Universidade de Brasília
IE – Departamento de Estatística
Estágio Supervisionado 2

AVALIAÇÃO DO RENDIMENTO DOS ALUNOS EM DISCIPLINAS OFERTADAS PELO DEPARTAMENTO DE ESTATÍSTICA PARA OUTROS CURSOS DA UNIVERSIDADE DE BRASÍLIA:
UMA APLICAÇÃO DE REGRESSÃO LOGÍSTICA MULTINÍVEL

Amanda Pereira Ferraz

Relatório Final do Projeto Final

Orientadora: Profª Maria Teresa Leão Costa

Brasília
Dezembro de 2013

Sumário

Lista de Ilustrações

Resumo

O desempenho dos estudantes nas universidades é uma preocupação atual tanto dos professores quanto dos próprios alunos, pois estes estão muito apreensivos com o ingresso no mercado de trabalho, visto que, nos dias de hoje, está cada vez mais difícil conseguir uma posição, pois além da concorrência, as empresas buscam, principalmente, os profissionais que se destacam e que são mais capacitados e qualificados.

Este trabalho visa analisar o desempenho dos estudantes da Universidade de Brasília que cursaram as disciplinas de serviço do Departamento de Estatística a partir da metodologia de Regressão Multinível, pois com ela é possível captar aspectos que influenciam os estudantes nos mais diversos níveis, sendo possível, dessa forma, avaliar tanto os fatores relacionados ao próprio aluno quanto fatores ligados à turma a que esse estudante pertence.

Com este trabalho, foi possível pontuar os aspectos que se mostram mais relevantes quando se deseja verificar os fatores que mais impactam o desempenho dos estudantes.

Palavras-chave: Regressão Multinível, Regressão Logística, Avaliação Educacional.

Abstract

The performance of students in universities is a present concern of both teachers and the students themselves, as they are more and more concerned about joining the labor market, since, these days, it is increasingly difficult to get a position because, in addition to the competition, the companies increasingly seek, professionals who stand out and end up being more capable and qualified.

This paper aims to analyze the performance of students from the University of Brasilia, who attended the disciplines of the Department of Statistics, from the Multilevel Regression Modeling, because, with it, it is possible to capture aspects that influence students in various levels and can thus analyze both factors related to the student himself as factors related to the class to which that student belongs.

This paper points the aspects shown as the more relevant when you want to check the factors which have more impact on student performance.

Keywords: Multilevel Regression, Logistic Regression, Educational Evaluation.

Introdução e Justificativa

O desempenho dos estudantes nas universidades é uma preocupação atual tanto dos professores quanto dos próprios alunos, pois estes estão muito apreensivos com o ingresso no mercado de trabalho, visto que, nos dias de hoje, está cada vez mais difícil conseguir uma posição, pois além da concorrência, as empresas buscam, principalmente, os profissionais que se destacam e que são mais capacitados e qualificados.

Atualmente, o Departamento de Estatística da Universidade de Brasília oferta três disciplinas para atender os estudantes dos mais variados cursos, são elas: Bioestatística, disciplina obrigatória para os alunos de Agronomia, Engenharia Florestal e Medicina Veterinária; Estatística Aplicada, que é disciplina obrigatória dos cursos de Administração, Arquivologia, Biblioteconomia, Ciência Política, Ciências Ambientais, Ciências Contábeis, Ciências Sociais, Geografia, Gestão de Agronegócios, Psicologia e Relações Internacionais; e Probabilidade e Estatística, que consta da grade curricular dos cursos de Ciência da Computação, Ciências Econômicas, Computação, Engenharia Civil, Engenharia de Computação, Engenharia Elétrica, Engenharia Mecânica, Engenharia Mecatrônica, Engenharia de Produção, Engenharia de Redes de Comunicação e Matemática.

Este trabalho visa analisar o desempenho dos estudantes da Universidade de Brasília que cursaram as disciplinas de serviço do Departamento de Estatística a partir da metodologia de Regressão Multinível. Com esta metodologia é possível captar os aspectos que estão associados ao rendimento dos alunos nos mais diversos níveis, sendo possível, dessa

forma, analisar tanto os fatores relacionados ao próprio estudante quanto fatores ligados à turma a que esse aluno pertence.

A motivação inicial para a escolha do tema foi a possibilidade de estudar a técnica de Regressão Multinível fazendo uma aplicação em um conjunto de dados real. Outro aspecto que foi levado em consideração foi a oportunidade de utilizar os resultados deste trabalho para avaliar o papel do Departamento de Estatística na formação dos estudantes de outras áreas. A análise feita pode ajudar o Departamento como um todo, pois os professores poderão repensar a forma como a disciplina é ministrada e também pode-se refletir sobre a elaboração de políticas que venham a contribuir de forma significativa para que os estudantes tenham um desempenho satisfatório nessas disciplinas.

Objetivos

Visando a compreensão do desempenho dos alunos da Universidade de Brasília que cursaram as disciplinas de serviço (Bioestatística, Estatística Aplicada e Probabilidade e Estatística) ofertadas pelo Departamento de Estatística, tem-se como finalidade definir a partir da aprovação dos alunos, os principais aspectos responsáveis pelas eventuais diferenças entre os desempenhos dos mesmos.

Os dois objetivos principais deste trabalho são: estudar a técnica de Análise Multinível e fazer uma aplicação da mesma, identificando as características do estudante e da turma/professor que estão associadas ao desempenho dos alunos nas disciplinas ofertadas pelo Departamento de Estatística para os mais diversos cursos da Universidade de Brasília.

Os objetivos específicos consistem em:

- ✓ Definir os principais fatores, tanto no nível dos estudantes quanto no das turmas, que são responsáveis pelas eventuais diferenças entre o rendimento dos alunos;
- ✓ Descrever um modelo que considere os aspectos relevantes para definir o desempenho dos estudantes em cada uma das três disciplinas de serviço. Uma vez que as disciplinas de serviço são separadas por áreas de conhecimento (Exatas, Humanas e Agrárias/Biológicas/Saúde), deve-se levar em consideração que, provavelmente, diferentes fatores serão responsáveis pela aprovação

dos alunos em cada um dos grupos, ou seja, em cada uma das disciplinas.

Referencial Teórico

Regressão Logística

O modelo de regressão logística e não linear binário é usado quando a variável resposta é qualitativa com duas possibilidades de resposta. Dessa forma, a variável resposta de interesse é representada por uma variável indicadora binária ou dicotômica, que assume os valores 0 ou 1.

Considerando o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad Y_i = 0, 1 \tag{1}$$

Quando a resposta Y_i é binária, assumindo valores 0 ou 1, a resposta esperada $E(Y_i)$ tem um significado especial. Assim, desde que $E(\varepsilon_i) = 0$, tem-se que:

$$Y_i = \beta_0 + \beta_1 X_i \tag{2}$$

Assim, a resposta média quando a variável é indicadora, sempre representa a probabilidade de $Y_i = 1$ para os níveis das variáveis preditoras.

Deve-se considerar que utilizar o modelo de regressão simples quando a variável resposta é binária, pode acarretar alguns problemas, tais como:

1. Não normalidade dos erros, logo um modelo que assume que os erros são normalmente distribuídos não é apropriado.

2. Variâncias heterogêneas, isso ocorre porque a variável resposta é indicadora, portanto, Variância(Y_i) = $\pi_i \times (1 - \pi_i)$.

3. Restrição na função resposta, como a resposta média representa a probabilidade de $Y_i = 1$, tem-se que por ser uma probabilidade esse valor deve ser maior ou igual a 0 e menor ou igual a 1. Por essa razão, uma função linear não é adequada para esse tipo de modelagem.

Assim, no caso da modelagem de uma variável resposta binária, deve-se considerar funções que são limitadas entre 0 e 1 e que tem uma característica curvilínea com o formato sigmoidal, ou seja, a função resposta tem a forma de S e tem assíntotas nos valores 0 e 1.

Logo, uma alternativa para a distribuição dos erros que é similar à forma da função de distribuição normal é a distribuição logística.

Regressão Logística Simples

Considerando que quando a variável resposta é binária e assume valores 1 e 0 com probabilidades π_i e $1 - \pi_i$ respectivamente, e que Y_i é uma variável aleatória com distribuição de Bernoulli com $E(Y_i) = \pi_i$. O modelo de regressão logística simples é definido da seguinte forma:

$$Y_i = E(Y_i) + \varepsilon_i \qquad (3)$$

Uma vez que a distribuição dos erros depende da variável resposta Y_i que tem distribuição Bernoulli, a função logística é apresentada no seguinte formato:

$$E(Y_i) = \pi(X_i) = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} \quad (4)$$

ou de forma equivalente:

$$E(Y_i) = [1 + \exp(-\beta_0-\beta_1 X_i)]^{-1} \quad (5)$$

A partir deste ponto, quando for utilizada a notação π_i, deve-se perceber que ela, na verdade, depende de X_i, ou seja, é uma função de X_i.

Uma propriedade que torna viável o uso da função logística é que ela pode ser linearizada. Assim, como a variável resposta é binária e a resposta média representa a probabilidade, $E(Y_i) = \pi_i$, tem-se que fazendo a transformação a seguir:

$$\pi_i' = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \quad (6)$$

obtém-se:

$$\pi_i' = \beta_0 + \beta_1 X_i \quad (7)$$

Essa função é chamada de transformação logito da probabilidade π_i. A razão $\pi_i/1-\pi_i$, que aparece na transformação logito é chamada de odds

(chance). A função resposta transformada é denominada função resposta logito e π_i' é a resposta logito média. É importante observar que na função resposta logito tanto π_i' quanto X_i estão definidos no intervalo de -∞ a ∞.

Na regressão logística simples, as estimativas dos parâmetros β_0 *e* β_1 são obtidas por meio da maximização do logaritmo da função de verossimilhança.

Considerando que $P(Y_i = 1) = \pi_i$ e $P(Y_i = 0) = 1 - \pi_i$ e que a distribuição de probabilidade de Bernoulli é dada por:

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1 - Y_i}, \qquad Y_i = 0,1;\ i = 1, 2, ..., n \tag{8}$$

Como as observações Y_i são independentes, a densidade conjunta é dada por:

$$g(Y_1, ..., Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1 - Y_i} \tag{9}$$

Aplicando o logaritmo neperiano à função de densidade conjunta, obtém-se a seguinte equação:

$$\ln g(Y_1, ..., Y_n) = \sum_{i=1}^{n}[Y_i \ln(\pi_i / 1 - \pi_i)] + \sum_{i=1}^{n} ln(1 - \pi_i) \tag{10}$$

Sabendo-se que $E(Y_i) = \pi_i$ para uma variável binária e considerando a expressão (4), tem-se que:

$$1 - \pi_i = (1 + \exp(\beta_0 + \beta_1 X_i))^{-1} \tag{11}$$

A partir das expressões definidas em (6) e (7), a função de verossimilhança é dada por:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i (\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} ln (1 + \exp (\beta_0 + \beta_1 X_i)) \quad (12)$$

Os estimadores de máxima verossimilhança $\beta_0 \, e \, \beta_1$ não podem ser encontrados de forma analítica, por isso é necessário utilizar métodos numéricos para encontrar as estimativas de máxima verossimilhança b_0 e b_1.

Encontradas as estimativas b_0 e b_1, deve-se substituir os valores na equação (4) com o objetivo de encontrar os valores ajustados. Assim, tem-se que o valor ajustado para o i-ésimo valor é dado por:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)} \quad (13)$$

Logo, a função resposta ajustada é dada por:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)} \quad (14)$$

E, usando a transformação logito, a função resposta ajustada é dada por:

$$\hat{\pi}' = b_0 + b_1 X \quad (15)$$

em que

$$\hat{\pi}' = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \quad (16)$$

Considerando o valor da função resposta ajustada na expressão (15) em $X = X_j$ e em $X = X_j + 1$, tem-se que:

$$\hat{\pi}'(X_j) = b_0 + b_1 X_j \quad (17)$$

$$\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1) \quad (18)$$

A diferença entre os dois valores é dada por:

$$\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1 \quad (19)$$

De acordo com (16), tem-se que a expressão (17) é o logaritmo da chance (odds) estimada quando $X = X_j$, e este é denominado $\ln(\text{chance}_1)$. De forma semelhante, a expressão (18) é o logaritmo da chance estimada quando $X = X_j + 1$, denominado $\ln(\text{chance}_2)$. Assim, a diferença entre os valores ajustados é dada por:

$$\ln(\text{chance}_2) - \ln(\text{chance}_1) = \ln(\text{chance}_2/\text{chance}_1) = b_1 \quad (20)$$

Aplicando o anti-logaritmo em cada lado da equação, é possível perceber que a razão das chances estimadas, denominada razão das chances (odds ratio), é dada por:

$$\widehat{OR} = \frac{chance_2}{chance_1} = \exp(b_1) \quad (21)$$

Assim, tem-se que a razão das chances estimada quando existe uma diferença de c unidades em X é igual a $\exp(cb_1)$, essa é a interpretação quando a variável explicativa é quantitativa. Quando a variável explicativa é qualitativa, a interpretação da razão de chances revela quantas vezes é mais provável ocorrer sucesso considerando um determinado grupo em relação a outro.

Regressão Logística Múltipla

No modelo de regressão logística múltipla, tem-se que a variável resposta binária é explicada por p-1 variáveis explicativas segundo a seguinte expressão:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon_i \quad (22)$$

A função definida pela expressão (4) pode ser generalizada como:

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})}{1+\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})} \quad (23)$$

De forma equivalente, tem-se que a expressão acima pode ser escrita como:

$$E(Y_i) = (1 + \exp(-(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})))^{-1} \quad (24)$$

A transformação logito definida na expressão (7) resulta em:

$$\pi_i' = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} \quad (25)$$

Assim, a formulação do modelo para as variáveis aleatórias Y_i independentemente distribuídas segundo a distribuição de Bernoulli com valores esperados $E(Y_i) = \pi_i$ é dada por:

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})}{1+\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})} \quad (26)$$

Deve-se considerar que as variáveis preditoras X_i podem ser qualitativas, quantitativas ou até mesmo indicadoras. Essa flexibilidade torna o modelo logístico múltiplo bastante útil nas análises estatísticas.

