**Universidade Federal de Minas Gerais**
**Curso de Especialização em Estatística**
**Departamento de Estatística**

# A Utilização do Modelo Discriminante Logístico na Análise de Dados de Avaliação Sistêmica

Monografia apresentada ao Departamento de Estatística do instituto de Ciências Exatas, como requisito parcial à obtenção do Título de Especialista em Estatística

Aluno: Lucas Rodrigues Duarte
Orientadora: Profa. PhD. Sueli Aparecida Mingoti

**Belo Horizonte**
**2013**

# A Utilização do Modelo Discriminante Logístico na Análise de Dados de Avaliação Sistêmica

**Resumo**

Neste trabalho avaliou-se a utilização do modelo discriminante de regressão logística como ferramenta de análise de dados, em um teste simulado de uma avaliação sistêmica, que foi aplicado em uma instituição de ensino superior de Belo Horizonte, em julho de 2012. A proposta consiste na busca de um modelo de regressão, que destaque fatores socioeconômicos capazes de explicar o desempenho dos alunos na prova.

**Palavras-chaves:** Modelo descriminante, regressão logística, avaliação sistêmica.

## 1 Introdução

A partir dos anos 90, o modelo de avaliação sistêmica vem ganhando cada vez mais espaço em diversos níveis de ensino, como uma modalidade de avaliação capaz de fornecer respostas a problemas educacionais, e ao mesmo tempo, com o objetivo de melhorar a qualidade do ensino e aprendizado (Vianna, 2003).

Atualmente existem diversas avaliações sistêmicas sendo utilizadas pelo ministério da Educação e demais secretarias de Educação estaduais, desde o ensino fundamental ao ensino superior. Em nível nacional: Enadem- Exame Nacional de Desempenho de Estudantes, avaliação aplicada ao ensino superior; Enem- Exame Nacional do Ensino Médio, avaliação em nível médio; Prova Brasil (Saeb)- Sistema Nacional da Educação Básica, avaliação realizada no último ano do ensino fundamental.

Embora cada avaliação esteja associada a diferentes modalidades de ensino, suas propostas são semelhantes, exceto Enem, que também passou a ser um exame de seleção para ingresso em instituições de ensino superior. As propostas destas avaliações têm como objetivo diagnosticar, em larga escala, a qualidade do ensino e aprendizado através de testes padronizados.

A crescente relevância das avaliações sistêmicas em vigor tem obrigado as instituições de ensino a criar alternativas e teste simulados, que de certa forma preparem, ou até mesmo condicionem seus egressos à realidade destas provas.

Uma questão importante vinculada a qualquer avaliação sistêmica é a metodologia utilizada para o tratamento e análise dos dados provenientes deste processo. Seguindo esta

questão, a proposta deste trabalho consiste na utilização do modelo discriminante de regressão logística como uma alternativa para análise de dados de uma avaliação sistêmica, aplicado em um teste simulado, em uma instituição de ensino superior de Belo Horizonte.

A utilização do modelo logístico pode ser bem aplicada à análise de desempenho de avaliações padronizadas através de medida de satisfação (BERTOLOTTI, 2003).

De acordo com esta proposta de pesquisa, surge a seguinte questão: "quais variáveis, ou fatores socioeconômicos interferem no desempenho do aluno, em um simulado de uma avaliação sistêmica?"

Diante desta questão investigativa, o objetivo geral deste trabalho é estimar um modelo discriminante de regressão logística que possibilite detectar alguns fatores que interferem no desempenho acadêmico.

Nesta realidade de pesquisa, pretende-se: avaliar os dados obtidos através da aplicação de um questionário socioeconômico que foi elaborado e que serviu de veículo para o levantamento de dados, dos alunos, durante a realização da prova; e a partir desses dados, ajustar um modelo de regressão logística que permita identificar os fatores relacionados ao desempenho dos alunos na prova.

## 2 Modelo Discriminante Logístico

Em muitas situações nas quais as variáveis explicativas que podem estar associadas a ocorrência de um determinado evento são binárias ou qualitativas ordinais ou nominais, o modelo de análise discriminante mais comumente utilizado é o modelo logístico (Mingoti, 2005).

O modelo logístico é utilizado para avaliar como a resposta do tipo binária se relaciona com um conjunto de variáveis explicativas. Considera-se a situação na qual tem-se duas populações, onde para cada elemento da população/amostra observa-se um vetor aleatório $X = \lfloor X_1, X_2, X_3, \ldots, X_p \rfloor$, tal que para um dado experimento $\xi$, há apenas dois resultados possíveis: sucesso ou fracasso; sim ou não; aprovado ou reprovado, por exemplo. Seja *Y* a variável aleatória definida como: 1 se o resultado for sucesso e 0 se for fracasso.

No experimento $\xi$, a variável *Y* tem uma distribuição Bernoulli $(Y \sim Bernoulli(p))$; assim sua distribuição de probabilidade é definida como:

$$P(Y = y) = \pi^y (1-\pi)^{1-y} \text{ , } y=0,1, \quad 0<\pi<1. \qquad (1.0)$$

sendo $\pi$ a probabilidade de que o resultado do experimento seja sucesso, ou seja a probabilidade de que $Y$ assuma o valor 1, e $(1-\pi)$ a probabilidade de que o resultado seja fracasso, ou seja a probabilidade de que $Y$ assuma o valor zero.

Com o modelo logístico, é possível estimar a probabilidade de cada elemento amostral pertença a cada uma das populações, através das expressões:

$$P(1) = \frac{e^{\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}}{1 + e^{\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}} \quad \text{e} \quad P(0) = \frac{1}{1 + e^{\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right)}} \tag{1.1}$$

onde a soma $P(1)+P(0)=1$. Quando se utiliza o logaritmo natural para comparar a razão entre as equações de $P(1)$ e $P(0)$, isto é,

$$\ln\left(\frac{P(1)}{P(0)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{1.2}$$

pode-se observar a resultante como um modelo de regressão linear múltipla, definido como função de ligação canônica. Uma ilustração do modelo logístico com uma variável explicativa é apresentada, no Gráfico 1.

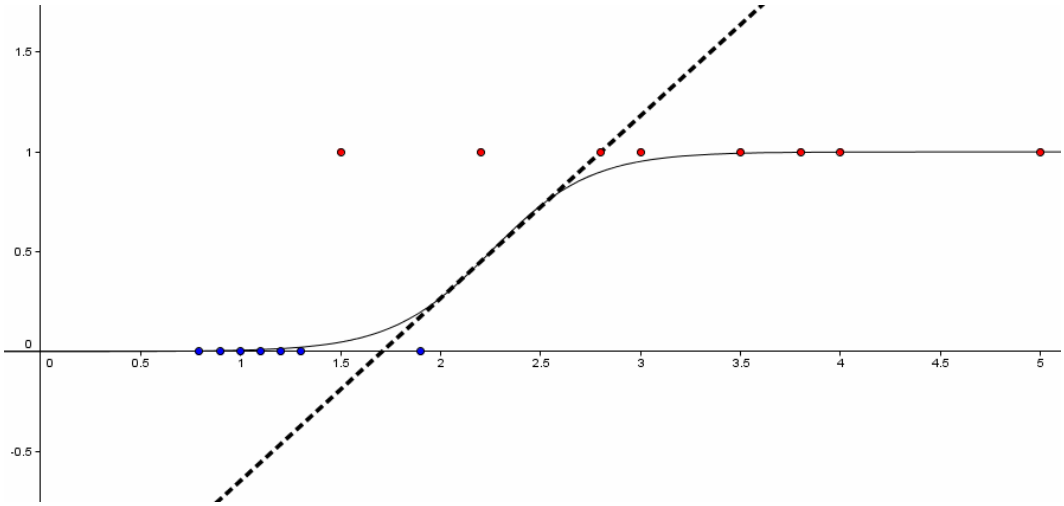


**Gráfico 1: Representação do modelo Logístico**

**Fonte: Elaborado com *Geogebra***

Quando avaliamos a razão entre as equações $P(1)$ e $P(0)$, estamos avaliando a razão entre a probabilidade do evento ocorrer e de não ocorrer, que é definida como razão das chances *(Odds Ratio)*:

$$\frac{P(1)}{P(0)} = \frac{\pi}{1-\pi} = e^{\beta_0 + \sum_{i=1}^{p} \beta_i x_i} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} \tag{1.3}$$

Na expressão (1.3) os coeficientes: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são interpretados como a variação na proporção das probabilidades definida como razão de desigualdades, e convenientemente expresso pelo logaritmo natural.

