados os valores de maior frequência observados na caraterização do perfil dos estudantes, enquanto que para as variáveis quantitativas foram consideradas as médias mais ou menos um erro padrão:

Tabela 51 – Perfil dos Estudantes da Amostra Final pelas Habilitações.

Variável	Perfil	
	Bacharelado	**Licenciatura**
Sexo	Masculino (61,16%)	Masculino (56,05%)
Turno	Diurno (100%)	Noturno (60%)
Área do Conhecimento	Ciências Exatas (34,71%)	Ciências Exatas (39,49%)
Monitoria	Não (68,18%)	Não (70,38%)
Idade de Ingresso	$21{,}89 \pm 5{,}43$ (C.V.= 24,38%)	$23 \pm 6{,}64$ (C.V. = 28,78%)
Porcentagem de Disciplinas Obrigatórias com Reprovação	$20\% \pm 29\%$ (C.V. = 145,4%)	$19{,}26\% \pm 25\%$ (C.V. = 134,64%)
Semestres de Permanência	$5{,}91 \pm 3{,}4$ (C.V. = 57,46%)	$6{,}28 \pm 2{,}94$ (C.V. = 46,93%)

Tabela 52 – Perfil dos Estudantes da Amostra Final pela Forma de Saída.

Variável	Perfil	
	Não Evadiu	**Evadiu**
Sexo	Igual (50%/50%)	Masculino (64,49%)
Turno	Diurno (74,37%)	Diurno (57,20%)
Área do Conhecimento	Ciências Biológicas (30,93%)	Ciências Exatas (54,23%)
Monitoria	Não (55,94%)	Não (87,71%)
Habilitação	Licenciatura (53,75%)	Licenciatura (60,16%)
Porcentagem de Disciplinas Obrigatórias com Reprovação	$5{,}33\% \pm 8{,}65\%$ (C.V.= 162,09%)	$39\% \pm 31{,}81\%$ (C.V.= 80,51%)
Semestres de Permanência	$6{,}85 \pm 3{,}19$ (C.V.= 46,60%)	$5{,}13 \pm 2{,}82$ (C.V.= 54,95%)

Os cursos de História, Geografia, Matemática, Física, Química, Ciências Biológicas, Língua Portuguesa e Respectiva Literatura e Língua Inglesa e Respectiva Literatura foram

os cursos selecionados e considerados os de maior peso e de formação básica nos processos seletivos de ingresso na Universidade de Brasília.

Observou-se uma quantidade maior de estudantes do sexo masculino tanto na coorte inicial que continha todos os alunos ingressantes em 2º/2004, quanto na amostra final que continha os cursos selecionados. A variável Sexo foi considerada uma variável importante no modelo final, onde foi constatado que a chance de uma estudante do sexo feminino evadir é quase a metade da chance de um estudante do sexo masculino.

O período estudado foi o primeiro no qual o sistema de cotas para negros foi implementado na Universidade de Brasília. Na amostra final os estudantes cotistas representavam aproximadamente 12% do total de estudantes. Apesar de ter sido observado notas de corte muito inferiores para os cotistas em relação ao sistema universal no vestibular, a variável Cota de Ingresso foi a primeira variável a ser removida do estudo, pois foi considerada a menos significante no modelo.

A variável Turno não foi considerada significante no estudo.

As Idades de Ingresso dos estudantes foram muito parecidas em todas as análises realizadas, sendo os estudantes de Licenciatura um pouco mais velhos que os estudantes de Bacharelado e também não foi um fator significante para a evasão.

O estudo mostrou como a realização ou não de uma atividade acadêmica extra pode influenciar na evasão, neste caso, a atividade extra foi representada pela Monitoria. Na amostra, menos de um terço dos estudantes foram monitores pelo menos uma vez, mas foi visto que os estudantes que não foram monitores nenhuma vez têm quase três vezes a chance de evadir do que estudantes que foram monitores pelo menos uma vez.

A maioria dos estudantes na amostra está no seu primeiro ingresso na universidade, entretanto, observa-se uma quantidade significativa de estudantes em ingressos posteriores. Para os estudantes em ingressos posteriores as maiores proporções são de estudantes ingressantes por Dupla Habilitação, Mudança de Curso e por um pequeno grupo de estudantes que parecem reingressar na universidade em um momento perto da conclusão do curso, cujo objetivo era a limpeza do histórico escolar. Assim, o Ingresso no qual o estudante se encontra foi outro fator considerado significativo no estudo, onde quanto maior o ingresso do estudante menor a chance de evasão, o que faz todo o sentido considerando-se as principais formas de ingresso nos ingressos posteriores ao primeiro. Informações semelhantes (mas com pesos diferentes) foram observadas para o aumento de uma posição na categoria do Fluxo e para o Número de Semestres de Permanência.

Apesar da dificuldade em se criar a variável que representa a proporção de Disciplinas Obrigatórias com Reprovação, esta se mostrou significante para o estudo. Observou-se uma proporção média praticamente igual entre as Habilitações, porém, a

distribuição é completamente diferente entre os estudantes que evadem e os estudantes que não evadem. A porcentagem (proporção X 100) média de reprovações para os estudantes que não evadem é de aproximadamente 5% enquanto que para os estudantes que evadem é de 39%. No modelo final, o aumento de 1% na porcentagem de reprovações implica em um aumento de 10,9% nas chances de evasão. Com base nesta informação observa-se a importância de uma maior atenção da universidade em relação ao rendimento dos seus estudantes.