A função de log-verossimilhança definida em (12) pode ser estendida para o modelo de regressão logística múltipla, logo tem-se que:

$$\ln L(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} Y_i (\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}) - \sum_{i=1}^{n} ln (1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1})) \quad (27)$$

Assim como ocorre na regressão logística simples, os estimadores $\beta_0, \beta_1, \ldots, \beta_{p-1}$ que maximizam a expressão (26) devem ser obtidos a partir de métodos numéricos. As estimativas de máxima verossimilhança são denotadas por b_0, b_1, ..., b_{p-1}.

A função resposta logística e os valores ajustados são dados por:

$$\hat{\pi} = \frac{\exp(\boldsymbol{b}'\boldsymbol{X})}{1+\exp(\boldsymbol{b}'\boldsymbol{X})} = (1 + \exp(-\boldsymbol{b}'\boldsymbol{X}))^{-1} \quad (28)$$

$$\hat{\pi}_i = \frac{\exp(\boldsymbol{b}'\boldsymbol{X}_i)}{1+\exp(\boldsymbol{b}'\boldsymbol{X}_i)} = (1 + \exp(-\boldsymbol{b}'\boldsymbol{X}_i))^{-1} \quad (29)$$

O próximo passo na construção do modelo logístico é a seleção de variáveis preditoras. Entre os métodos de seleção de variáveis, tem-se o forward, o backward e o stepwise. Neste trabalho, o processo utilizado é o stepwise, que consiste em adicionar e remover variáveis explicativas do modelo. O método de seleção é o mesmo utilizado no caso da regressão linear.

Como o interesse é verificar se um subconjunto das variáveis X_i pode ser retirado do modelo de regressão logística múltipla, deve-se testar se os coeficientes de regressão β_k são iguais a zero. Para isso, o teste utilizado é o da razão de verossimilhança, que é baseado na estatística de deviance do modelo.

A deviance de um modelo compara o logaritmo da verossimilhança deste modelo com o logaritmo da verossimilhança do modelo completo. O modelo completo é o que se ajusta completamente aos dados, ou seja, para cada observação existe um parâmetro. Assim, a deviance do modelo logístico definido na expressão (26) é dada por:

$$DEV(X_0, X_1, ...,X_{p-1}) = - 2 LL \quad (30)$$

em que LL representa o logaritmo da verossimilhança do modelo logístico definido em (29).

A interpretação da deviance é a seguinte: se a deviance é pequena, a explicação do modelo ajustado é praticamente igual a do modelo completo, assim o modelo encontrado pode ser utilizado no ajuste dos dados, isso é uma vantagem, pois normalmente o modelo ajustado é mais simples, ou seja, tem uma quantidade menor de parâmetros. No caso de a deviance ser grande, o modelo ajustado explica de forma não satisfatória os dados, assim, não é adequado utilizar esse modelo.

Outra medida que pode ser calculada é a deviance parcial, que consiste na diferença entre as deviances de dois modelos, por meio dela é possível testar se determinadas variáveis explicativas podem ser retiradas do modelo.

Considerando o modelo logístico completo com função resposta dada por:

$$\pi_i = [1 + \exp(-(\beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}))]^{-1} \quad (31)$$

Após calcular as estimativas de máxima verossimilhança ($\mathbf{b}_C$) e a deviance do modelo, que é definida como DEV(X_0, X_1, ...,X_{p-1}). Deve-se proceder ao teste das seguintes hipóteses:

H_0: $\beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$

H_1: pelo menos um β_k é diferente de zero

O modelo de regressão logística tem a seguinte função resposta:

$$\pi_i = [1 + \exp(-\beta_0 + \beta_1 X_1 + \ldots + \beta_{q-1} X_{q-1})]^{-1} \quad (32)$$

Calculando as estimativas de máxima verossimilhança ($\mathbf{b}_R$) e a deviance deste modelo, que é definida como $DEV(X_0, X_1, ..., X_{q-1})$. É possível comparar o modelo logístico completo com o modelo reduzido e observar o seguinte:

- Se a deviance residual do modelo reduzido não é muito maior que a deviance residual do modelo completo, a conclusão a que se chega é que as variáveis X_q, X_{q+1}, ..., X_{p-1} podem ser retiradas do modelo logístico múltiplo sem perda de informação.
- Uma grande diferença entre as duas deviances residuais significa que as variáveis preditoras X_q, X_{q+1}, ..., X_{p-1} devem ser mantidas no modelo, pois elas melhoram o ajuste do modelo, ou seja, melhoram a explicação dada pelo mesmo.

A diferença entre duas deviances, a deviance parcial, é dada por:

$$DEV(X_q, X_{q+1}, ..., X_{p-1} | X_0, X_1, ..., X_{q-1}) = DEV(X_0, X_1, ..., X_{q-1}) - DEV(X_0, X_1, ..., X_{p-1}) \quad (33)$$

A deviance parcial segue, aproximadamente, para um n razoavelmente grande, uma distribuição qui-quadrado com p-q graus de liberdade. Os graus de liberdade correspondem à diferença entre os graus de

liberdade dos erros para os dois modelos ajustados, ou seja, (n-q)-(n-p) = p-q.

Utilizando a regra de decisão de aproximação pelo Qui-Quadrado, tem-se que:

- Se DEV(X_q, X_{q+1}, ..., X_{p-1}| X_0, X_1, ..., X_{q-1})$\leq \chi^2$(p-q), a hipótese nula não deve ser rejeitada, ou seja, as variáveis X_q, X_{q+1}, ..., X_{p-1} podem ser retiradas do modelo logístico múltiplo sem prejuízo na informação.
- Se DEV(X_q, X_{q+1}, ..., X_{p-1}| X_0, X_1, ..., X_{q-1})$> \chi^2$(p-q), a hipótese nula deve ser rejeitada, ou seja, as variáveis preditoras X_q, X_{q+1}, ..., X_{p-1} devem ser mantidas no modelo, pois elas melhoram o ajuste daquele modelo.

Após todos esses passos, deve-se proceder ao diagnóstico do modelo, que é feito a partir da verificação do ajuste do modelo, da verificação do ajuste da parte linear do modelo de regressão logística e da identificação da deviance residual dos valores extremos (outliers). Os outliers são observações bem afastadas do restante dos dados e que, por consequência, têm resíduos grandes, por essa razão, essas observações têm um efeito muito grande sobre a função de regressão de mínimos quadrados ajustada.

Os pontos cruciais no diagnóstico do modelo são verificar se a função estimada é monotônica e em forma sigmoidal, a presença de pontos influentes e se o modelo logístico ajustado é adequado.

Regressão Multinível

O modelo de regressão multinível ou hierárquico consegue incorporar em sua formulação a natureza de agrupamento da população em estudo, uma vez que diferentemente do modelo de regressão tradicional leva em consideração a correlação entre indivíduos associados a um mesmo nível de agregação.

A vantagem de se utilizar o modelo de regressão multinível é que ele possibilita o estudo da interação entre as variáveis em seus diferentes níveis.

Outro aspecto que torna a regressão multinível bastante aplicável é que pela variância estar decomposta nos diversos níveis, pode-se ter uma melhor compreensão do processo em estudo como um todo.

Para compreender a regressão multinível é importante dominar dois conceitos fundamentais, quer sejam:

- Correlação intra-classe que no modelo nulo, ou seja, sem a presença de variáveis explicativas, pode ser utilizada na mensuração da homogeneidade de duas ou mais medidas e é interpretada como a estimativa da proporção da variabilidade total que é atribuída ao objeto em estudo. Assim, aplicando o conceito à situação que será analisada, tem-se que a correlação intra-classe mede a proporção da variância entre as turmas em relação à variância total, isto é, essa correlação mostra o quanto da variação no desempenho dos estudantes é explicada por diferenças entre as turmas de cada um deles. Esse coeficiente varia entre 0 e 1. Quando o seu valor é próximo de zero, isso significa que as turmas são homogêneas entre si e que o

desempenho do aluno independe da turma a que ele pertence. Na situação de o coeficiente ser próximo de 1, tem-se que toda a variabilidade no desempenho dos estudantes se deve às diferenças existentes entre as turmas, nesse caso, as características individuais não contribuem para o desempenho dos alunos, mas também não atrapalham, já que o desempenho é afetado apenas pelas características da turma a que ele pertence.

- Interação inter-nível que mede a interação entre variáveis medidas em dois níveis diferentes de um conjunto de dados com estrutura hierárquica.

Para definir um modelo multinível com dois níveis é necessário especificar duas equações, uma para cada um dos níveis que estão sendo considerados, dessa forma, tem-se que os estudantes pertencem ao nível micro enquanto as turmas compõem o nível macro.

Ao considerar o modelo multinível com dois níveis, deve-se observar que dependendo das características e peculiaridades de cada um deles, estes podem ser divididos em:

1. Modelo multinível de componentes de variância.
2. Modelo multinível de coeficientes aleatórios.
3. Modelo multinível com mais de uma variável explicativa.

O modelo de regressão multinível, considerando a existência de dois níveis, traz o aluno como a unidade do nível 1, identificado pelo índice i, e a

turma como a unidade do nível 2, identificada pelo índice k. Para o presente estudo, serão consideradas K turmas, k = 1, 2, ..., K cada uma delas com n_k estudantes, i = 1, 2, ..., n_k. Logo, considerando um modelo para p variáveis explicativas do nível 1 e q variáveis explicativas do nível 2, tem-se que o modelo obtido é dado por:

$$Y_{ik} = \gamma_{00} + \gamma_{p0} X_{pik} + \gamma_{0q} W_{qk} + \gamma_{pq} W_{qk} * X_{pik} + u_{pk} * X_{pik} + e_{ik} + u_{0k} \qquad (34)$$

O modelo definido acima é um modelo misto, pois em sua formulação há tanto variáveis que contém efeitos fixos quanto variáveis que apresentam efeitos aleatórios.

A parte fixa do modelo definido em (34) é dada por:

$$\gamma_{00} + \gamma_{p0} X_{pik} + \gamma_{0q} W_{qk} + \gamma_{pq} W_{qk} * X_{pik} \qquad (35)$$

que além das variáveis explicativas pertencentes ao nível 1 (X_{pik}) e ao nível 2 (W_{qk}), apresenta também um termo ($W_{qk} * X_{pik}$), que mostra a interação entre os dois níveis.

A parte aleatória do modelo definido em (34) é dada por:

$$u_{pk} * X_{pik} + e_{ik} + u_{0k} \qquad (36)$$

que representa os efeitos aleatórios que influenciam no desempenho do estudante, e que atuam tanto no nível do aluno quanto no nível da turma, mas não são captados pela parte determinística do modelo. A componente

aleatória do modelo é decomposta tanto no erro do nível 1, e_{ik}, quanto no erro do nível 2, u_{0k}.

Para estimar os parâmetros do modelo misto será utilizado o método de máxima verossimilhança, pois ele produz estimadores com propriedades importantes, tais como consistência e eficiência e esses estimadores também possuem variância mínima.

Outra técnica que pode ser usada na estimação dos coeficientes é a máxima verossimilhança restrita que se assemelha ao método de máxima verossimilhança para a estimação dos coeficientes fixos. Entretanto, quando se trata da estimação dos coeficientes aleatórios, os dois métodos diferem, pois o método de máxima verossimilhança restrita considera os graus de liberdade perdidos na estimação dos coeficientes fixos, enquanto no método de máxima verossimilhança isso não ocorre.

O passo que se segue à estimação dos parâmetros é justamente a verificação da significância dos mesmos. Assim, o teste de Wald é utilizado para avaliar se o parâmetro é estatisticamente significativo para um determinado nível de significância. Logo, as hipóteses a serem testadas são:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

A estatística do teste é obtida por meio da razão do coeficiente pelo seu respectivo erro padrão, tem-se assim, que essa estatística apresenta distribuição Binomial. É necessário considerar que o teste de Wald é menos preciso que o teste de máxima verossimilhança, por isso, para os coeficientes

identificados como não significativos, ou seja, para os casos em que a hipótese nula for rejeitada, o ideal é testar os parâmetros considerando o teste da razão de verossimilhança.

O teste de razão de verossimilhança permite testar modelos encaixados, ou seja, ele testa uma hipótese nula contra uma hipótese alternativa que apresenta um maior número de parâmetros que a primeira. Assim, a estatística do teste compara o valor observado com a distribuição Qui-Quadrado com número de graus de liberdade igual a diferença no número de parâmetros dos dois modelos considerados nas hipóteses.

Outro estudo que pode ser feito é a comparação entre os modelos que foram ajustados para os dados. Por meio da deviance é possível medir o grau de desajuste do modelo. A deviance é definida por:

$$\text{Deviance} = -2\ln(L_0) - [-2\ln(L_1)] = -2\,LL \qquad (37)$$

em que L_0 é a verossimilhança do modelo nulo, ou seja, sem a presença de covariáveis, L_1 é a verossimilhança do modelo completo e LL representa o logaritmo da verossimilhança.

Assim, tem-se que o modelo que apresentar a menor deviance é aquele que melhor se ajusta ao conjunto de dados. Espera-se que a introdução das variáveis explicativas melhore o ajuste do modelo. Caso os modelos que se deseja comparar não sejam encaixados, pode-se utilizar os critérios de informação de Akaike, de Akaike corrigido e o bayesiano para compará-los.

Um aspecto que deve ser considerado é que quando se deseja verificar a diferença entre modelos hierárquicos encaixados, deve-se comparar a deviance entre modelos. Logo, a diferença entre deviances possui uma distribuição aproximadamente Qui-Quadrado em que os graus de liberdade são iguais à diferença entre o número de parâmetros que foram testados em cada um dos modelos considerados. Assim, é possível comparar dois modelos por meio do teste do Qui-Quadrado.

Após os testes de hipóteses, deve-se proceder à análise dos resíduos, que representam o quanto as estimativas médias estão afastadas da média geral e que confirmarão se os erros do modelo ajustado respeitam os seguintes pressupostos para o nível da turma:

- Seguem uma distribuição Normal, essa condição pode ser verificada por meio de um gráfico de probabilidade Normal. Assim, se os erros forem mesmo normalmente distribuídos, os pontos do gráfico devem se localizar o mais próximo possível de uma reta.
- Têm média igual a zero.
- Têm variância constante, ou seja, os erros são homocedásticos.
- São independentes.