A taxa de aumento/redução da probabilidade de sucesso *(odds ratio)*, quando há um aumento de uma unidade na variável preditora $(x_1 + 1)$, considerando os valores das outras variáveis preditoras fixos, está associada ao coeficiente $\beta_1$; o fator $e^{\beta_1}$ indica um aumento/redução da probabilidade de sucesso para cada unidade adicionada a variável preditora. Assim, se por exemplo o valor do fator for $e^{0,0935} = 1{,}10$, para um acréscimo de uma unidade na variável preditora respectiva, há um aumento de 10% na probabilidade de sucesso. Os parâmetros do modelo são estimados através do método estatístico de máxima verossimilhança (Paula, 2013).

## 2.1 Estimador de Máxima Verossimilhança

Seja *X* uma variável aleatória com função densidade $f(x, \theta)$, com o vetor de parâmetros desconhecido $\theta$ . Dada uma amostra aleatória com $(x_1, x_2, \ldots, x_n)$ observações, então chama-se *L(θ)* função de máxima verossimilhança definida como:

$$L(\theta) = f(x_1; \theta).f(x_2; \theta)\ldots f(x_n; \theta) \qquad (1.4)$$

O estimador de máxima verossimilhança de $\theta$ é o vetor de valores $\hat{\theta}$ que maximiza a função verossimilhança *L(θ)* (Montgomery, 2003). É conveniente definir *log-verossimilhança* por $\ln(\theta|x)$, que também tem seu máximo em $\hat{\theta}$ .

## 2.2 Estimando os coeficientes $\beta_i$

Dada uma amostra aleatória de *n* elementos tais que $n_1$ possuem o atributo e $n_2$ não possuem o atributo, pode-se usar o método estatístico de máxima verossimilhança para estimar os parâmetros $(\beta_1, \beta_2, \ldots, \beta_p)$. Para cada elemento amostral $j$ tem-se $x = (x_{0j}, x_{1j}, \ldots, x_{kj})'$, sendo $x_{0j} = 1$ para todo $j = 1,2,3,\ldots,n$ .

Se $y_j = 1$ o elemento tem atributo e $y_j = 0$ o elemento não tem atributo Nessa situação a função de verossimilhança é dada por:

$$L(y_1, y_2, \ldots y_n) = \prod_{j=1}^{n} (\pi_j)^{y_j}.(1-\pi_j)^{1-y_j} \qquad (1.5)$$

sendo $\pi_j = P[Y_j = 1]$ e $1-\pi_j = P[Y_j = 0]$, *j=1,2,...,n.*

Usando o modelo logístico para $\pi_j$:

$$L(y_n) = \prod_{j=1}^{n}\left[\frac{e^{\beta_0+\beta' x_j}}{1+e^{\beta_0+\beta' x_j}}\right]^{y_j}\cdot\left[\frac{1}{1+e^{\beta_0+\beta' x_j}}\right]^{1-y_j}$$

$x_j = (x_{0j}, x_{1j,\ldots,} x_{pj})'$ $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$, sendo *p* o número de variáveis explicativas.

As estimativas correspondem à solução da derivada do logaritmo da função de verossimilhança em relação aos *p+1* parâmetros desconhecidos e podem ser obtidas através do algoritmo iterativo conhecido como método escore de Fisher (Casella,2002).

Um teste estatístico é feito para avaliar se o modelo é significativo ou não, a hipótese nula é a de que os parâmetros do modelo relativos as variáveis explicativas são todos iguais a zero e a hipótese alternativa é a de que nem todos os coeficientes são iguais a zero. A estatística de teste, chamada de função G dada em (1.6), compara o logaritmo da função de verossimilhança calculada nas estimativas de máxima verossimilhança dos parâmetros do modelo, com o logaritmo da função de verossimilhança considerando um modelo que tem apenas o intercepto (modelo nulo). Valores elevados da função G indicam que não se tem um modelo nulo, o que é bom. Sob a hipótese nula a estatística G tem aproximadamente uma distribuição qui-quadrado com *p* graus de liberdade, sendo *p* o número de variáveis explicativas do modelo ajustado. Assim, dado o nível de significância α, 0< α <1, do teste, a hipótese nula será rejeitada se o valor da estatística G for maior que o valor de referência $\chi^2_{1-\alpha,p}$ obtido através da distribuição qui-quadrado, valor a partir da qual se tem α% dos valores da distribuição. Um valor pequeno da probabilidade de significância indica rejeição da hipótese nula, o que é bom.

$$G = -2[\ln(L(\hat{\beta}_\alpha)) - \ln(L(\hat{\beta}^*))] \qquad (1.6)$$

sendo $\hat{\beta}_\alpha = (\hat{\beta}_0, \hat{\beta}_1, ...., \hat{\beta}_p)$ o vetor com as estimativas de máxima verossimilhança do modelo com as p variáveis explicativas e $\hat{\beta}^* = (\hat{\beta}_0)$ a estimativa de máxima verossimilhança de $\beta_0$ considerando um modelo no qual $\beta_k = 0$, $k = 1,2,...p.$ A estatística G é proveniente do teste de hipótese da razão de verossimilhança (Casella,2002).

## 2.3 Medidas de Qualidade de Ajuste do Modelo Logístico

### 2.3.1 Estatística *Deviance*

Existem várias medidas que podem ser usadas para quantificar a qualidade de ajuste do modelo logístico dentre elas a estatística chamada de *deviance* (Paula, 2013).

Para avaliar a qualidade do ajuste do modelo logístico utiliza-se a estatística do desvio *deviance*, que compara o modelo ajustado com dado modelo saturado. Na equação (1.7) $y = (y_1, y_2, ..., y_n)$ são os valores amostrais observados e *L(y)* é a função de verossimilhança calculada nesses valores. $L(\hat{\beta})$ é a função de verossimilhança calculada nas estimativas de máxima verossimilhança obtidas com o modelo de *p+1* parâmetros sendo $\hat{\beta}$ as estimativas dos parâmetros do modelo.

$$D\left(\hat{\beta}\right) = 2.\left[\ln\left(L(y)\right) - \ln\left(L(\hat{\beta})\right)\right] \qquad (1.7)$$

A hipótese nula é que o modelo está bem ajustado e a hipótese alternativa é a de que o modelo não está bem ajustado.

Para uma amostra de tamanho *n*, se o modelo de regressão logístico está bem ajustado, então a estatística *D deviance* tem distribuição qui-quadrado com *n-p* graus de liberdade. Valores pequenos de *deviance* ou valores elevados de probabilidade de significância do teste são desejáveis para que o modelo seja considerado bem ajustado.

A regra de decisão do teste qui-quadrado é:

$$\begin{cases} Se\ D\left(\hat{\beta}\right) \leq \chi^2_{\alpha,n-p},\ não\ rejeita\ H_0 \\ Se\ D\left(\hat{\beta}\right) > \chi^2_{\alpha,n-p}, rejeita\ H_0 \end{cases}$$

### 2.3.2 Estatística de *Pearson*

É uma medida que compara o número esperado de observações para cada nível das covariáveis, sob o modelo ajustado, com o valor de fato observado na amostra. A hipótese

nula é de que o modelo está bem ajustado (Paula, 2013), ou seja, que os valores observados para cada nível das covariáveis está próximo dos valores ajustados pelo modelo logístico.