Por fim, analisou-se uma das principais variáveis de interesse deste estudo sobre evasão: as Habilitações. Na amostra em estudo a maior proporção de estudantes se encontra nas licenciaturas. Observou-se que os estudantes das licenciaturas são um pouco mais velhos que os de bacharelado e que apesar de terem taxas de reprovações parecidas, os estudantes de licenciatura evadem um pouco mais do que os estudantes de bacharelado. Através da Regressão Logística foi possível averiguar a existência de efeito significante das habilitações em relação à evasão, porém este efeito não ocorreu sozinho. Suspeitava-se da existência de uma possível interação entre as habilitações e os cursos e tal suspeita foi confirmada durante a modelagem. Portanto, as análises dos efeitos das habilitações e dos cursos na evasão não puderam ser realizadas separadamente, pois não faria sentido. Analisando-se as interações (utilizando como referência o curso de Química) observou-se que, no bacharelado, um estudante dos cursos de História, Língua Portuguesa e Matemática, possuía muito mais chances de evasão que um estudante de Química, chegando até aproximadamente 53 vezes de chances a mais de evadir. Já para a licenciatura a situação se inverte. Um estudante de Química possui chances de evasão maiores que os estudantes dos cursos de História, Língua Portuguesa e Matemática, sendo a maior destas chances de quase 13 vezes mais. Os demais cursos não apresentaram diferença significativa em relação ao curso de Química. Observou-se assim que nos bacharelados Química é um curso com menos chances de evasão, mas na licenciatura, é um curso com mais chances de evadir em comparação com os demais cursos significativos.

Através das estatísticas descritivas e da técnica de Regressão Logística foi possível traçar um perfil dos estudantes e comparar as suas características, principalmente entre as habilitações de licenciatura e bacharelado. Estudos posteriores podem ser realizados com base neste trabalho, onde, por exemplo, podem-se verificar possíveis mudanças em períodos posteriores ao 2º/2004, inclusão de novas variáveis que poderiam ser obtidas através de uma pesquisa de campo.

Referências Bibliográficas

AGRESTI, Alan. **An Introduction to Categorical Data Analysis,** John Wiley & Sons, New York, 1996.

ALMEIDA, Edson Pacheco; VELOSO, Tereza Christina M. A. **Evasão nos cursos de graduação da Universidade Federal de Mato Grosso, campus universitário de Cuiabá:** Um processo de exclusão**.** Cuiabá: UFMT. 2002.

COLLET, D., **Modelling Binary Data**, Chapman e Hall, 1994.

DE PAULA, Caio César M. Ribeiro. **Estudo da evasão na Universidade de Brasília (UnB).** Brasília: UnB. 2010.M

HOSMER, D. W. e LEMESHOW, S.. **Applied Logistic Regression,**.John Wiley & Sons, New York, 1989.

MINISTÉRIO DE EDUCAÇÃO E CULTURA. Secretaria de Ensino Superior. **Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras.** ANDIFES/ABRUEM,SESu,MEC, Brasília, 1996. 134p.

_____. Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira. **Resumo Técnico – Censo da Educação Superior de 2009.** Brasília, 2010. 37 p.

_____. Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira. **Resumo Técnico – Censo Escolar 2010.** Brasília, 2011. 42 p.

_____. Instituto Nacional De Estudos E Pesquisas Educacionais Anísio Teixeira. **Estudo Exploratório sobre o Professor Brasileiro:** Com base nos resultados do Censo Escolar da Educação Básica de 2007 **.**Brasília, 2011. 68 p.

MIRANDA, Derlinéa P. M.. **Gestão da evasão nas instituições de ensino superior privado:** um estudo sobre cursos de administração no estado do espírito santo.2006. 104 f. Dissertação (Mestrado em gestão e Estratégia em Negócios) – Instituto de Ciências Humanas e Sociais, Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, 2006.

STOKES, M. E., DAVIS, C. S. e KOCH, G.G., **Categorical Data Analysis Using the SAS System***,* SAS Institute, 1995.

REDAÇÃO, MEC quer aluno por mais tempo na escola. Destak, São Paulo, 14 de setembro de 2011. Disponível em: <http://www.destakjornal.com.br/readContent.aspx?id=14%2C109178>. Acesso em: 19 de setembro 2011.

TIGRINHO, Luiz Maurício. **Evasão Escolar nas instituições de ensino superior**. Gestão Universitária, 17 de setembro de 2008. Disponível em: <http://www.gestaouniversitaria.com.br/index.php?option=com_content&view=article&id=649:evasao-escolar-nas-instituicoes-de-ensino-superior&catid=135:173&Itemid=21>. Acesso em: 19 de outubro de 2011.

VIDA UNIVERSITARIA. Evasão é relacionada à má escolha da carreira. Disponível em: <http://www.vidauniversitaria.com.br/blog/?p=16464>. Acesso em 15 de setembro 2011.