A verificação dos três últimos pressupostos também se baseiam na análise gráfica. Assim, plotando em um gráfico os resíduos em função dos valores estimados da variável dependente. Para atender os pressupostos, é necessário que os pontos se distribuam de forma aleatória em torno da reta que corresponde ao resíduo nulo.

Metodologia

Os dados utilizados como referência no trabalho são provenientes das Listas de Menções, dos históricos dos estudantes e das listas de Ofertas. Esses dados foram extraídos do SIGRA (Sistema de Informação Acadêmica de Graduação) e a partir deles pode-se analisar e descrever o desempenho dos estudantes.

Inicialmente foi realizada uma análise descritiva dos dados que permitiu apresentar o perfil dos estudantes e assim, observar e discutir aspectos que são realmente relevantes na definição do desempenho desses alunos.

Utilizou-se a Regressão Logística Multinível para construir um modelo que explicassse o rendimento dos estudantes. Para isso, considerou-se as características que foram relevantes na etapa descritiva tanto no nível do aluno quanto no nível da turma.

A Regressão Logística foi adotada no desenvolvimento do trabalho, pois a variável resposta do estudo, que no caso, é o desempenho do aluno é categorizada binária, com a aprovação como a resposta referência. Já a opção pela abordagem da Regressão Multinível se deve ao fato que este método considera a estrutura hierárquica dos dados, estabelecendo as relações entre as variáveis e também agregando informações referentes à correlação existente entre indivíduos associados a um mesmo nível de agregação. Neste trabalho, foram considerados dois níveis, o nível macro que são as turmas e o nível micro, os alunos. Entre os principais aspectos que foram considerados no nível do estudante estão: sexo, país de

nascimento, naturalidade, cidade e unidade da federação em que reside, ano e semestre em que ingressou na universidade, curso, forma de ingresso e de saída, semestre em que cursou a disciplina, modalidade da disciplina, entre outros. Já em relação às turmas, considerou-se os seguintes fatores: turno para o qual a turma é destinada, horário e local em que é ofertada, qual professor lecionou para cada turma, a situação do professor (quadro/substituto), entre outros.

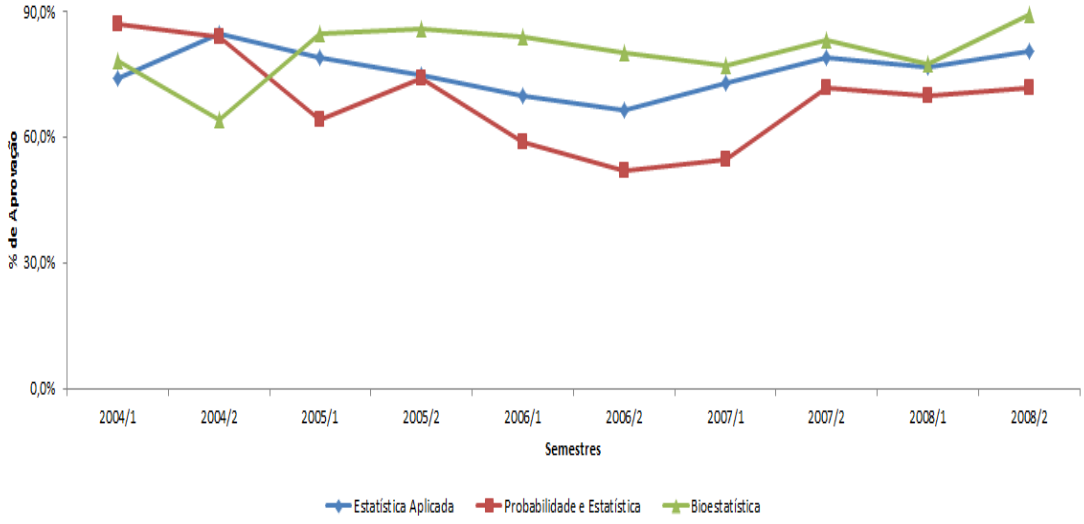
O Software Estatístico utilizado para fazer a análise descritiva dos dados foi o SAS 9.3. Já na etapa de modelagem foi utilizado o MLwiN, que é um programa desenvolvido com o propósito de analisar estudos próprios de Regressão Multinível.

Análise Descritiva

Panorama Geral dos Anos de 2004 a 2008

Fazendo um panorama geral em relação aos resultados obtidos pelos alunos nos anos de 2004 a 2008, tem-se que o número de aprovações e reprovações nos semestres que compõem esse período revela o desempenho dos alunos que cursaram as disciplinas de Bioestatística, Estatística Aplicada e Probabilidade e Estatística.

Figura 1 - Percentual de Aprovação segundo a Disciplina por Semestre nos Anos de 2004 a 2008



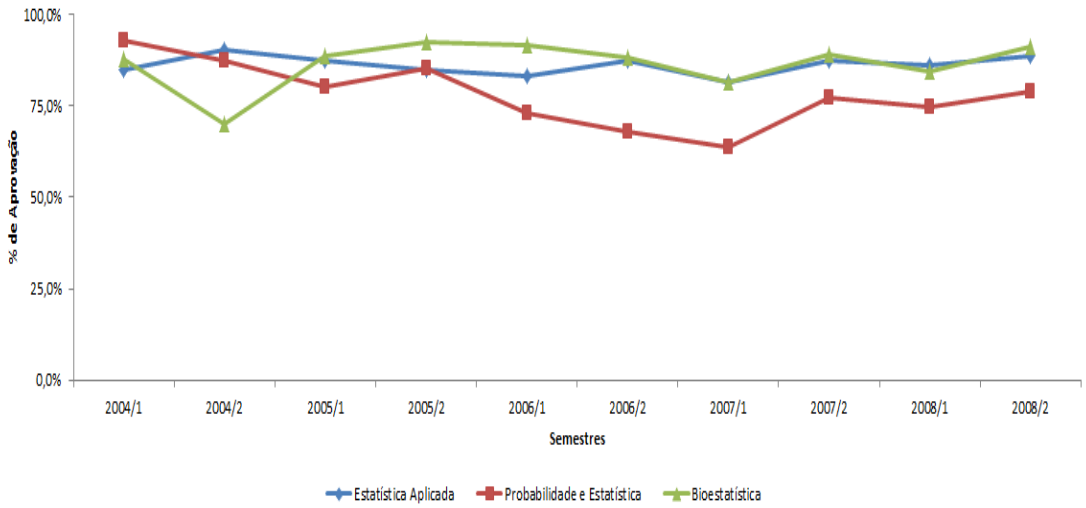
A partir da Figura 1, é possível observar que, na maior parte dos períodos, os alunos de Bioestatística são os que têm o melhor desempenho, atingindo 89,3% de aprovação no 2º semestre de 2008, enquanto os de Probabilidade e Estatística são os que apresentam o pior rendimento, atingindo 52% de aprovação no 2º semestre de 2006.

Tabela 1 - Percentual de SR's por Semestre nos Anos de 2004 a 2008

Período	Disciplinas			Total
	Estatística Aplicada	Probabilidade e Estatística	Bioestatística	
2004/1	12,1%	5,9%	10,6%	9,4%
2004/2	6,2%	3,9%	8,1%	5,8%
2005/1	9,3%	18,4%	4,1%	10,9%
2005/2	11,4%	12,5%	7,0%	10,4%
2006/1	15,8%	18,8%	8,4%	14,9%
2006/2	23,2%	22,6%	8,9%	20,4%
2007/1	10,1%	14,0%	5,4%	10,4%
2007/2	9,4%	6,8%	6,5%	8,0%
2008/1	10,4%	6,2%	8,3%	8,6%
2008/2	8,9%	8,9%	1,9%	8,7%
Total	11,9%	11,6%	7,1%	10,8%

Como o objetivo é verificar se existe diferença entre o rendimento dos alunos considerando as turmas das quais eles fazem parte, foi calculado um segundo percentual de aprovação, que desconsiderou os estudantes que tiveram menção SR, ou seja, Sem Rendimento. Isso foi feito porque uma boa parcela dos alunos que obtém SR abandona a disciplina em que está matriculado, dessa forma, não seria conveniente considerá-lo como parte da turma.

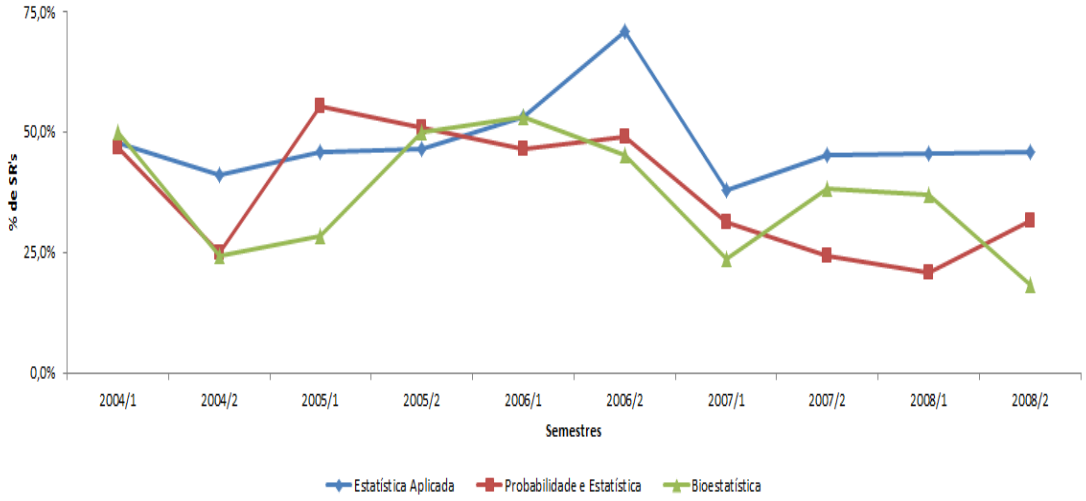
Figura 2 - Percentual de Aprovação segundo a Disciplina por Semestre nos Anos de 2004 a 2008 (desconsiderando os Alunos com Menção SR)



A partir da Figura 2, é possível observar que o desempenho dos alunos melhora consideravelmente ao se desprezar na análise os estudantes que obtiveram Menção SR. Dessa forma, na maior parte dos períodos, os alunos de Bioestatística têm um desempenho semelhante aos de Estatística Aplicada, enquanto os estudantes de Probabilidade e Estatística continuam apresentando o pior rendimento, atingindo 63,5% de aprovação no 1º semestre de 2007.

Como o percentual de aprovações difere substancialmente quando as reprovações com SR são ou não consideradas, viu-se o efeito de contabilizar o percentual de SR's nas reprovações.

Figura 3 - Percentual de SR's nas Reprovações segundo a Disciplina por Semestre nos Anos de 2004 a 2008



A partir da Figura 3, é possível perceber que, na maior parte dos períodos, os alunos de Estatística Aplicada são os que têm os maiores percentuais de SR's ao se considerar o número de reprovações, atingindo 70,8% de SR's no 2º semestre de 2006, enquanto os de Probabilidade e Estatística e Bioestatística apresentam os menores percentuais em

semestres alternados, entretanto o menor percentual, 18,2%, é atingido pelos alunos de Bioestatística no 2º semestre de 2008.

O cenário apresentado mostra a importância de se entender quais são as características que podem estar eventualmente associadas ao rendimento dos estudantes dessas disciplinas. Além disso, pode-se identificar se esses fatores são os mesmos para as três disciplinas.

Sendo assim optou-se por utilizar neste estudo dados referentes aos alunos que cursaram e também aos professores que lecionaram as disciplinas de Bioestatística, Estatística Aplicada e Probabilidade e Estatística no 1º e no 2º semestres do ano de 2008.

No 1º semestre de 2008, foram ofertadas 18 turmas, sendo nove de Estatística Aplicada, seis de Probabilidade e Estatística e três de Bioestatística. E, no 2º semestre de 2008, foram ofertadas 17 turmas, a única diferença é que houve a redução de uma turma de Bioestatística do 1º para o 2º semestre de 2008.

Tabela 2 - Percentual de Trancamentos por Semestre nos Anos de 2004 a 2008

Período	Disciplinas			Total
	Estatística Aplicada	Probabilidade e Estatística	Bioestatística	
2004/1	2,5%	2,3%	3,9%	2,5%
2004/2	1,8%	2,8%	6,9%	2,9%
2005/1	2,6%	7,6%	7,6%	5,1%
2005/2	2,6%	5,1%	1,5%	2,8%
2006/1	1,4%	1,6%	2,0%	1,5%
2006/2	2,2%	3,9%	1,9%	2,6%
2007/1	2,3%	1,3%	1,2%	1,7%
2007/2	0,0%	0,3%	0,0%	0,1%
2008/1	1,5%	1,1%	1,7%	1,3%
2008/2	0,5%	0,0%	0,0%	0,3%
Total	1,8%	2,5%	3,0%	2,1%

Foram selecionados apenas os estudantes que cursaram a disciplina integralmente, ou seja, os alunos que fizeram o trancamento não foram contabilizados. A partir dos percentuais da Tabela 2, pode-se perceber que os trancamentos não representam uma parcela significativa dos estudantes, por isso, não há prejuízo na análise ao desconsiderá-los. Assim, o banco de dados é formado por 222 estudantes de Bioestatística (119 no 1º semestre de 2008 e 103 no 2º semestre de 2008), 1076 de Estatística Aplicada (529 no 1º semestre de 2008 e 547 no 2º semestre de 2008) e 679 de Probabilidade e Estatística (352 no 1º semestre de 2008 e 327 no 2º semestre de 2008).

As informações disponíveis em relação aos alunos são: sexo, país de nascimento, naturalidade, cidade e unidade da federação em que reside, ano e semestre em que ingressou na universidade, curso, forma de ingresso e de saída, semestre em que cursou a disciplina, modalidade da disciplina, turma, menção obtida e percentual de faltas.

As informações disponíveis em relação às turmas e aos professores são: número de vagas, turno para o qual a turma é destinada, horário e local em que é ofertada, qual professor lecionou para cada turma e a situação do professor (quadro/substituto). É importante perceber que existe uma estreita relação entre professor e turma, essas variáveis só não estão mais associadas porque alguns dos professores lecionam para mais de uma turma.

Características Sociodemográficas dos Estudantes

A partir das informações sociodemográficas dos estudantes, é possível perceber que os estudantes de Probabilidade e Estatística são em sua maioria absoluta do sexo masculino. Já as disciplinas de Estatística Aplicada e Bioestatística apresentam uma distribuição equitativa no que diz respeito ao sexo dos alunos.