Valores elevados dessa estatística, assim como valores pequenos de probabilidade significância, indicam um modelo inadequado.

### 2.3.3 Estatística *Hosmer-Lemeshow*

Os de dados são dividido em *k* conjuntos, de modo que cada conjunto tem um número de observações $(n_1, n_{2,} n_3, \ldots, n_k)$. Seja $(e_1, e_2, e_3, \ldots, e_k)$ o número esperado de elementos amostrais em cada conjunto calculados com base no modelo logístico ajustado.

Utiliza-se o teste qui-quadrado para comparar $n_j$ com $e_j$, *j=1,2,3,..., k*. Logo, valores elevados, da estatística *Hosmer-Lemeshow*, indicam um modelo não adequado (HOSMER, D.W. LEMESHOW, 1980).

## 2.4 Análise de Significância dos Parâmetros do Modelo

Os testes estatísticos para avaliar a significância dos coeficientes individuais dos parâmetros do modelo são realizados de forma análoga ao teste da distribuição *t-Student,* pois para grandes amostras, a distribuição do estimador de máxima verossimilhança do parâmetro respectivo é aproximadamente normal. Cada coeficiente é testado mediante hipótese nula de que $\beta_i = 0$, e hipótese alternativa de que $\beta_i \neq 0$, *i=0,1,2,...p*.

Por se tratar de um modelo discriminante, é importante, também, avaliar a qualidade do ajuste a partir da classificação e capacidade de discriminação. Logo, as medidas de sensibilidade e especificidade serão utilizadas como estimativa das probabilidades de classificações correta.

No caso de duas populações, para cada elemento da amostra, calcula-se o escore numérico da função logística ajustada, e assim, cada elemento é discriminado a partir de um valor de probabilidade estimado de pertinência para cada uma das populações.

Em Bioestatística a sensibilidade é definida como o percentual de indivíduos doentes (população 1), que são classificados como doentes, e a especificidade sendo o percentual de indivíduos não doentes (população 0), que são classificados como não doentes. Todavia, podemos utilizar a sensibilidade para avaliar o percentual de elementos da população 1 que são classificados pelo modelo logístico ajustado como indivíduos procedentes da população 1,

e a especificidade para avaliar o percentual de elementos da população 0 que são classificados como pertencentes à população 0.

Sejam $C_{11}$ - o número de indivíduos doentes classificados como doentes; $C_{10}$ - o número de indivíduos doentes classificados como não doentes; $C_{00}$ - o número de indivíduos não doentes classificados com não doentes e $C_{01}$ - o número de indivíduos não doentes classificados como doentes. A sensibilidade e a especificidade são definidas como (Mingoti, 2005):

:

$$\begin{cases} sensibilidade\ do\ teste: s = P(1|1) = \dfrac{C_{11}}{C_{11} + C_{10}} \\ especificidade\ do\ teste: e = P(0|0) = \dfrac{C_{00}}{C_{00} + C_{01}} \end{cases}$$

sendo $P(1|1)$ a probabilidade de que o indivíduo pertencente a população 1 (doentes), seja classificado corretamente como procedentes da população 1; $P(0|0)$ a probabilidade de que o indivíduo pertencente a população 0 ( não doentes) seja classificado corretamente como procedente da população 0.

Finalmente, a estimativa da probabilidade global de acertos indica a razão entre o número total de indivíduos classificados corretamente, em ambas as populações 1 e 0, e o número total de elementos, ou de classificações realizadas.

$$P(acerto) = \frac{C_{11} + C_{00}}{C_{11} + C_{10} + C_{00} + C_{01}}$$

## 3 Metodologia

Uma instituição de ensino superior, da cidade de Belo Horizonte, deseja desenvolver um modelo de avaliação sistêmica que possa simular o desempenho de seus alunos no ENADE, Exame Nacional de Desenvolvimento Estudantil, realizado pelo Ministério da Educação.

Para tanto, foi realizado uma avaliação piloto no final do 1° semestre de 2012, com alunos de apenas um curso, em três diferentes Campus. Juntamente com a prova, foi aplicado

um questionário experimental com 28 perguntas -O modelo do questionário está apresentado no Anexo A.

Basicamente, o questionário foi elaborado por sete diferentes grupos de perguntas: as quatro primeiras variáveis de ordem demográficas, com duas quantitativas e as outras duas qualitativas nominais; seis variáveis qualitativas nominais e uma qualitativa, relacionadas ao ensino médio; duas variáveis qualitativas ordinais, que avaliaram a disponibilidade de estudo extra-classe; uma variável quantitativa e três variáveis binárias, relacionadas a questão da empregabilidade e tipo de emprego; duas variáveis quantitativas e uma qualitativa, associadas a questão de transporte; quatro variáveis qualitativas nominais, relacionadas aos hábitos e lazer; três últimas qualitativas nominais, relacionadas a percepção e avaliação da prova.

Para utilização do modelo de regressão logístico, todas as variáveis qualitativas nominais e quantitativas foram codificadas como variáveis binárias. Finalmente, o banco de dados utilizado conteve 382 informações (alunos que prestaram a prova) de 26 variáveis binárias, 5 qualitativas ordinais e uma variável quantitativa.

A variável resposta foi definida com base no desempenho de cada aluno na prova. Como os modelos de provas aplicadas foram compostos por dez questões objetivas com cinco diferentes itens (a,b,c,d,e), havia a possibilidade do acerto eventual (quando aluno marca um item sem saber corretamente a resposta).

Neste experimento aleatório de responder cada item da prova, havia a possibilidade dos alunos acertarem cada uma das dez respostas da prova eventualmente, ou não, que é um experimento associado ao modelo de probabilidade Binomial com *(n=10)* tentativas independentes e com probabilidade *($\pi$=0,20)* de acerto eventual (evento).

O Gráfico 2 mostra a distribuição de probabilidade do número de acertos eventuais. Observa-se que a partir de $x = 4$ os acertos aleatórios começam a serem difíceis de ocorrer, pois apresentam probabilidades relativamente pequenas. O gráfico ainda mostra que, em termos de acertos aleatórios, o mais provável é a ocorrência de dois acertos *[P(2)=0,30]*, pois a esperança matemática do número de acertos eventuais é $E(X) = 10.(0,2) = 2$ acertos.

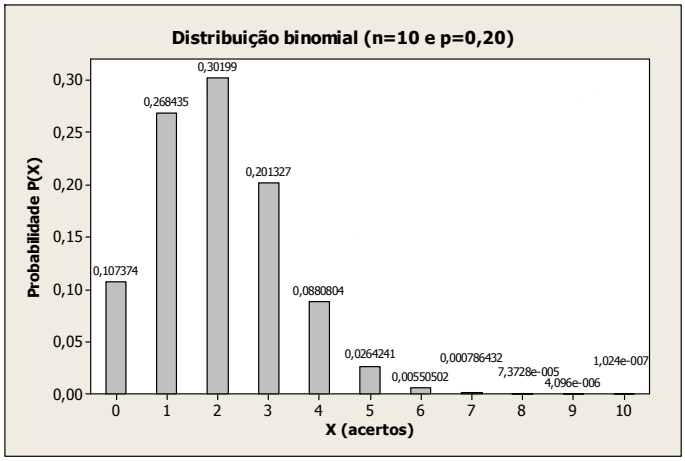


**Gráfico 2: Probabilidade de *x* acertos eventuais**

**Fonte: Elaborado com *Minitab 16***

O Gráfico 3 foi utilizado como base para definirmos o ponto de corte para discriminação das populações (0 e 1) iniciais. Logo, alunos com até 4 acertos na prova, foram classificados como pertencentes à população (0), pois sabiam até 4 questões ou acertaram, eventualmente, até 4 questões com uma probabilidade $\sum_{x=0}^{4} P(X = x) = 0{,}9672$.