Tabela 3 - Características Sociodemográficas dos Estudantes - 2008

Características		Disciplinas			Total
		Estatística Aplicada	Probabilidade e Estatística	Bioestatística	
Sexo					
	Feminino	46,8%	13,8%	56,3%	36,5%
	Masculino	53,3%	86,2%	43,7%	63,5%
Local de Nascimento					
	Brasil	97,7%	99,0%	99,6%	98,3%
	Exterior	2,3%	1,0%	0,5%	1,7%
Naturalidade (para os Brasileiros)					
	Distrito Federal	66,5%	64,6%	72,0%	66,5%
	Goiás	6,9%	10,9%	8,1%	8,4%
	Minas Gerais	5,5%	4,2%	5,9%	5,1%
	Rio de Janeiro	3,5%	4,2%	3,6%	3,8%
	São Paulo	3,0%	4,0%	1,8%	3,2%
	Outras	14,6%	12,2%	8,6%	13,1%
UF de Residência					
	Distrito Federal	96,8%	94,1%	91,9%	95,3%
	Goiás	2,2%	5,3%	6,3%	3,7%
	Outras	1,0%	0,6%	1,8%	1,0%
Cidade de Residência					
	Brasília	60,7%	54,4%	57,7%	58,2%
	Taguatinga	9,1%	10,8%	13,5%	10,2%
	Sobradinho	3,4%	3,5%	5,4%	3,6%
	Outras	26,9%	31,3%	23,4%	28,0%
Local de Residência (*)					
	DF - Alta Renda	62,6%	61,6%	59,9%	61,0%
	DF - Média Renda	24,7%	26,6%	28,4%	26,9%
	DF - Baixa Renda	10,0%	6,2%	5,0%	8,0%
	GO - Entorno	1,1%	1,3%	2,7%	1,4%
	Outros	1,6%	4,3%	4,1%	2,8%

(*) O DF - Alta Renda é composto pelas cidades de Brasília, Asa Sul, Asa Norte, Sudoeste, Área Octogonal Sul, Lago Sul, Lago Norte, Park Way e Jardim Botânico. O DF - Média Renda engloba as cidades de Taguatinga, Gama, Vicente Pires, Águas Claras, Sobradinho, Sobradinho II, Núcleo Bandeirante, Guará I, Guará II, Lúcio Costa, Cruzeiro, São Sebastião, Riacho Fundo I, Candangolândia e Vila Planalto. O DF - Baixa Renda é formado pelas cidades de Brazlândia, Itapoã, Arapoanga, Planaltina, Paranoá, Ceilândia, Samambaia, Santa Maria, Recanto das Emas, Riacho Fundo II, Valparaíso e Valparaíso II. As cidades de Goiás que constituem o Entorno do Distrito Federal são Águas Lindas, Cidade Ocidental, Cristalina, Formosa, Luziânia, Novo Gama, Padre Bernardo, Planaltina de Goiás, Santo Antônio Descoberto e Valparaíso de Goiás.

Em todas as disciplinas, predominam os estudantes brasileiros, entretanto, na disciplina de Estatística Aplicada, há um número expressivo de alunos africanos, principalmente dos países de Cabo Verde e de Guiné-Bissau.

Os estudantes brasileiros nasceram em sua maioria no Distrito Federal, entretanto, outros estados que se destacam são Goiás, Minas Gerais, Rio de Janeiro e São Paulo.

Os estudantes dessas disciplinas moram principalmente no Distrito Federal, mais de 50% deles reside nas Regiões Administrativas que englobam a área de alta renda do Distrito Federal, nas cidades de Brasília, Asa Sul, Asa Norte, Sudoeste, Área Octogonal Sul, Lago Sul, Lago Norte, Park Way e Jardim Botânico.

Vida Acadêmica dos Estudantes

A partir das informações sobre a vida acadêmica dos estudantes, é possível perceber que os alunos de Estatística Aplicada e Probabilidade e Estatística ingressaram na Universidade, principalmente, nos anos de 2007 e 2008, enquanto os de Bioestatística entraram majoritariamente no ano de 2007.

Tabela 4 - Vida Acadêmica dos Estudantes

Fatores		Disciplinas					
		Estatística Aplicada		Probabilidade e Estatística		Bioestatística	
Semestre		1º/2008	2º/2008	1º/2008	2º/2008	1º/2008	2º/2008
Ano de Ingresso							
	2004	2,3%	1,8%	8,2%	4,6%	8,4%	1,0%
	2005	11,2%	7,7%	10,2%	12,9%	13,4%	4,9%
	2006	19,9%	13,5%	13,4%	9,8%	11,8%	5,8%
	2007	59,2%	20,7%	57,7%	24,5%	64,7%	60,2%
	2008	4,5%	53,9%	5,1%	46,8%	0,9%	25,2%
	Outros	3,0%	2,4%	5,4%	1,5%	0,8%	2,9%
Curso							
	Administração	22,1%	18,1%	-	-	-	-
	Agronomia	-	0,6%	-	-	21,9%	23,3%
	Arquivologia	8,5%	10,1%	-	-	-	-
	Biblioteconomia	10,0%	10,4%	-	-	-	-
	Ciência da Computação	-	-	10,2%	10,4%	-	-
	Ciência Política	7,8%	7,7%	-	-	-	-
	Ciências Biológicas	-	-	0,3%	-	12,6%	1,9%
	Ciências Contábeis	14,2%	15,2%	0,6%	0,3%	-	-
	Ciências Sociais	11,3%	7,5%	-	-	-	-
	Computação	-	-	10,8%	10,1%	-	-
	Engenharia Civil	-	-	16,8%	18,7%	-	-
	Engenharia de Redes de Comunicação	-	-	8,2%	10,7%	-	-
	Engenharia Elétrica	-	-	11,4%	12,5%	-	-
	Engenharia Florestal	0,2%	0,2%	-	-	28,6%	35,0%
	Engenharia Mecânica	-	-	11,1%	12,5%	0,8%	1,9%
	Engenharia Mecatrônica	-	-	7,4%	8,6%	-	-
	Farmácia	-	-	-	-	12,6%	7,8%
	Geografia	6,4%	6,8%	-	-	-	-
	Matemática	0,8%	1,7%	19,0%	12,2%	0,8%	-
	Medicina Veterinária	-	-	-	-	20,2%	27,2%
	Psicologia	3,2%	7,9%	-	-	-	-
	Relações Internacionais	7,8%	8,4%	-	-	-	-
	Outros	7,7%	5,7%	4,3%	4,0%	2,5%	2,9%
Forma de Entrada							
	Vestibular	79,2%	59,6%	83,8%	69,1%	73,1%	75,7%
	Programa de Avaliação Seriada	10,0%	32,4%	11,4%	25,7%	24,4%	22,3%
	Transferência Facultativa	4,2%	1,8%	0,6%	0,3%	1,7%	1,0%
	Transferência Obrigatória	2,7%	3,3%	2,8%	4,6%	0,8%	1,0%
	Outras	4,0%	2,9%	1,4%	0,3%	-	-
Forma de Saída							
	Cursando	90,7%	94,2%	88,9%	94,2%	89,1%	99,0%
	Desligamento por Rendimento	2,5%	2,4%	4,8%	2,5%	3,4%	-
	Formatura	2,5%	1,7%	2,8%	1,5%	5,9%	1,0%
	Outras	4,4%	1,8%	3,4%	1,8%	1,7%	-
Modalidade da Disciplina							
	Módulo Livre	2,7%	1,6%	0,3%	-	0,8%	-
	Obrigatória	92,3%	94,0%	90,9%	93,6%	72,3%	86,4%
	Optativa	5,1%	4,4%	8,8%	6,4%	26,9%	13,6%
Menção							
	SR - Sem Rendimento	10,6%	9,0%	6,3%	8,9%	8,4%	1,9%
	II - Inferior	4,5%	5,8%	16,2%	4,3%	7,6%	3,9%
	MI - Médio Inferior	8,1%	4,8%	7,7%	15,0%	6,7%	4,9%
	MM - Médio	43,1%	33,5%	38,1%	37,6%	41,2%	53,4%
	MS - Médio Superior	26,7%	31,6%	25,0%	28,8%	27,7%	32,0%
	SS - Superior	7,0%	15,4%	6,8%	5,5%	8,4%	3,9%
Percentual de Faltas							
	0 a 25%	89,6%	91,1%	93,8%	91,1%	91,6%	98,1%
	26 a 50%	4,9%	5,1%	1,7%	4,6%	3,4%	-
	51 a 75%	3,4%	1,8%	0,3%	-	0,8%	1,0%
	76 a 100%	2,1%	2,0%	4,3%	4,3%	4,2%	1,0%

Em relação aos cursos de graduação, é possível perceber que os estudantes de Estatística Aplicada são provindos, principalmente, dos cursos da área de Humanas, tais como: Administração, Ciências Contábeis, Ciências Sociais, Biblioteconomia, Arquivologia, Relações Internacionais e Geografia, mas também estão presentes estudantes do curso de Psicologia, que é da área de Saúde. Já os alunos de Probabilidade e Estatística fazem, em sua maioria, cursos na área de Exatas, tais como: Matemática, Engenharia Civil, Engenharia Elétrica, Engenharia Mecânica, Computação, Ciência da Computação, Engenharia de Redes de Comunicação e Engenharia Mecatrônica. Por outro lado, os estudantes de Bioestatística são advindos, predominantemente, dos cursos na área de Saúde e Ciências Agrárias, tais como: Engenharia Florestal, Medicina Veterinária, Agronomia, entretanto, alunos de outros cursos, como: Farmácia e Ciências Biológicas.

Para todas as disciplinas, as formas de ingresso dos estudantes na Universidade foram, principalmente, o Vestibular e o Programa de Avaliação Seriada.

É possível perceber que, em todas as disciplinas, uma boa parcela dos alunos ainda está cursando a graduação. Para aqueles que saíram, as formas de saída que se destacaram foram a formatura e o desligamento por rendimento.

As disciplinas de Estatística Aplicada e Probabilidade e Estatística apresentam, principalmente, estudantes cujos cursos têm essas matérias como obrigatórias. Já a disciplina de Bioestatística, apesar de ter, em sua maioria, alunos cujos cursos têm essa matéria como obrigatória, também

apresenta uma boa parcela de estudantes que cursam essa disciplina como optativa.

Para todas as disciplinas, as menções que se destacam são o MM, que representa o desempenho mínimo exigido para a aprovação e o MS, que representa um desempenho melhor que o MM.

Em relação ao percentual de faltas, é possível perceber que a maioria dos estudantes faltou até 25% das aulas, que é o percentual máximo permitido para não ser reprovado com SR.

Tabela 5 - Distribuição dos Estudantes nas Turmas

Fatores		Disciplinas					
		Estatística Aplicada		Probabilidade e Estatística		Bioestatística	
Semestre		1º/2008	2º/2008	1º/2008	2º/2008	1º/2008	2º/2008
Turmas							
	A	6,1%	7,7%	19,9%	19,6%	49,6%	60,2%
	B	12,9%	11,7%	18,8%	19,0%	-	-
	C	11,9%	11,9%	13,1%	14,1%	26,9%	-
	D	12,1%	11,7%	21,3%	24,2%	-	-
	E	10,8%	11,3%	16,5%	19,6%	23,5%	39,8%
	F	12,3%	12,3%	10,5%	-	-	-
	G	9,6%	11,9%	-	3,7%	-	-
	H	12,5%	12,1%	-	-	-	-
	I	11,9%	9,5%	-	-	-	-
Turno							
	Diurno	59,2%	59,8%	76,4%	82,3%	73,1%	100,0%
	Noturno	28,7%	28,5%	-	-	-	-
	Ambos	12,1%	11,7%	23,6%	17,7%	26,9%	-
Horário							
	08:00 às 09:50	37,1%	35,8%	-	-	49,6%	60,2%
	10:00 às 11:50	21,7%	23,6%	18,8%	19,0%	26,9%	-
	14:00 às 15:50	12,5%	12,1%	37,8%	43,7%	23,5%	39,8%
	16:00 às 17:50	-	-	19,9%	19,6%	-	-
	19:00 às 20:50	16,8%	19,0%	-	-	-	-
	20:50 às 22:40	11,9%	9,5%	23,6%	17,7%	-	-
Local							
	Anfiteatro/Sala Grande	88,1%	90,5%	73,0%	96,3%	49,6%	60,2%
	Sala Pequena	11,9%	9,5%	27,0%	3,7%	50,4%	39,8%
Professor							
	Quadro	40,1%	31,4%	86,9%	100,0%	100,0%	100,0%
	Substituto	59,9%	68,6%	13,1%	-	-	-

Comparando o perfil dos estudantes e as turmas que foram ofertadas no 1º e no 2º semestres, percebe-se uma grande semelhança nos números referentes aos dois semestres.

Em relação à disciplina de Estatística Aplicada, é possível observar a preferência dos estudantes por determinadas turmas, visto que todas elas apresentam o mesmo número de vagas. A maioria dos alunos cursa a disciplina no período diurno, o que pode ser explicitado ao se detalhar o horário das turmas. Como grande parte das turmas é alocada em anfiteatros ou salas grandes e os professores que ministram a disciplina são principalmente substitutos, é natural que a maior parte dos estudantes também tenham aulas nesses locais e com professores substitutos.

Em relação à disciplina de Probabilidade e Estatística, é possível observar que os estudantes estão bem distribuídos entre as turmas e que algumas delas se destacam em relação a outras. A maioria dos alunos cursa a disciplina no período diurno, o que pode ser explicitado ao se detalhar o horário das turmas. Como grande parte das turmas é alocada em anfiteatros ou salas grandes e os professores que ministram a disciplina são principalmente do quadro, é natural que a maior parte dos estudantes também tenha aulas nesses locais e com professores do quadro, já no 2º semestre apenas os professores do quadro ministraram a disciplina.

Em relação à disciplina de Bioestatística, é possível observar que os estudantes estão bem distribuídos entre as turmas e que algumas delas se destacam em relação a outras. A maioria dos alunos cursa a disciplina no período diurno, o que pode ser explicitado ao se detalhar o horário das turmas, já 2º semestre foram ofertadas apenas turmas no período diurno. No

1º semestre, o percentual de estudantes que tiveram aula em anfiteatros ou salas grandes e em salas pequenas foi praticamente o mesmo, já no 2º semestre, o percentual de alunos em sala grandes ou anfiteatros se destacou. Apenas os professores do quadro ministraram a disciplina.

Professores e Turmas

Tabela 6 - Distribuição das Turmas

Fatores		Disciplinas					
		Estatística Aplicada		Probabilidade e Estatística		Bioestatística	
Semestre		1º/2008	2º/2008	1º/2008	2º/2008	1º/2008	2º/2008
Turno							
	Diurno	55,6%	55,6%	66,7%	66,7%	66,7%	100,0%
	Noturno	33,3%	33,3%	-	-	-	-
	Ambos	11,1%	11,1%	33,3%	33,3%	33,3%	-
Horário							
	08:00 às 09:50	33,3%	33,3%	-	-	33,3%	50,0%
	10:00 às 11:50	22,2%	22,2%	16,7%	16,7%	33,3%	-
	14:00 às 15:50	11,1%	11,1%	33,3%	33,3%	33,3%	50,0%
	16:00 às 17:50	-	-	16,7%	16,7%	-	-
	19:00 às 20:50	22,2%	22,2%	-	-	-	-
	20:50 às 22:40	11,1%	11,1%	33,3%	33,3%	-	-
Local							
	Anfiteatro/Sala Grande	88,9%	88,9%	66,7%	83,3%	33,3%	50,0%
	Sala Pequena	11,1%	11,1%	33,3%	16,7%	66,7%	50,0%
Professor							
	Quadro	44,4%	33,3%	83,3%	100,0%	100,0%	100,0%
	Substituto	55,6%	66,7%	16,7%	-	-	-

Comparando o perfil das turmas que foram ofertadas no 1º e no 2º semestres, percebe-se uma grande semelhança nos números referentes aos dois semestres.

Em relação às turmas de Estatística Aplicada, tem-se que foram ofertadas turmas dessa disciplina em todos os períodos, mas predominam as do turno diurno, o que pode ser detalhado ao se observar os horários das mesmas. As aulas são ministradas, predominantemente, em anfiteatros ou

salas grandes, o que pode ser explicado pelo tamanho das turmas que são todas para 65 alunos. Os professores que deram aulas dessa disciplina foram principalmente os substitutos.

Em relação às turmas de Probabilidade e Estatística, tem-se que foram ofertadas turmas dessa disciplina em todos os períodos, mas predominam as do turno diurno, o que pode ser detalhado ao se observar os horários das mesmas, entretanto não foram ofertadas turmas exclusivas para o turno noturno. As aulas são ministradas principalmente em anfiteatros ou salas grandes, apesar de o tamanho das turmas ser bem variado. Os professores que deram aulas dessa disciplina foram principalmente os do quadro, já no 2º semestre apenas os professores do quadro ministraram a disciplina.

Em relação às turmas de Bioestatística, tem-se que foram ofertadas turmas dessa disciplina em todos os períodos, mas predominam as do turno diurno, o que pode ser detalhado ao se observar os horários das mesmas, entretanto não foram ofertadas turmas exclusivas para o turno noturno, já no 2º semestre foram ofertadas apenas turmas no período diurno. As aulas são ministradas, predominantemente, em salas pequenas, já no 2º semestre a divisão das turmas entre salas grandes e pequenas foi semelhante, o que pode ser explicado pelo tamanho das turmas. Apenas professores do quadro deram aulas dessa disciplina.

Análise Bivariada

Com a análise bivariada, é possível verificar quais as variáveis estão mais relacionadas com a aprovação dos estudantes e quais delas podem ser as variáveis explicativas do modelo que explica as diferenças entre as turmas e que são determinantes ao se considerar a aprovação dos alunos.

Tabela 7 - Análise Bivariada da Aprovação em relação às demais Variáveis - Estatística Aplicada

Características		Percentual de Aprovação em relação a cada Categoria	Estatística do Teste	p-valor
Sexo			19,3556	<0,0001
	Feminino	84,5%		
	Masculino	73,5%		
Local de Nascimento			19,6795	0,1406
	Brasil	79,0%		
	Exterior	64,0%		
Naturalidade (para os Brasileiros)			18,7165	0,7176
	Distrito Federal	78,0%		
	Goiás	79,2%		
	Minas Gerais	81,0%		
	Rio de Janeiro	75,7%		
	São Paulo	87,5%		
	Outras	81,7%		
UF de Residência			4,0423	0,7749
	Distrito Federal	78,7%		
	Goiás	75,0%		
	Outras	81,8%		
Cidade de Residência			42,9620	0,6789
	Brasília	80,4%		
	Taguatinga	75,5%		
	Sobradinho	83,3%		
	Outras	75,1%		
Local de Residência			6,5794	0,1599
	DF - Alta Renda	80,4%		
	DF - Média Renda	77,6%		
	DF - Baixa Renda	69,8%		
	GO - Entorno	83,3%		
	Outros	82,4%		
Ano de Ingresso			26,2476	0,0506
	2004	68,2%		
	2005	74,3%		
	2006	74,9%		
	2007	78,4%		
	2008	83,4%		
	Outros	75,9%		
Curso			86,3401	<0,0001
	Administração	87,5%		
	Arquivologia	63,0%		
	Biblioteconomia	80,9%		
	Ciência Política	77,1%		
	Ciências Contábeis	78,5%		
	Ciências Sociais	69,3%		
	Geografia	77,5%		
	Psicologia	86,7%		
	Relações Internacionais	87,4%		
	Outros	77,1%		
Forma de Entrada			18,3605	0,0187
	Vestibular	78,0%		
	Programa de Avaliação Seriada	84,4%		
	Transferência Facultativa	78,1%		
	Transferência Obrigatória	65,6%		
	Outras	67,6%		
Forma de Saída			81,1606	<0,0001
	Cursando	80,6%		
	Desligamento por Rendimento	23,1%		
	Formatura	90,9%		
	Outras	46,5%		
Modalidade da Disciplina			52,4398	<0,0001
	Módulo Livre	17,4%		
	Obrigatória	79,9%		
	Optativa	80,4%		
Semestre			2,1790	0,1399
	1º/2008	76,8%		
	2º/2008	80,4%		
Turma (*)			36,6945/	<0,0001
	A	71,9%/57,1%	52,4287	
	B	77,9%/89,1%		
	C	54,0%/81,5%		
	D	93,8%/98,4%		
	E	71,9%/61,3%		
	F	84,6%/92,5%		
	G	88,2%/80,0%		
	H	72,7%/75,8%		
	I	74,6%/78,9%		
Turno			38,8788	<0,0001
	Diurno	79,5%		
	Noturno	69,5%		
	Ambos	96,1%		
Horário			42,0869	<0,0001
	08:00 às 09:50	80,1%		
	10:00 às 11:50	90,2%		
	14:00 às 15:50	74,2%		
	19:00 às 20:50	65,3%		
	20:50 às 22:40	76,5%		
Local			0,3388	0,5605
	Anfiteatro/Sala Grande	78,9%		
	Sala Pequena	76,5%		
Professor			75,4523	<0,0001
	1	63,5%		
	2	83,2%		
	3	54,0%		
	4	71,4%		
	5	84,6%		
	6	88,7%		
	7	72,7%		
	13	95,4%		
	14	75,8%		
Professor			10,5427	0,0012
	Quadro	73,2%		
	Substituto	81,7%		

(*) A variável Turma apresenta os percentuais de aprovação do 1º e do 2º semestres de 2008, por isso há dois valores na 3ª coluna.

A um nível de significância de 5%, é possível concluir que as variáveis significativas no contexto da aprovação dos estudantes de Estatística Aplicada, considerando o nível dos alunos, são: sexo, curso, forma de entrada, forma de saída e modalidade da disciplina. É importante perceber que a variável ano de ingresso fica no limite entre a significância e a não significância. Já, em relação ao nível da turma, as variáveis significativas são: turno, horário, professor e situação do professor. Como a turma é o nível mais agregado na Regressão Multinível, ela não será considerada como uma possível variável explicativa do modelo, o fato de ser significativa será utilizado apenas como um indício de que o modelo Multinível é adequado para modelar a aprovação dos estudantes de Estatística Aplicada.

Com o objetivo de modelar a aprovação dos estudantes de maneira mais criteriosa, todas as variáveis com p-valor até 0,25 foram consideradas na modelagem inicial. Assim, o local de nascimento, o local de residência e o semestre, que são variáveis do nível dos alunos, também constaram na lista de possíveis variáveis explicativas do modelo que descreve a aprovação dos estudantes de Estatística Aplicada.

Tabela 8 - Análise Bivariada da Aprovação em relação às demais Variáveis - Probabilidade e Estatística

Características		Percentual de Aprovação em relação a cada Categoria	Estatística do Teste	p-valor
Sexo			1,7501	0,1859
	Feminino	76,6%		
	Masculino	69,9%		
Local de Nascimento			2,9115	0,8199
	Brasil	70,5%		
	Exterior	100,0%		
Naturalidade (para os Brasileiros)			23,8668	0,2996
	Distrito Federal	70,3%		
	Goiás	72,6%		
	Minas Gerais	71,4%		
	Rio de Janeiro	60,7%		
	São Paulo	74,1%		
	Outras	72,0%		
UF de Residência			2,002	0,8489
	Distrito Federal	70,4%		
	Goiás	75,0%		
	Outras	100,0%		
Cidade de Residência			32,2226	0,6924
	Brasília	70,2%		
	Taguatinga	67,1%		
	Sobradinho	66,7%		
	Outras	73,6%		
Local de Residência			4,8851	0,2993
	DF - Alta Renda	71,3%		
	DF - Média Renda	67,0%		
	DF - Baixa Renda	76,2%		
	GO - Entorno	55,6%		
	Outros	82,8%		
Curso			28,983	0,0486
	Ciência da Computação	80,0%		
	Computação	77,5%		
	Engenharia Civil	61,7%		
	Engenharia de Redes de Comunicação	75,0%		
	Engenharia Elétrica	76,5%		
	Engenharia Mecânica	72,5%		
	Engenharia Mecatrônica	77,8%		
	Matemática	68,2%		
	Outros	71,9%		
Forma de Entrada			10,7956	0,1478
	Vestibular	69,7%		
	Programa de Avaliação Seriada	78,2%		
	Transferência Facultativa	33,3%		
	Transferência Obrigatória	56,0%		
	Outras	100,0%		
Forma de Saída			40,6261	<0,0001
	Cursando	72,8%		
	Desligamento por Rendimento	24,0%		
	Formatura	80,0%		
	Outras	47,8%		
Modalidade da Disciplina			0,4791	0,7870
	Módulo Livre	100,0%		
	Obrigatória	70,9%		
	Optativa	69,2%		
Semestre			0,3214	0,5708
	1º/2008	69,9%		
	2º/2008	71,9%		
Turma (*)			13,2375/	0,0213/
	A	75,7%/79,7%	12,3525	0,0303
	B	60,6%/62,9%		
	C	76,1%/80,4%		
	D	76,0%/67,1%		
	E	55,2%/67,2%		
	F	78,4%/-		
	G	-/100,0%		
Vagas			21,2031	0,0007
	45	83,7%		
	46	78,3%		
	59	55,2%		
	65	63,5%		
	70	77,6%		
	80	71,4%		
Turno			7,4546	0,0063
	Diurno	68,4%		
	Ambos	80,1%		
Horário			15,9765	0,0011
	10:00 às 11:50	61,7%		
	14:00 às 15:50	67,0%		
	16:00 às 17:50	77,6%		
	20:50 às 22:40	80,1%		
Local			0,4205	0,5167
	Anfiteatro/Sala Grande	71,3%		
	Sala Pequena	68,2%		
Professor			16,1728	0,0129
	1	67,1%		
	5	77,1%		
	6	80,4%		
	8	74,2%		
	9	58,1%		
	10	76,0%		
	15	68,9%		
Professor			0,6577	0,4174
	Quadro	70,5%		
	Substituto	76,1%		

(*) A variável Turma apresenta os percentuais de aprovação do 1º e do 2º semestres de 2008, por isso há dois valores na 3ª coluna.

A um nível de significância de 5%, é possível concluir que as variáveis significativas no contexto da aprovação dos estudantes de Probabilidade e Estatística, considerando o nível dos alunos, são: curso e forma de saída. Já, em relação ao nível da turma, as variáveis significativas são: vagas, turno, horário e professor. Como a turma é o nível mais agregado na Regressão Multinível, ela não será considerada como uma possível variável explicativa do modelo, o fato de ser significativa será utilizado apenas como um indício de que o modelo Multinível é adequado para modelar a aprovação dos estudantes de Probabilidade e Estatística.

Com o objetivo de modelar a aprovação dos estudantes de maneira mais criteriosa, todas as variáveis com p-valor até 0,25 foram consideradas na modelagem inicial. Assim, o sexo e a forma de entrada, que são variáveis do nível dos alunos, também constaram na lista de possíveis variáveis explicativas do modelo que descreve a aprovação dos alunos de Probabilidade e Estatística.