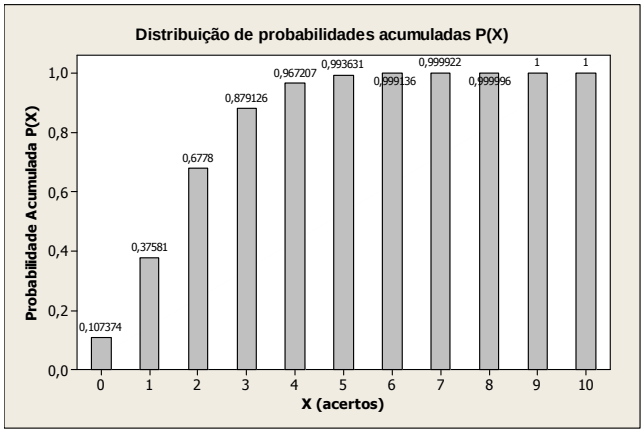


**Gráfico 3: Probabilidade de até *x* acertos eventuais ocorrerem**

**Fonte: Elaborado com *Minitab 16***

A população 1 foi então composta por alunos que acertaram 5 ou mais questões, onde a probabilidade de acerto eventual é $\sum_{x=5}^{10} P(X = x) = 1 - 0{,}9672 = 0{,}0328$, que indica um conjunto de alunos que sabia boa parte das questões da prova, em termos de probabilidade. Assim, a população 1 ficou constituída de 278 estudantes correspondente a 27% dos alunos analisados, enquanto a população 0 ficou 104 estudantes perfazendo um total de 382 indivíduos (73%).

## 4 Análise Exploratória de dados e Ajuste do Modelo Logístico

O Quadro 1 mostra o número de alunos por campus, que realizaram a prova, de acordo com a população em que foram designados pelo critério explicado na seção 3. São apresentados os percentuais de alunos por Campus e por populações

**Quadro 1: total de alunos em cada população por campus**

| Campus | População- 0 | População- 1 | total |
|---|---|---|---|
| A | 18(50%) | 18(50%) | 36(9,4%) |
| B | 32(26%) | 92(74%) | 124(32,5%) |
| C | 54(24%) | 168(76%) | 222(58,1%) |
| total | 104(27%) | 278(73%) | 382 |

**Fonte: Elaborado pelo autor**

Primeiramente foi ajustado o modelo de regressão logística com 32 variáveis explicativas, onde foram codificadas as variáveis qualitativas nominais e a quantitativas, a partir das 28 variáveis originais mais 4 variáveis adicionais: período, campus, turma e turno. No Quadro 2 são apresentados os critérios de codificação para cada variável.

Eliminando as variáveis não significativas ao nível de significância de 6,3%, o modelo foi novamente ajustado com 6 variáveis explicativas: (1) percepção da prova; (2) idade superior a 20 e até 30 anos; (3) trabalhou em mais de 2 diferentes empregos nos últimos anos; (4) prática esporte; (5) é aluno do campus A; (6) percorre os deslocamentos trabalho/escola e escola/casa com média de até 2 horas. Os resultados do ajuste estão no Quadro 3.

**Quadro 2: Descrição da codificação das variáveis**

| | Variável | Classificação inicial | Codificação (nova classificação) | Descrição |
|---|---|---|---|---|
| 1 | Sexo | qualitativa nominal | binária | 0-Masculino 1-Feminino |
| 2 | Idade | quantitativa | binária | 0-(x<=20 ou x>30) 1-(20<x<=30) x é idade em anos |
| 3 | Estado civil | qualitativa nominal | binária | 0-(solteiro,separado ,divorciado , união est.) 1-(casado) |
| 4 | Numero de filhos | quantitativa | binária | 0-(nenhum filho) 1-( um ou mais filhos) |
| 5 | Local de conclusão do ensino médio | qualitativa nominal | binária | 0-(outro estados, interior) 1- (capital) |
| 6 | Escola de conclusão do ensino médio | qualitativa nominal | binária | 0-(publica) 1-(particular) |
| 7 | Disciplina mais difícil | qualitativa nominal | binária | 0-(mat, port, fis, qui, hist, geo, NDA) 1-( lingua estrangeira) |
| 8 | Disciplina que gosta mais | qualitativa nominal | binária | 0-(mat, port, fis, qui, hist, geo, NDA) 1-( lingua estrangeira) |
| 9 | Dsciplina mais fácil | qualitativa nominal | binária | 0-(mat, port, fis, qui, hist, geo, NDA) 1-( lingua estrangeira) |
| 10 | Dsciplina mais importante | qualitativa nominal | binária | 0-(mat, port, fis, qui, hist, geo, NDA) 1-( lingua estrangeira) |
| 11 | Idade de conclusão do ensino médio | quantitativa | binária | 0-(até 18 anos) e 1-( mais de 18 anos) |
| 12 | Tempo de estudo extra classe nos dias de semana | qualitativa ordinal | qualitativa ordinal | qualitativa ordinal |
| 13 | Tempo de estudo extra classe nos finais de semana | qualitativa ordinal | qualitativa ordinal | qualitativa ordinal |
| 14 | Tipo de trabalho | qualitativa nominal | binária | 0-(não trabalha) 1-(trabalha) |
| 15 | Numero de empregos diferentes no últimos anos | quantitativa | binária | 0-( até dois empregos) 1-(mais de 2 empregos) |
| 16 | Trabalha na área do curso | qualitativa nominal | binária | 0-(não trabalha na área do curso) 1-(trabalha na área do curso) |
| 17 | Exerce cargo de Lider | qualitativa nominal | binária | 0-( não exerce cargo de lider) 1-(exerce) |
| 18 | Utiliza transporte público | qualitativa nominal | binária | 0-(não) 1-(sim) |
| 19 | Tipo de transporte | qualitativa nominal | binária | 0-(a pé, ônibus, metrô, carona) 1-(carro, moto) |
| 20 | Tempo de deslocamento trabalho/escola | quantitativa | binária | 0-(tempo< 2 horas) 1-(tempo >=2 horas) |
| 21 | Tempo de deslocamento escola/casa | quantitativa | binária | 0-(tempo< 2 horas) 1-(tempo >=2 horas) |
| 22 | Prática de esporte | qualitativa nominal | binária | 0-(não) 1-(sim) |
| 23 | Consumo de bebida alcoólica | qualitativa nominal | binária | 0-(não) 1-(sim) |
| 24 | Tabagismo | qualitativa nominal | binária | 0-(não fuma) 1-(fuma) |
| 25 | Tipo de lazer | qualitativa nominal | qualitativa nominal | qualitativa nominal |
| 26 | Grau de dificuldade da prova | qualitativa ordinal | qualitativa ordinal | qualitativa ordinal |
| 27 | Extensão da prova | qualitativa ordinal | qualitativa ordinal | qualitativa ordinal |
| 28 | Dificuldade em responder à prova | qualitativa ordinal | qualitativa ordinal | qualitativa ordinal |
| | | | | |
| | Demais variáveis | | | |
| 29 | Período | quantitativa | quantitativa discrita | 1,2,3,4 |
| 30 | Turma | qualitativa nominal | qualitativa nominal | A,B,C |
| 31 | Campus | qualitativa nominal | binária | 0-(não é aluno do campus A) 1-(aluno do campus A) |
| 32 | turno | qualitativa nominal | binária | 0-(vespetino e matutino) 1-(noturno) |