Tabela 9 - Análise Bivariada da Aprovação em relação às demais Variáveis - Bioestatística

Características		Percentual de Aprovação em relação a cada Categoria	Estatística do Teste	p-valor
Sexo			2,4944	0,1143
	Feminino	86,4%		
	Masculino	78,4%		
Local de Nascimento			0,2075	0,6488
	Brasil	82,8%		
	Exterior	100,0%		
Naturalidade (para os Brasileiros)			16,9319	0,3230
	Distrito Federal	82,4%		
	Goiás	77,8%		
	Minas Gerais	100,0%		
	Rio de Janeiro	87,5%		
	São Paulo	75,0%		
	Outras	79,0%		
UF de Residência			0,8419	0,8394
	Distrito Federal	82,8%		
	Goiás	85,7%		
	Outras	75,0%		
Cidade de Residência			25,3396	0,6093
	Brasília	82,0%		
	Taguatinga	90,0%		
	Sobradinho	83,3%		
	Outras	80,8%		
Local de Residência			0,2466	0,9930
	DF - Alta Renda	82,7%		
	DF - Média Renda	84,1%		
	DF - Baixa Renda	81,8%		
	GO - Entorno	83,3%		
	Outros	77,8%		
Ano de Ingresso			27,1171	0,0074
	2004	54,6%		
	2005	81,0%		
	2006	75,0%		
	2007	87,8%		
	2008	81,5%		
	Outros	50,0%		
Curso			13,8211	0,2431
	Agronomia	84,0%		
	Ciências Biológicas	94,1%		
	Engenharia Florestal	81,4%		
	Farmácia	95,7%		
	Medicina Veterinária	76,9%		
	Outros	70,0%		
Forma de Entrada			2,5946	0,4584
	Vestibular	82,4%		
	Programa de Avaliação Seriada	86,5%		
	Transferência Facultativa	66,7%		
	Transferência Obrigatória	50,0%		
Forma de Saída			22,7746	0,0001
	Cursando	84,1%		
	Formatura	100,0%		
	Outras	16,7%		
Modalidade da Disciplina			9,1411	0,0104
	Módulo Livre	0,0%		
	Obrigatória	80,6%		
	Optativa	93,5%		
Semestre			5,6128	0,0178
	1º/2008	77,3%		
	2º/2008	89,3%		
Turma (*)			11,7655/	0,0028/
	A	84,8%/85,5%	2,4033	0,1211
	C	84,4%/-		
	E	53,6%/95,1%		
Vagas			0,9415	0,3319
	45	85,1%		
	60	80,2%		
Turno			0,0587	0,8086
	Diurno	82,6%		
	Ambos	84,4%		
Horário			1,5176	0,4682
	08:00 às 09:50	85,1%		
	10:00 às 11:50	84,4%		
	14:00 às 15:50	78,3%		
Local			0,9415	0,3319
	Anfiteatro/Sala Grande	85,1%		
	Sala Pequena	80,2%		
Professor			21,7641	<0,0001
	3	85,1%		
	8	53,6%		
	11	84,4%		
	16	95,1%		

(*) A variável Turma apresenta os percentuais de aprovação do 1º e do 2º semestres de 2008, por isso há dois valores na 3ª coluna.

A um nível de significância de 5%, é possível concluir que as variáveis significativas no contexto da aprovação dos estudantes de Bioestatística, considerando o nível dos alunos, são: ano de ingresso, forma de saída, modalidade da disciplina, semestre. Já, em relação ao nível da turma, a variável significativa é professor. Como a turma é o nível mais agregado na Regressão Multinível, ela não será considerada como uma possível variável explicativa do modelo, o fato de ser não significativa será utilizado apenas como um indício de que o modelo Multinível é adequado para modelar a aprovação dos estudantes de Bioestatística. Observando a Tabela 7, é possível perceber que a turma foi significativa no 1º semestre de 2008, mas não foi significativa no 2º semestre.

Com o objetivo de modelar a aprovação dos estudantes de maneira mais criteriosa, todas as variáveis com p-valor até 0,25 foram consideradas na modelagem inicial. Assim, o sexo e o curso, que são variáveis do nível dos alunos, também constaram na lista de possíveis variáveis explicativas do modelo que descreve a aprovação dos alunos de Bioestatística.

Modelagem Estatística

O MLwiN possui uma metodologia própria para a modelagem de problemas com distribuição binomial e função de ligação logito. Assim, para a elaboração dos modelos, utilizou-se, primeiramente, o método Iterativo de Mínimos Quadrados Generalizados – IGLS. Entretanto, pela resposta ser binária e para melhorar a estimativa obtida inicialmente, foi necessário utilizar também o método de estimação de Monte Carlo via Cadeia de Markov, com o procedimento do algoritmo de Metropolis-Hastings, que é particularmente útil para Modelos Lineares Generalizados Multinível.

Por razões técnicas, foi utilizado o modelo preditivo de quase verossimilhança de 2ª ordem, visto que ele apresenta estimativas melhores e mais precisas de componentes de variância que o modelo marginal de quase verossimilhança de 1ª ordem. Isso ocorre porque o modelo preditivo inclui os resíduos estimados no processo iterativo e a 2ª ordem melhora o controle que se deseja ter do grau de aproximação.

Estatística Aplicada

Definição do Modelo

Com o objetivo de elaborar um modelo multinível que explique a aprovação dos estudantes de Estatística Aplicada foi formulado um passo a passo, que orienta essa modelagem:

- Passo 1:

Primeiramente, analisa-se o modelo sem nenhuma variável explicativa. Esse modelo, conhecido como nulo ou vazio, é dado por:

$$\text{logito}(\pi_{ik}) = \gamma_{00} + e_{ik} + u_{0k} \quad (38)$$

em que γ_{00} é igual a 1,412 com um erro padrão associado de 0,179 e representa a aprovação média dos estudantes. O resíduo do nível da turma, u_{0k}, tem distribuição Normal com média zero e variância igual a 0,659 com um erro padrão associado de 0,332. Já o resíduo do nível do aluno, e_{ik}, tem, por construção, média nula e variância igual a 1. Assim, o modelo nulo é útil porque proporciona uma estimativa do coeficiente de correlação intraclasse, logo $\rho = 0{,}397$. A partir disso, é possível concluir que, aproximadamente, 40% da variância no desempenho dos estudantes de Estatística Aplicada pode ser atribuída ao nível da turma. O valor elevado do coeficiente de correlação intraclasse justifica a utilização da abordagem multinível.

O modelo vazio proporciona também uma medida de referência da deviance, que representa uma medida do grau de desajuste do modelo e que pode ser usada para comparar modelos. No caso, quanto menor a deviance, maior o ajuste do modelo. No caso do modelo nulo definido acima, tem-se que a deviance é igual a 1039,626.

- Passo 2:

Analisa-se o modelo incluindo, separadamente, cada uma das variáveis explicativas fixas do nível mais baixo, que é o nível do aluno. Isso significa que os componentes de variância correspondentes aos coeficientes são fixados em zero. Inserem-se, inicialmente, as variáveis do nível menos agregado, porque existe um maior número de observações disponíveis neste nível, o que gera coeficientes mais precisos.

Foram inseridas uma a uma as variáveis significativas com p-valor até 0,25. Assim, as variáveis sexo, curso, modalidade da disciplina, forma de saída, forma de entrada, ano de ingresso, semestre, local de nascimento e local de residência foram incluídas no modelo para avaliar se o seu ajuste melhoraria com a introdução das variáveis explicativas.

Para verificar quais variáveis têm coeficientes significativos, deve-se comparar o valor absoluto do coeficiente com duas vezes o valor do seu erro padrão. Utiliza-se o valor 2, porque essa é uma aproximação bastante útil do valor 1,96, que corresponde ao valor tabelado da Normal para uma confiança de 95%. Esse procedimento corresponde ao teste de Wald.

Considerando esse fato, tem-se que as variáveis que apresentam coeficientes significativos são: sexo, modalidade da disciplina e forma de saída. Entretanto, apesar de haver uma melhora no ajuste do modelo com o acréscimo desta última variável, como mais de 90% dos alunos de Estatística Aplicada ainda estão cursando a graduação, essa variável não será considerada no modelo que explica a aprovação desses estudantes.

Assim, o modelo definido a partir da inclusão das variáveis explicativas do nível do aluno é dado por:

$$\text{logito}(\pi_{ik}) = 1{,}217 + 0{,}718\ \text{feminino}_{ik} + 0{,}057\ \text{optativa}_{ik} - 3{,}227\ \text{módulo livre}_{ik} + e_{ik} + u_{0k} \quad (39)$$

Em relação a esse modelo, tem-se que o erro padrão associado ao intercepto é igual 0,225. O resíduo do nível da turma, u_{0k}, tem distribuição Normal com média zero e variância igual a 0,657 com um erro padrão associado de 0,338. Considerando a variável sexo tem-se que a categoria de referência é o sexo masculino e o erro padrão associado ao sexo feminino é igual a 0,165. Avaliando a variável modalidade da disciplina, tem-se que a categoria de referência é obrigatória e o erro padrão associado à optativa e a módulo livre são, respectivamente, iguais a 0,383 e 0,632.

Em comparação com o modelo nulo, a variância do nível da turma teve uma pequena redução que não chegou a modificar, consideravelmente, o coeficiente de correlação intraclasse que passou a ser igual a 0,396.

A deviance deste modelo é igual a 986,617, o que representa uma diminuição de 53,009 em relação ao anterior. Foram estimados cinco parâmetros neste modelo, o que em comparação com os dois parâmetros do modelo vazio resulta em três graus de liberdade. A diferença entre as deviances é superior ao valor tabelado de uma distribuição χ^2 com 3 graus de liberdade, que é igual a 7,815. Esse resultado significa que este modelo se ajusta muito melhor aos dados que o modelo nulo.

- Passo 3:

Analisa-se o modelo incluindo, separadamente, cada uma das variáveis explicativas fixas do nível mais agregado, que é o nível da turma. Isso significa que os componentes de variância correspondentes aos coeficientes são fixados em zero.

Foram inseridas uma a uma as variáveis significativas com p-valor até 0,25. Assim, as variáveis turno, horário, professor e situação do professor foram incluídas no modelo para avaliar se o seu ajuste melhoraria com a introdução das variáveis explicativas.

Para verificar quais variáveis têm coeficientes significativos, deve-se comparar o valor absoluto do coeficiente com duas vezes o valor do seu erro padrão. Utiliza-se o valor 2, porque essa é uma aproximação bastante útil do valor 1,96, que corresponde ao valor tabelado da Normal para uma confiança de 95%. Esse procedimento corresponde ao teste de Wald.

Considerando esse fato, tem-se que nenhuma das variáveis apresentou todos os coeficientes significativos, apesar disso, verificou-se entre elas qual tinha maior inclinação à significância. Dessa forma, optou-se por incluir a variável situação do professor, para que houvesse uma variável do nível da turma, no modelo que explica a aprovação desses estudantes.

Assim, o modelo definido a partir da inclusão das variáveis explicativas dos níveis do aluno e da turma é dado por:

$$\text{logito}(\pi_{ik}) = 0{,}807 + 0{,}716\ \text{feminino}_{ik} + 0{,}076\ \text{optativa}_{ik} - 3{,}260\ \text{módulo livre}_{ik} + 0{,}650\ \text{substituto}_{k} + e_{ik} + u_{0k} \quad (40)$$

Em relação a esse modelo, tem-se que o erro padrão associado ao intercepto é igual 0,351. O resíduo do nível da turma, u_{0k}, tem distribuição Normal com média zero e variância igual a 0,645 com um erro padrão associado de 0,343. Considerando a variável sexo tem-se que a categoria de referência é o sexo masculino e o erro padrão associado ao sexo feminino é igual a 0,175. Avaliando a variável modalidade da disciplina, tem-se que a categoria de referência é obrigatória e os erros padrão associados à optativa e a módulo livre são, respectivamente, iguais a 0,383 e 0,660. Observando a variável situação do professor, tem-se que a categoria de referência é professor do quadro e o erro padrão associado a professor substituto é igual a 0,401.

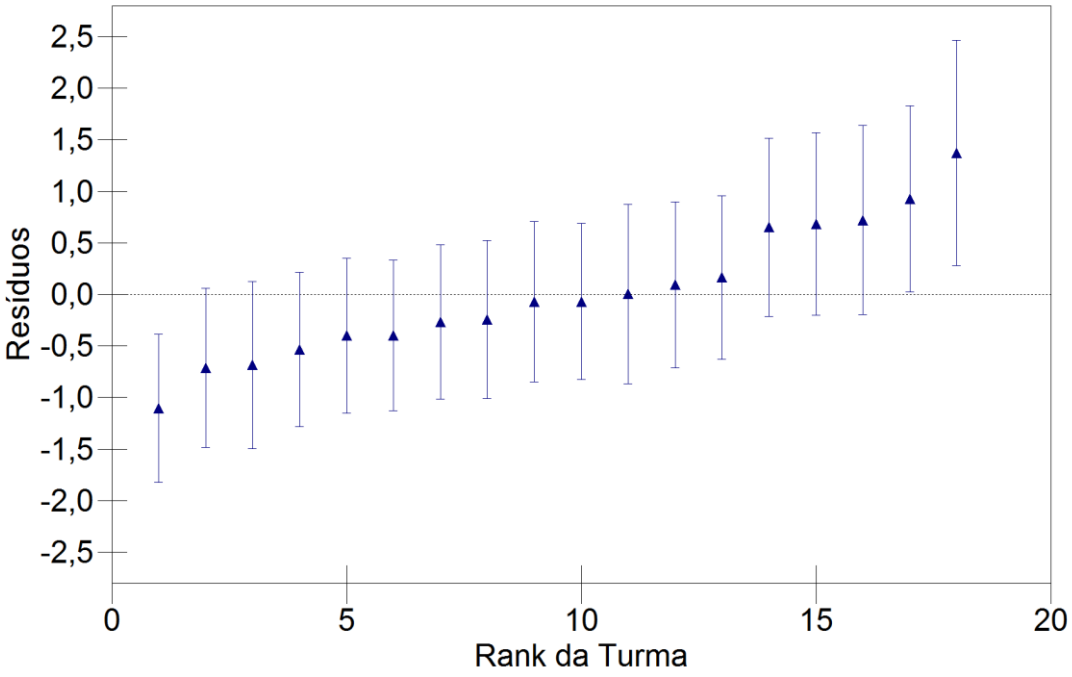
Em comparação com o modelo de variáveis explicativas do nível do aluno, a variância do nível da turma teve uma pequena redução que provocou uma diminuição no coeficiente de correlação intraclasse que passou a ser igual a 0,392.

A deviance deste modelo é igual a 986,193, o que representa uma diminuição de 0,424 em relação ao anterior. Foram estimados seis parâmetros neste modelo, o que em comparação com os cinco parâmetros do modelo de variáveis explicativas do nível do aluno resulta em um grau de liberdade. A diferença entre as deviances é menor que o valor tabelado de uma distribuição χ^2 com 1 grau de liberdade, que é igual a 3,841. Esse resultado significa que o modelo anterior se ajusta melhor aos dados que este modelo. Entretanto, apesar de não melhorar o ajuste do modelo, optou-se por utilizá-lo, visto que ele apresenta variáveis relacionadas a todos os níveis considerados.

Os modelos dos passos 2 e 3 são chamados modelos de componentes de variância, por decomporem a variância do intercepto em componentes distintos da variância para cada nível hierárquico. Nesses modelos, considera-se que o intercepto varia entre as turmas, ou seja, a variância do intercepto mostra que a aprovação média dos estudantes não é igual para todas as turmas, mas os coeficientes de regressão são fixos.

Análise de Resíduos

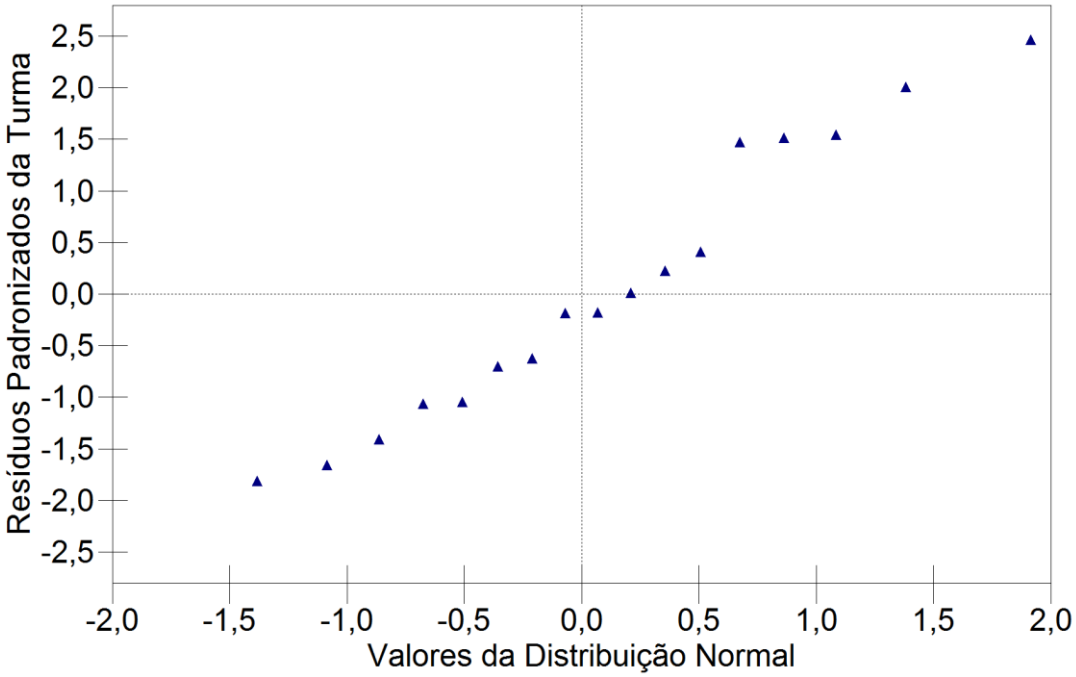
Gráfico 4 - Resíduos do Nível da Turma



A Figura 4 apresenta os resíduos plotados em ordem crescente de magnitude com seus respectivos intervalos de confiança. O intervalo que intercepta o zero, mostra que o desempenho daquela turma não é significativamente diferente do desempenho global das turmas. Se o intervalo de confiança é inteiramente abaixo da linha pontilhada, a aprovação dos

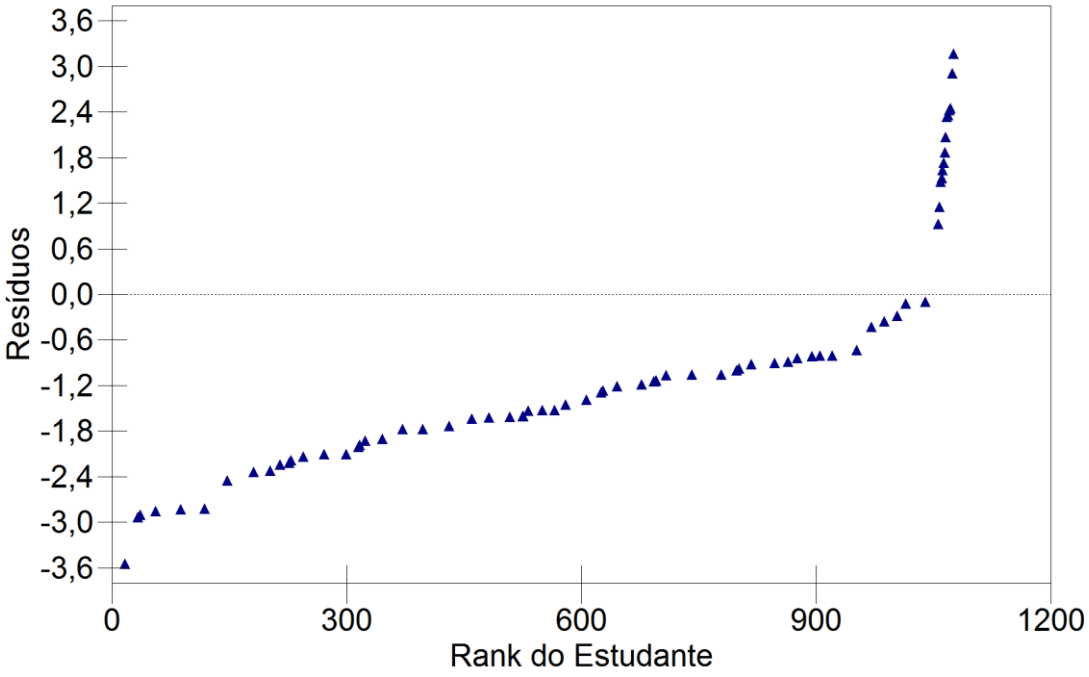
estudantes é significativamente menor para essa turma, situação que ocorre na turma E do 2º semestre de 2008; já se o intervalo de confiança é totalmente acima da linha pontilhada, a aprovação dos alunos é significativamente maior para aquela turma, situação que ocorre de forma mais significativa na turma D do 2º semestre de 2008, mas também ocorre na turma B do 2º semestre de 2008.

Figura 5 - Resíduos Padronizados do Nível da Turma em relação aos respectivos Valores da Distribuição Normal



A Figura 5 traz os resíduos padronizados plotados em relação aos valores da distribuição Normal, para se verificar o pressuposto da normalidade no nível da turma, os resíduos deveriam estar distribuídos ao longo de uma linha reta. Entretanto, apesar de isso não ocorrer, é possível observar que a violação ocorre mais intensamente nas observações finais e não chega a inviabilizar a utilização do modelo escolhido anteriormente.

Gráfico 6 - Resíduos do Nível do Estudante



Os pontos cruciais a serem verificados no diagnóstico do modelo considerando o nível do aluno são observar se a função estimada é monotônica e se tem forma sigmoidal. A partir da Figura 6, é possível analisar que esses pressupostos são atendidos.

Interpretação dos Resultados

Após o diagnóstico do modelo, é possível interpretar os resultados do modelo escolhido:

$$\text{logito}(\pi_{ik}) = 0{,}807 + 0{,}716\ \text{feminino}_{ik} + 0{,}076\ \text{optativa}_{ik} - 3{,}260\ \text{módulo livre}_{ik} + 0{,}650\ \text{substituto}_{k} + e_{ik} + u_{0k} \quad (41)$$

A média geral de aprovação na escala logito é igual a 0,807, convertendo esse valor para probabilidade, tem-se que 69,1% dos alunos foram aprovados na disciplina de Estatística Aplicada no conjunto das turmas, considerando um intervalo com 95% de confiança tem-se que o percentual de estudantes aprovados na disciplina está entre 53,0% e 81,7%.

Em relação à variável sexo, tem-se que a razão de chances é igual a 2,046, ou seja, a chance de aprovação entre os estudantes do sexo feminino é 2,046 vezes a chance de aprovação entre os alunos do sexo masculino, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 1,454 a 2,881. Assim, tem-se que a aprovação entre os estudantes do sexo feminino é 104,6% maior do que a aprovação entre os alunos do sexo masculino.

Considerando a categoria optativa, em relação à variável modalidade da disciplina, tem-se que a razão de chances é igual a 1,079, ou seja, a chance de aprovação dos que cursam a disciplina como optativa é 1,079 vezes a chance de aprovação dos que cursam a disciplina como obrigatória, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,509 a 2,286. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina como optativa é 7,9% maior do que a aprovação entre os alunos a cursam como obrigatória.

Considerando a categoria módulo livre, em relação à variável modalidade da disciplina, tem-se que a razão de chances é igual a 0,038, ou seja, a chance de aprovação dos que cursam a disciplina como módulo livre é 0,038 vezes a chance de aprovação dos que cursam a disciplina como obrigatória, considerando um intervalo com 95% de confiança tem-se que a

razão de chances varia de 0,011 a 0,140. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina como obrigatória é 96,2% maior do que a aprovação entre os alunos que a cursam como módulo livre.

Em relação à variável situação do professor, tem-se que a razão de chances é igual a 1,916, ou seja, a chance de aprovação tendo aula com o professor substituto é 1,916 vezes a chance de aprovação tendo aula com o professor do quadro, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,874 a 4,200. Assim, tem-se que a aprovação entre os estudantes que tem aula com professores substitutos é 91,6% maior do que a aprovação entre os alunos que tem aula com professores do quadro.

A modelagem estatística para os outros casos foi feita seguindo o mesmo procedimento do modelo de Estatística Aplicada definido anteriormente. Por essa razão, o passo a passo será omitido e serão apresentadas apenas as tabelas com o resumo dos modelos.

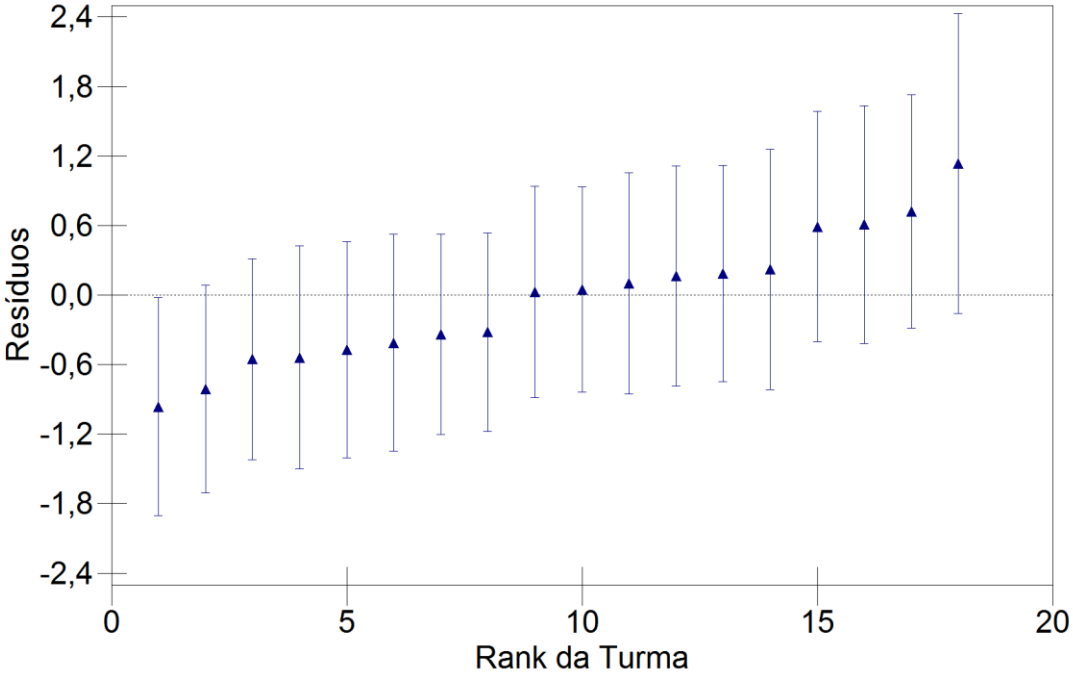
Estatística Aplicada (desconsiderando os alunos com menção SR)

Com o objetivo de fazer a modelagem da aprovação para os estudantes que não obtiveram menção SR, reduziu-se de 1076 para 971 o universo dos alunos que foram considerados no modelo, o que mostra que no 1º e no 2º semestres de 2008, 105 estudantes ficaram com SR na disciplina de Estatística Aplicada.

Tabela 10 - Modelagem Estatística - Estatística Aplicada - Sem SR's
Coeficiente de Correlação Intraclasse = 0,343
Variável do Nível do Estudante: Sexo e Modalidade da Disciplina
Variável do Nível da Turma: Situação do Professor
Modelo Final: logito (π_{ik}) = 1,471 + 0,709 $feminino_{ik}$ + 0,600 $optativa_{ik}$ – 2,423 $módulo\ livre_{ik}$ + 0,551 $substituto_k$ + e_{ik} + u_{0k}
Média geral de aprovação = 81,3% (66,1% a 90,7%)
Razão de chances (feminino) = 2,032 (1,359 a 3,037)
Razão de chances (optativa) = 1,822 (0,584 a 5,680)
Razão de chances (módulo livre) = 0,089 (0,021 a 0,378)
Razão de chances (substituto) = 1,735 (0,694 a 4,336)

Análise de Resíduos

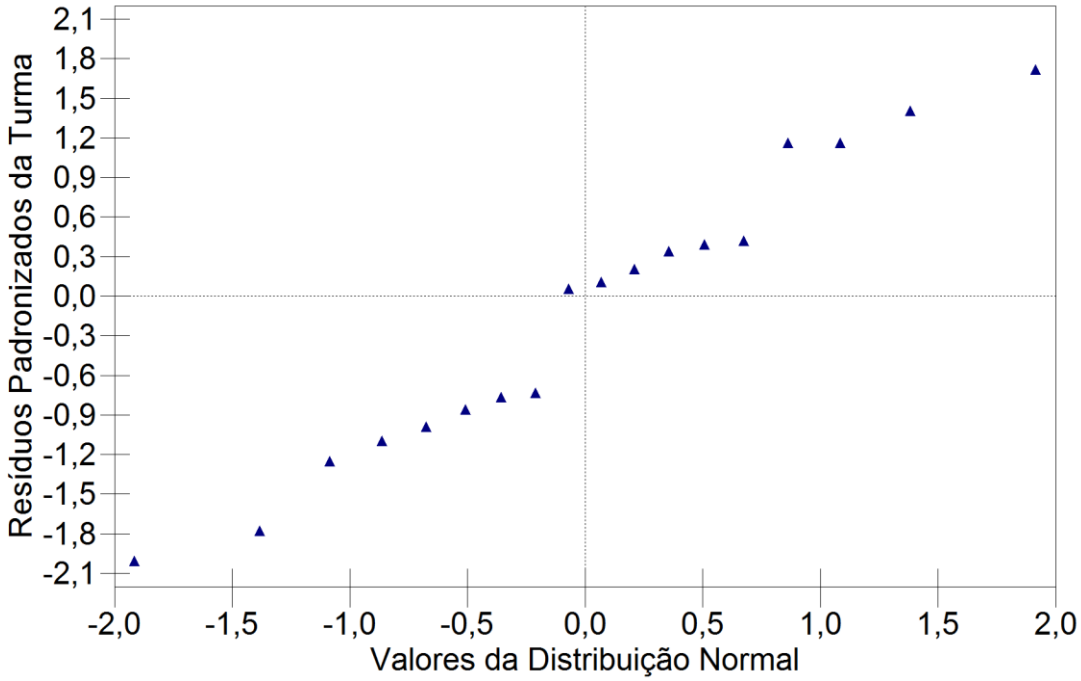
Figura 7 - Resíduos do Nível da Turma



A Figura 7 apresenta os resíduos plotados em ordem crescente de magnitude com seus respectivos intervalos de confiança. O intervalo que intercepta o zero, mostra que o desempenho daquela turma não é significativamente diferente do desempenho global das turmas. Se o intervalo