**Fonte: Elaborado pelo autor**

**Quadro 3: Modelo Logístico ajustado**

Ajuste de modelo de regressão Logístico

| Variáveis Preditoras | [2]Coef | [3]SE Coef | [4]P | Razão Chances | [1]IC (95%) Limite inferior | Limite superior |
|---|---|---|---|---|---|---|
| Constant | -0,412279 | 1,37011 | 0,763 | | | |
| P26 (Percepção da Prova) | -0,335827 | 0,180829 | 0,063 | 0,71 | 0,50 | 1,02 |
| (20<idade<=30) | 0,714096 | 0,247481 | 0,004 | 2,04 | 1,26 | 3,32 |
| mais de 2 empregos diferentes | -0,599765 | 0,263158 | 0,023 | 0,55 | 0,33 | 0,92 |
| Pratica esporte | -0,639975 | 0,342854 | 0,062 | 0,53 | 0,27 | 1,03 |
| É aluno do campus liberdade | -0,896048 | 0,373901 | 0,017 | 0,41 | 0,20 | 0,85 |
| tempo médio/deslocamento <=2h | 2,65003 | 1,26277 | 0,036 | 14,15 | 1,19 | 168,18 |

Log-verossimilhança = -207,446
Teste para verificar se os parâmetros são iguais a zero:G = 32,420,GL =6,P-Valor =0,000

Qualidade de ajuste

| Método | Qui-Quadrado | DF | P |
|---|---|---|---|
| Pearson | 50,9104 | 40 | 0,116 |
| Deviance | 62,5170 | 40 | 0,013 |
| Hosmer-Lemeshow | 4,8741 | 5 | 0,431 |

**Fonte: dados da pesquisa saída do *Minitab 16***

Analisando o modelo ajustado (Quadro 3), observa-se que todas as variáveis explicativas são significativas ao nível de significância de 6,3%, com exceção da constante.

[1] Intervalo de confiança ao nível de 95%
[2] Coeficientes do modelo logístico
[3] Erro padrão dos coeficientes
[4] Probabilidade de significância (p-valor)

Quanto a avaliação geral do modelo, na análise da função *log* de verossimilhança o *p-valor* foi próximo de zero, o que nos levaria rejeitar a hipótese de um modelo nulo. No entanto, quando avaliamos as estatísticas *qui-quadrado: Pearson, Deviance e Hosmer-Lemeshow*, nos quais a hipótese nula é de que o modelo está bem ajustado, rejeitaríamos a boa qualidade de ajuste do modelo ao nível de significância 5% se considerarmos apenas a estatística *Deviance* (0,013), pois as demais estatísticas apresentam *p-valor* superior a 0,05: *Pearson* (0,116) *e Hosmer-Lemeshow* (0,431).

Em termos de discriminação o modelo ajustado gerou estimativas elevadas para a probabilidade do estudante pertencer à população 1, logo, grande parte dos valores preditos apresentaram uma elevada probabilidade estimada de pertencer a esse grupo, o que dificultou a definição do nível de corte do valor da probabilidade para realizar a discriminação (classificação) dos alunos (ver Gráfico 4).

Assim, para contornar este problema foi realizada uma mudança no ponto de corte convencional de probabilidade (0,50) usado para classificação de elementos nas populações, isto é: o indivíduo com probabilidade de pertinência estimada abaixo de 0,50 seria classificado como da população 0; e aqueles com probabilidade de pertinência acima de 0,50 seria classificado como da população 1. Através de algumas tentativas de mudanças de ponto de corte, e análises do poder de discriminante resultante estabeleceu-se como ponto de corte final para a probabilidade o valor 0,74, sendo então utilizada a seguinte regra de classificação: os alunos com probabilidade de pertinência estimada pelo modelo logístico menor que 0,74 foram classificados como elementos da população 0 enquanto os com probabilidade estimada maior ou igual a 0,74 foram estimados como pertencentes à população 1. As estimativas das probabilidades de erros e acertos na classificação dos estudantes usando a regra estabelecida estão no Quadro 4. Como ilustração apresenta-se no Quadro 5 o percentual total de acertos e erros, nas classificações realizadas para alguns níveis de corte de probabilidade diferentes: 0,5;0,6 e 0,8. É importante observar que para todas os diferentes níveis de corte testados o valor de 0,74 apresentou maior índice de especificidade e sensibilidade, o que justificou a escolha deste valor.

**Quadro 4: Comparação entre a Classificação real e a Classificação resultante da aplicação do modelo logístico estimado**

| Classificação | Resultado da prova-0 | Resultado da prova-1 | total |
|---|---|---|---|
| **Estimado-0** | **73 (70,19%)** | **31(29,81)** | **104(100%)** |
| **Estimado-1** | **104(37,41%)** | **174(62,59%)** | **278(100%)** |
| **total** | **177** | **205** | |

**Fonte: Elaborado pelo autor com *Minitab 16***

Ao avaliarmos a *sensibilidade*, *especificidade* e o percentual geral de acertos (ver Quadro 4), nota-se que o modelo logístico ajustado e com ponto de corte igual a 0,74 para classificação apresenta melhor desempenho preditivo para alunos da população (0), pois a especificidade é de $e = P(0|0) = 0{,}7019$ (ou seja 70,19%) e sensibilidade $s = P(1|1) = 0{,}6223$ (ou seja 62,23). A porcentagem global de acertos registrou um valor de aproximadamente $65\%$.

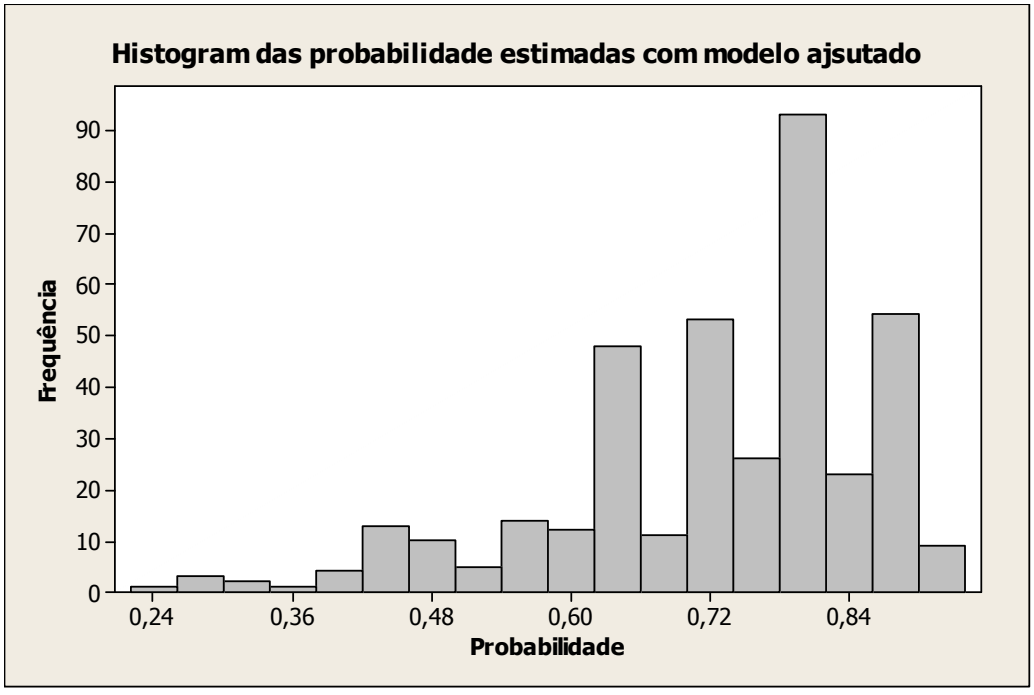


**Gráfico 4: Probabilidades estimadas para cada um do 382 elementos do banco de dados**

**Fonte: Elaborado pelo autor com *Minitab 16***

**Quadro 5: Comparação entre a Classificação real e a Classificação resultante da aplicação do modelo logístico estimado com três diferentes níveis de corte**

| Classificação nível 0,5 | Resultado da prova-0 | Resultado da prova-1 |
|---|---|---|
| **Estimativa-0** | **20(19,23%)** | **84(80,77%)** |
| **Estimativa-1** | **14(5,04%)** | **264(94,96%)** |
| **Classificação nível 0,6** | | |
| **Estimativa-0** | **28(26,92%)** | **76(73,08%)** |
| **Estimativa-1** | **27(9,71%)** | **251(90,29%)** |
| **Classificação nível 0,8** | | |
| **Estimativa-0** | **92(88,46%)** | **12(11,54%)** |
| **Estimativa-1** | **204(73,38%)** | **74(26,62%)** |

**Fonte: Elaborado pelo autor com *Minitab 16***

Ao observar os valores da *odds ratio* (razão das chances) no Quadro 3, é possível identificar quais fatores interferem na classificação dos elementos. Logo, indivíduos que têm "idade superior a 20 anos e até 30 anos" e apresentam "tempo médio de deslocamento de até 2 horas" têm razão de chances maior que 1, o que faz com que esses elementos tenham maior chance de serem estimados/classificados como elementos da população (1).

Por outro lado, os fatores: "ter mais de 2 empregos nos últimos anos", "praticar esporte" e "ser aluno do campus A" apresentam razão de chances menor que 1, o que reduz a chance de serem estimados/classificados como elementos da população (1). É importante observar que razão de chances para a variável "prática de esporte" é não significativa a 5% de significância, pois os limites inferior e superior da *odds ratio* são 0,27 e 1,03, respectivamente e incluem o valor 1, no entanto, destaca-se que o limite superior do intervalo de confiança é bom próximo de 1.

A variável "percepção da prova" é um caso a parte, pois é a única variável qualitativa ordinal com as seguintes opções: 1) muito fácil; 2) fácil; 3) regular; 4) difícil; 5) muito difícil. Como a razão das chances é menor que 1, há uma indicação de que quem teve uma percepção da prova como difícil, tem menor chance de ser estimado/classificado como elemento da população 1. Assim como a variável "prática de esporte", a "percepção da prova" também apresenta limites inferior e superior da *odds ratio* iguais a 0,50 e 1,02, o que indica razão das chances não significativa a nível 5% de significância. No entanto, como no caso anterior é importante ressaltar que o valor do limite superior de confiança é bem próximo de 1.

O Quadro 6 apresenta os resultados da análise de variância (Montgomery, 2003) realizada para comparação das médias das notas dos alunos na prova entre os diferentes níveis percepções da avaliação: 1-muito fácil; 2- fácil; 3-médio; 4- difícil e 5-muito difícil. Descritivamente observa-se que há uma diferença pontual das médias de cada nível, porém a diferença não é significativa ao nível de significância 5%, já que o *p-valor* 0,214 é maior que 0,05, levando a não rejeição da hipótese nula de igualdade de médias.

**Quadro 6: Análise de variância da nota da prova pela percepção da prova: 1) muito fácil; 2) fácil; 3) regular; 4) difícil; 5) muito difícil**

```
Análise de Variância    graus de
                        Liberdade   SQ    QM     F      P
P26 (Percepção da Prova)    4    15,60  3,90  1,46  0,214
Erro                      377  1007,69  2,67
Total                     381  1023,29

[5]S = 1,635   R-Sq = 1,52%   R-Sq(adj) = 0,48%


                          Intervalos Individuais de Confiança de
95%
                          Para Média, Baseados no des. Padrão
combinados
Nível    N  Média  D.p.   -------+---------+---------+---------+--
1       10  6,200  1,549             (------------*-----------)
2       16  6,125  1,586               (---------*---------)
3      247  5,543  1,620               (-*--)
4       95  5,368  1,695           (---*---)
5       14  5,071  1,592  (---------*----------)
                          -------+---------+---------+---------+--
                               4,80      5,60      6,40      7,20
[6]Desvio padrão combinados = 1,635
```

**Fonte: Elaborado pelo autor com *Minitab 16***

O Quadro 7 mostra o resultado do teste *t* de *student* para igualdade das médias, em relação ao fator idade e, concluímos, que os indivíduos com idade superior a 20 e até 30 anos tem média (na prova do teste simulado) superior aos demais alunos sendo 0- alunos que não tem idade superior a 20 e até 30 anos; 1- alunos que tem idade superior a 20 e até 30 anos.

[5] Desvio padrão combinados
[6] O desvio padrão combinado é a raiz quadrada do Quadrado médio residual

**Quadro 7: Teste *t-Student* para desigualdade das médias provão pela variável binária idade**

```
Teste t-Student para comparação de médias da variável: Prova do teste Simulado por
idade.

(20<idade<=30)    n  Média  Desvio padrão
Não 0           128  5,20   1,72
Sim 1           254  5,69   1,58


t-Test da diferença da médias = 0 (vs <): T-Valor = -2,66 P-Valor = 0,004  GL = 236
```

**Fonte: Elaborado pelo autor com *Minitab 16***

Em relação a variável "ter mais de 2 empregos", a hipótese nula de igualdade de médias das notas na prova dos estudantes que não tiveram mais de dois empregos diferentes nos últimos anos (codificada como 0) em relação aos que tiveram mais de dois empregos diferentes nos últimos anos (codificada como 1), não foi rejeitada (*p-valor=0,26*-Quadro 8)

**Quadro 8: Teste *t-Student* para diferença das médias provão pela variável binária emprego**

```
Teste t-Student para comparação de médias da variável: Prova do teste Simulado por
mais de 2 empregos diferentes ou não.

mais de 2 empregos diferentes            n  Média  Desvio Padrão

0(não teve mais de dois empregos)      146  5,64   1,64
1(teve mais de dois empregos)          236  5,45   1,64


t-Teste da diferença da médias= 0(vs not =):T-Valor = 1,13 P-Valor = 0,260  GL =306
```

**Fonte: Elaborado pelo autor com *Minitab 16***

Em relação a variável "prática de esporte", o Quadro 9mostra que as médias das notas dos indivíduos que praticam esporte é menor do que os não praticantes, ao nível de significância de 5%. (0- alunos que não pratica esporte; 1- alunos praticantes de esportes)

**Quadro 9: Teste *t-Student* para desigualdade das médias provão pela variável binária prática de esporte**

```
Teste t-Student para comparação de médias da variável: Prova do teste Simulado por
prática de esporte.

Prática de esporte                n  Média  Desvio Padrão

0 (não pratica esporte)       334  5,62   1,62
1 (pratica esporte)            48  4,88   1,67


T-Teste da diferença da média = 0 (vs >): T-Valor = 2,89  P-Valor = 0,003  GL = 60
```

**Fonte: Elaborado pelo autor com *Minitab 16***

Os dados do Quadro 10, mostram que há evidência estatística para afirmar que a média das notas da prova dos alunos do Campus A é inferior dos demais Campus, pois o *p-valor* =0,003 é inferior 0,05. (0- não é aluno do campus A; 1- é aluno do campus A)

**Quadro 10: Teste *t-Student* para desigualdade das médias provão pela variável binária campus.**

Teste t-Student para comparação de médias da variável: Prova do teste Simulado por campus.

| É aluno do Campus liberdade | n | Média | Desvio Padrão |
|---|---|---|---|
| 0 (não é aluno do campus Liberdade) | 346 | 5,59 | 1,64 |
| 1 (é aluno do campus Liberdade) | 36 | 4,86 | 1,46 |

T-Teste da diferença da média = 0 (vs >): T-Valor = 2,83 P-Valor = 0,003 GL = 44

**Fonte: Elaborado pelo autor com *Minitab 16***

O Quadro 11 os resultados do teste t-Student que indicam a não rejeição da hipótese nula de igualdade entre as médias das notas dos alunos cujo tempo médio de deslocamento para irem a escola é de até duas horas (codificado como 0) e aqueles cujos tempos médios de deslocamento é superior a duas horas (codificado como 1). É importante observar que há uma grande diferença entre o tamanho das amostras, o que prejudica, de certa forma, o poder do teste estatístico.

**Quadro 11: Teste *t-Student* para desigualdade das médias provão pela variável binária tempo de deslocamento médio de até 2horas.**

Teste t-Student para comparação de médias da variável: Prova do teste Simulado por tempo de deslocamento.

| tempo médio/deslocamento <=2h | n | Média | Desvio Padrão |
|---|---|---|---|
| 0 (media de deslocamento maior 2 h) | 3 | 4,00 | 1,00 |
| 1 (media de deslocamento menor ou igual 2h) | 379 | 5,54 | 1,64 |

T-Teste da diferença da média = 0 (vs <): T-Valor = -2,63 P-Valor = 0,060 GL = 2

**Fonte: Elaborado pelo autor com *Minitab 16***

Diante da suspeita de que o "tempo médio de deslocamento" interfere no desempenho dos alunos na prova, no Quadro 12 mostra-se os resultados de outro teste de hipótese considerando os indivíduos que se deslocam com tempo médio de até 1,5 hora (0) e acima de 1,5 hora (1). Diante desta nova perspectiva, constatou-se a rejeição da hipótese nula de igualdade de médias entre os dois grupos, ao nível de significância de 0,05, pois no Quadro 12

o *p-valor* é de 0,002. No entanto, mesmo com a rejeição da hipótese nula, é possível observar que ainda há uma grande diferença entre os tamanhos das amostras, o que compromete o poder do teste estatístico.

**Quadro 12: Teste *t-Student* para desigualdade das médias provão pela variável binária tempo de deslocamento médio de até 1,5 hora.**

```
Teste t-Student comparação de médias da variável: Prova do teste Simulado por tempo
de deslocamento.

Tempo de deslocamento <=1,5                  N   Média  Desvio Padrão
0 (media de deslocamento maior 2 h)          13  4,615  0,961
1 (media de deslocamento menor ou igual 2h) 369   5,56   1,65


T-Teste da diferença da Média = 0 (vs <): T-Valor = -3,36  P-Valor = 0,002  GL = 14
```

**Fonte: Elaborado pelo autor com *Minitab 16***

O Gráfico 6 mostra o *Box-Plot* das notas dos alunos na prova estratificadas pelos grupos: 0- tempo médio de deslocamento superior a 1,5 hora e 1- tempo médio de deslocamento de até 2 horas. Nesse gráfico apresenta-se a linha da média de cada categoria, e é possível observar que há uma tendência crescente entre as médias dos grupos, logo, sendo que os alunos que apresentam tempo médio de deslocamento inferior a 1,5 hora têm nota média superior na prova em relação àqueles com tempo médio de deslocamento superior a 1,5 horas. É importante observar que há uma grande diferença entre os tamanhos das amostras, o que prejudica, de certa forma, a extensão dos resultados do *Box-Plot*.

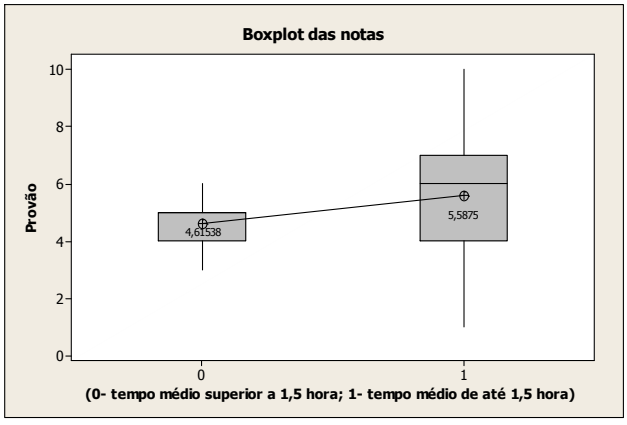


**Gráfico 6: *Boxplot* das notas da prova, com linha de médias**

**Fonte: Elaborado pelo autor com *Minitab 16***

## 5 Considerações Finais

Em virtude dos dados e estatística levantados, e apesar do modelo discriminante de regressão logística validado apresentar um percentual global de acertos relativamente baixo de 70%, foi possível identificar fatores que interferem no desempenho dos alunos que participaram da prova.

A dificuldade em estimar um modelo que tivesse um desempenho melhor em termos de classificação dos alunos, pode ser explicada por dificuldades metodológicas no processo de levantamento dos dados, na parte inicial da pesquisa. Nesta etapa, foram identificados alguns problemas nas respostas dos questionários como: questionários incompletos; ou mal respondidos; ou com erro de identificações. Na análise apresentada nesse trabalho utilizou-se apenas os 382 registros que estavam com informações completas em todas as 32 variáveis mencionadas. Vale lembrar que a aplicação dos questionários foi feita para um conjunto de 446 estudantes.

Outro grande problema constatado foi a dificuldade em digitar as respostas dos quase 450 questionários, o que consumiu muito tempo, pois o próprio pesquisador (autor desse artigo), realizou toda esta tarefa individualmente. Uma solução para esse problema seria ter utilizado um questionário com leitor óptico, o que requer recurso extra. Outra opção seria ter utilizado o modelo de questionário eletrônico via *Google Doc's*, mas essa possibilidade seria inviável, pois queríamos registrar as informações dentro da sala de aula no dia da prova.

Quanto ao questionário, foram observados alguns problemas que serão ajustados para as próximas coletas de dados que serão efetuadas nessa instituição de ensino, tais como: inclusão de algumas outras variáveis como a questão do tempo médio de sono, ou que os alunos dormem à noite, variáveis associadas ao grau de escolaridade dos pais; ajuste de opções de resposta, como na variável meio de transporte e principalmente realizar testes pilotos, com novo questionário, antes do levantamento de dados. Uma alternativa para o teste desses questionários seria aplicá-los aos demais alunos do Campus, sem que estes sejam alunos que vão realizar a prova.

Em termos do desempenho do modelo logístico validado, observa-se que o mesmo é mais eficiente para detectar o aluno que tem um desempenho inadequado, pois se constatou uma especificidade de 70,2% contra sensibilidade de 62,6%.

Das seis variáveis mantidas no modelo, duas delas: "percepção da prova" e "prática de esportes", apresentaram razão de chances (*odds ratio)* não significativas a 5% de significância, porém o *p-valor* de ambas as variáveis está próximo desse valor 6,3% e 6,2%

respectivamente. Por outro lado o teste *t-Student* mostrou o fato de que ser praticante de esporte interfere no valor da média das notas dos alunos. No enatnto, ambas foram mantidas no modelo, pois apresentaram p-valor próximo de 5%, e a ausência das mesmas reduziu a especificidade, e também, prejudicou a qualidade de ajuste já que as estatísticas *Deviance, Pearson Hosmer-Lemeshow* tiveram seus valores reduzidos. É importante ressaltar que as populações 0 e 1 foram constituídas a partir de um critério fundamentado no comportamento do número de acertos considerando acertos aleatórios ( ver seção 3). Neste artigo apresenta resultados considerando o valor de 4 acertos eventuais como ponto de corte para construir as duas populações. No entanto, outros estudos foram realizados considerando mudanças nesse ponto de corte e os modelos logísticos resultantes foram insatisfatórios para efeito de discriminação.

Quanto a variável percepção da prova, na análise de variância a hipótese de igualdade de médias das notas dos alunos entre os níveis de percepção não foi rejeitada. Para as demais quatro variáveis: “idade”, “emprego”, “ser aluno do campus A, “tempo de deslocamento”, a diferença entre as médias das notas dos grupos respectivos foi significativa ao nível de significância de 5%.

Do ponto de vista da Instituição de Ensino se o propósito for o de estimar ( ou prever) a população na qual o aluno será classificado antes da realização efetiva da prova, o melhor é ter um modelo logístico sem a variável “percepção da prova” já que essa somente é conhecida após a realização da mesma. No entanto, quando o modelo logístico foi ajustado sem a variável “percepção da prova” há uma variação na qualidade de ajuste com respectivos valores, *Deviance, Pearson, Hosmer-Lemeshow*: 8,2%; 2,9%; 26%, mas a principal mudança é que a especificidade de 70,2% reduz para 53,9% e sensibilidade de 62,6% aumenta para 75,5% e percentual global de acerto continua menor que 70% sendo igual a 69,6%. É importante destacar que a ausência da variável “percepção da prova” torna o modelo logístico menos eficiente para detectar alunos que tendem a um desempenho baixo, e sendo mais eficiente para identificar alunos com bom desempenho na prova.

Finalmente, apesar de alguns contrapontos mencionados, e conservando o modelo logístico com as seis variáveis, foi possível identificar os alunos que tendem a apresentar um desempenho baixo na avaliação, como sendo aqueles alunos que conjugaram as seguintes características: não ter idade superior a 20 anos e até 30 anos; ser aluno do campus A; deslocar-se nos trajetos trabalho-escola e escola-casa com tempo médio superior a duas horas; ter mais de 2 empregos diferentes nos últimos anos; ser praticante de esportes; ter uma

percepção muito difícil da prova. É importante destacar que há uma dificuldade de robustez do modelo, pois dependendo do ponto de corte escolhido para a construção da regra de classificação dos alunos, pode-se encontrar valores preditivos diferentes. Os resultados apresentados neste artigo são exploratórios e servirão de base para as novas coletas de dados que serão realizadas pela Instituição de Ensino já eu a intenção e dar continuidade a esse projeto.

**Referências:**

BARTOLOTTI, S. V; MOREIRA Jr, F; SOUSA Jr, A; ANDRADE, D. F. *Teoria da resposta ao item- Medida de Satisfação por meio do modelo Logístico de dois parâmetros*. 2003 Disponívelem<:http://www.ime.unicamp.br/sinape/sites/default/files/Artigo_SINAPE_MEDIDA%20DE%20SATISFA%C3%87AO_TRI.pdf>. Acesso em: 27 ago. 2013

CASELLA, G; BERGER, R. I Statistical inference. California: Duxbury Thmpson Learning, 2002. 660p.

HOSMER, D.W. LEMESHOW S. "*A goodness-of-fit test for the multiple logistic regression model*." (1980) Communications in Statistics A10:1043-1069

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.* 1° edição. Belo Horizonte: Editora UFMG, 2005. 297 p.

MONTGOMERY, D. C; RUNGER, G. C. *Estatística Aplicada e Probabilidade para Engenheiros.* 2° edição. Rio de Janeiro: LCT, 2003. 463 p.

PAULA, G. A. *MODELOS DE REGRESSÃO: com apoio computacional.* Apostila/ notas de Aula, 2013. Disponível em : < http://www.ime.usp.br/~giapaula/texto_2013.pdf > . Acesso em: 09 mai. 2013.

P De Jong, GZ Heller, *Generalized linear models for insurance data*. The Edinburgh Building, Cambridge: 2008. 196p.

VIANNA, H. M. Avaliação Nacional em larga escala: análise e proposta. *Estudo em Avaliação Educacional*, n. 27, jan-jun/2003. Disponível em: < http://educa.fcc.org.br/pdf/eae/n27/n27a02.pdf>. Acesso em: 09 mai. 2012.

# ANEXO A

Nome: RA: data:
Curso: Período: Campus:

1. Sexo:
1) Masculino 2) Feminino 3)_____

2. Idade:

3. Estado civil:
1) Solteiro(a) 2) Separado(a) 3) Casado(a) 4) Divorciado(a) 5)Viúvo(a) 6) União Estável

4. Número de Filhos:

5. Onde você concluiu o ensino médio?
1) Capital 2) Interior 3) Outro estado

6. Em que tipo de escola você concluiu o ensino médio?
1)Pública 2) Particular

Com relação às disciplinas do ensino médio.
7. Qual disciplina você acha mais difícil?
1) Mat. 2) Port. 3) Fís. 4) Quí. 5) Hist. 6) Geo. 7) L. Estrangeira. 8)

8. Qual disciplina você gosta mais?
1) Mat. 2) Port. 3) Fís. 4) Quí. 5) Hist. 6) Geo. 7) L. Estrangeira. 8)

9. Qual disciplina você acha mais fácil?
1) Mat. 2) Port. 3) Fís. 4) Quí. 5) Hist. 6) Geo. 7) L. Estrangeira. 8)

10. Qual disciplina você acha mais importante?
1) Mat. 2) Port. 3) Fís. 4) Quí. 5) Hist. 6) Geo. 7) L. Estrangeira. 8)

11. Qual era sua idade quando concluiu o ensino médio?

12. Durantes os dias de semana, quantas horas você dedica para estudos extra classe?
1) Não consigo estudar durante os dias da semana.
2) Menos de 60 minutos.
3) Entre 60 a 120 minutos.
4) Mais de 120 minutos.
5)

13. Durantes os finais de semana, quantas horas você dedica para estudos extra classe?
1) Não consigo estudar durante os finais da semana.
2) Menos de 60 minutos.
3) Entre 60 a 120 minutos.
4) Mais de 120 minutos.
5)

14. Atualmente realiza algum tipo de trabalho remunerado?
1) Não
2) Sim, estágio.
3) Sim, trabalho com carteira assinada.
4) Sim, autônomo.
5)

15. Quantos trabalhos/empregos diferentes você teve nos últimos dez anos?

16. Trabalha na área do seu curso?
1) Sim.
2) Não.

17. No seu trabalho, exerce cargo de líder?
1) Sim.
2) Não.

18. Você utiliza transporte público para ir à escola?
1) Sim.
2) Não.

19. Que tipo de condução (ou meio de transporte) você utiliza para ir à escola?
1) A pé.
2) Ônibus.
3) Metrô.
4) Moto.
5) Carro.
6) Transporte escolar.
7)

20. Em condições normais, quanto tempo você gasta no deslocamento do seu trabalho (ou casa) para a escola?

21. Em condições normais, quanto tempo você gasta no deslocamento da sua escola para casa (ou trabalho)?

22. Pratica algum tipo de esporte com freqüência mínima de 3 vezes por semana?
1) Sim.
2) Não.
3)

23. Ao menos uma vez por semana, você consome bebidas alcoólicas?
1) Sim.
2) Não.
3)

24. Você é fumante?
1) Sim.
2) Não.

25. Em seus momentos de lazer, qual é sua opção favorita de entretenimento?
1) Praticar esportes.
2) Assistir programas de tv.
3) Assistir partidas de futebol.
4) Leituras.
5) Fazer compras.
6) Cinema.
7) Acessar a Internet.
8)

Com relação à prova realizada.

26. Qual grau de dificuldade desta prova?
1) Muito fácil.
2) Fácil.
3) Médio.
4) Difícil.
5) Muito difícil.

27. Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi:
1) muito longa.
2) longa.
3) adequada.
4) curta.
5) muito curta.

28. Você deparou com alguma dificuldade ao responder à prova. Qual?
1) Desconhecimento de conteúdo.
2) Forma diferente de abordagem do conteúdo.
3) Espaço insuficiente para responder às questões.
4) Falta de motivação para fazer a prova.
5) Não tive dificuldade para responder à prova.