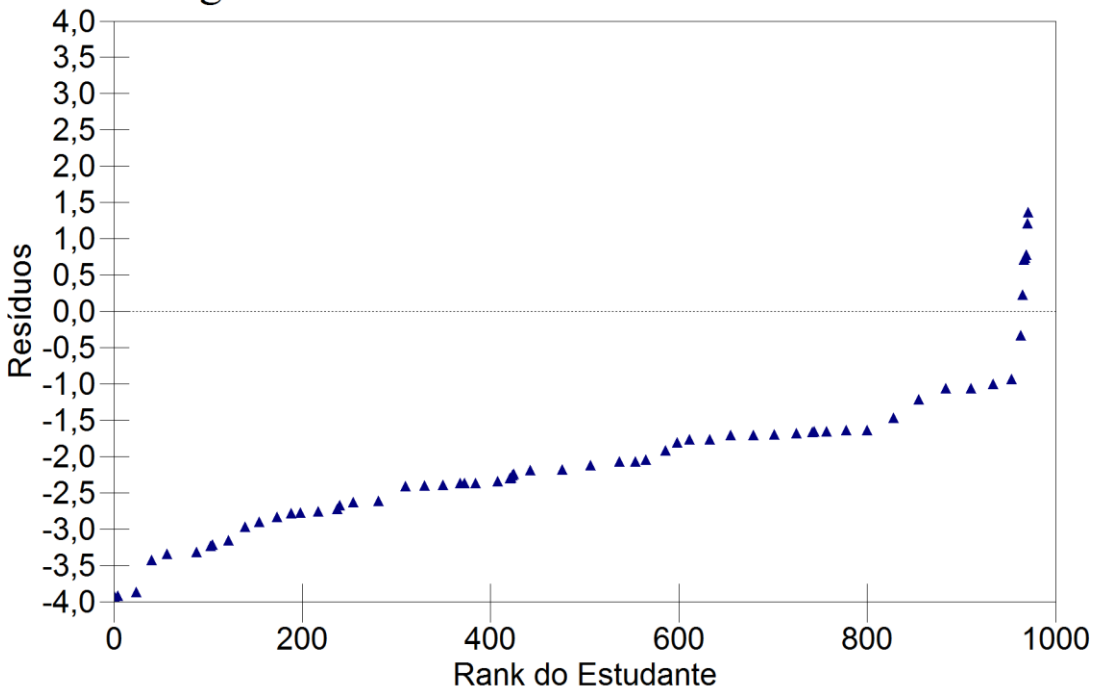
de confiança é inteiramente abaixo da linha pontilhada, a aprovação dos estudantes é significativamente menor para essa turma, situação que ocorre na turma E do 2º semestre de 2008, entretanto, é possível perceber que o desempenho desta turma não chega a ser tão diferente das demais. Já, se o intervalo de confiança é totalmente acima da linha pontilhada, a aprovação dos alunos é significativamente maior para aquela turma, essa situação não ocorre em nenhuma das turmas de Estatística Aplicada.

Figura 8 - Resíduos Padronizados do Nível da Turma em relação aos respectivos Valores da Distribuição Normal



Na Figura 8, têm-se os resíduos padronizados plotados em relação aos valores da distribuição Normal, para se verificar o pressuposto da normalidade no nível da turma, os resíduos deveriam estar distribuídos ao longo de uma linha reta. Entretanto, apesar de isso não ocorrer, é possível observar que a violação ocorre mais intensamente nas observações iniciais e não chega a inviabilizar a utilização do modelo escolhido anteriormente.

Figura 9 - Resíduos do Nível do Estudante



Os pontos cruciais a serem verificados no diagnóstico do modelo considerando o nível do aluno são observar se a função estimada é monotônica e se tem forma sigmoidal. A partir da Figura 9, é possível analisar que esses pressupostos são atendidos.

Interpretação dos Resultados

Após o diagnóstico do modelo, é possível interpretar os resultados do modelo escolhido:

$$\text{logito}(\pi_{ik}) = 1{,}471 + 0{,}709\ \text{feminino}_{ik} + 0{,}600\ \text{optativa}_{ik} - 2{,}423\ \text{módulo livre}_{ik} + 0{,}551\ \text{substituto}_{k} + e_{ik} + u_{0k} \quad (50)$$

A média geral de aprovação na escala logito é igual a 1,471, convertendo esse valor para probabilidade, tem-se que 81,3% dos alunos foram aprovados na disciplina de Estatística Aplicada no conjunto das turmas, considerando um intervalo com 95% de confiança tem-se que o percentual de estudantes aprovados na disciplina está entre 66,1% e 90,7%.

Em relação à variável sexo, tem-se que a razão de chances é igual a 2,032, ou seja, a chance de aprovação entre os estudantes do sexo feminino é 2,032 vezes a chance de aprovação entre os alunos do sexo masculino, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 1,359 a 3,037. Assim, tem-se que a aprovação entre os estudantes do sexo feminino é 103,2% maior do que a aprovação entre os alunos do sexo masculino.

Considerando a categoria optativa, em relação à variável modalidade da disciplina, tem-se que a razão de chances é igual a 1,822, ou seja, a chance de aprovação dos que cursam a disciplina como optativa é 1,822 vezes a chance de aprovação dos que cursam a disciplina como obrigatória, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,584 a 5,680. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina como optativa é 82,2% maior do que a aprovação entre os alunos que a cursam como obrigatória.

Considerando a categoria módulo livre, em relação à variável modalidade da disciplina, tem-se que a razão de chances é igual a 0,089, ou seja, a chance de aprovação dos que cursam a disciplina como módulo livre é 0,089 vezes a chance de aprovação dos que cursam a disciplina como obrigatória, considerando um intervalo com 95% de confiança tem-se que a

razão de chances varia de 0,021 a 0,378. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina como obrigatória é 91,1% maior do que a aprovação entre os alunos que a cursam como módulo livre.

Em relação à variável situação do professor, tem-se que a razão de chances é igual a 1,735, ou seja, a chance de aprovação tendo aula com o professor substituto é 1,735 vezes a chance de aprovação tendo aula com o professor do quadro, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,694 a 4,336. Assim, tem-se que a aprovação entre os estudantes que tem aula com professores substitutos é 73,5% maior do que a aprovação entre os alunos que tem aula com professores do quadro.

Probabilidade e Estatística

Como o coeficiente de correlação intraclasse é igual a 0,11, e esse valor é próximo de zero, isso significa que as turmas são homogêneas entre si e que o desempenho do aluno independe da turma a que ele pertence. Assim, as características individuais dos estudantes tendem a definir o seu próprio desempenho.

Como o valor do coeficiente de correlação intraclasse não justifica o uso da abordagem multinível, não será dada continuidade à modelagem da aprovação para os alunos de Probabilidade e Estatística.

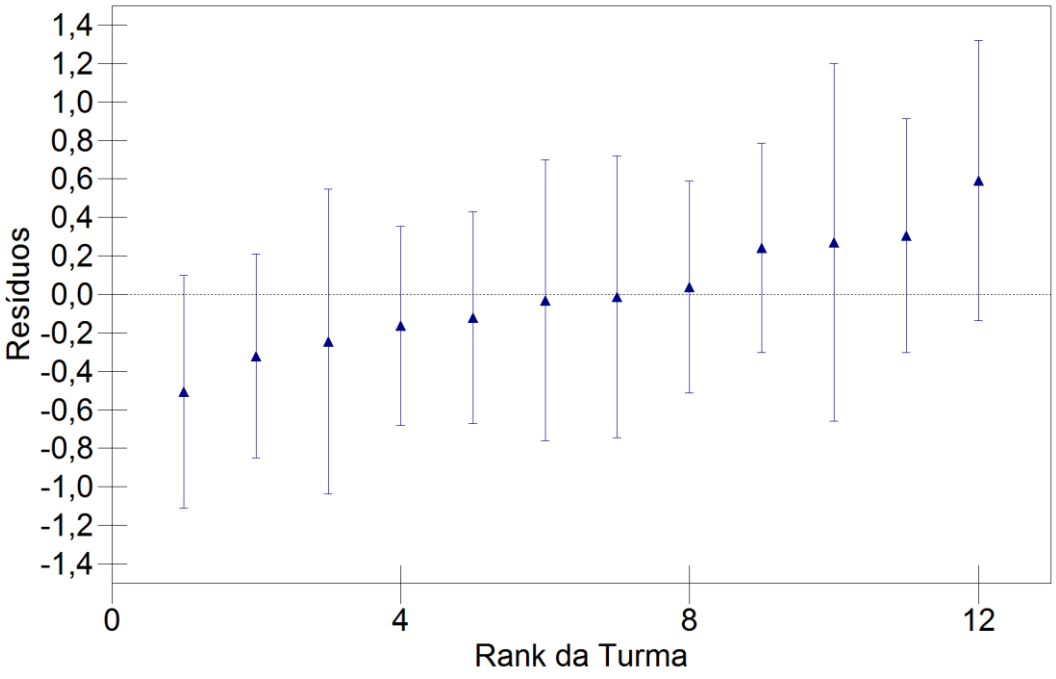
Probabilidade e Estatística (desconsiderando os alunos com menção SR)

Com o objetivo de fazer a modelagem da aprovação para os estudantes que não obtiveram menção SR, reduziu-se de 679 para 628 o universo dos alunos que foram considerados no modelo, o que mostra que no 1º e no 2º semestres de 2008, 51 estudantes ficaram com SR na disciplina de Probabilidade e Estatística.

Tabela 11 - Modelagem Estatística - Probabilidade e Estatística - Sem SR's
Coeficiente de Correlação Intraclasse = 0,23
Variável do Nível do Estudante: Forma de Saída
Variável do Nível da Turma: Turno
Modelo Final: logito (π_{ik}) = 1,134 – 0,069 outras formas de saída$_{ik}$ + 0,956 ambos os turnos$_k$ + e$_{ik}$ + u$_{0k}$
Média geral de aprovação = 75,7% (68,0% a 82,0%)
Razão de chances (saiu da graduação) = 0,933 (0,890 a 0,978)
Razão de chances (ambos os turnos) = 2,601 (1,141 a 5,930)

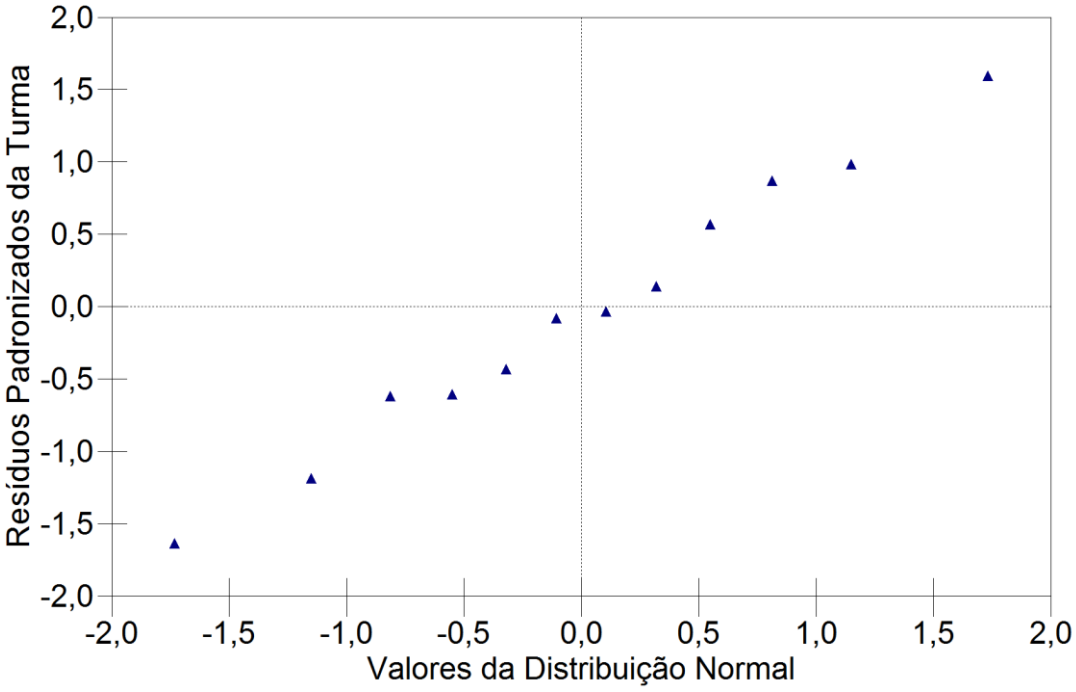
Análise de Resíduos

Figura 10 - Resíduos do Nível da Turma



A Figura 10 apresenta os resíduos plotados em ordem crescente de magnitude com seus respectivos intervalos de confiança. O intervalo que intercepta o zero, mostra que o desempenho daquela turma não é significativamente diferente do desempenho global das turmas. Se o intervalo de confiança é inteiramente abaixo da linha pontilhada, a aprovação dos estudantes é significativamente menor para essa turma. Já, se o intervalo de confiança é totalmente acima da linha pontilhada, a aprovação dos alunos é significativamente maior para aquela turma. Como não ocorre nenhuma dessas duas situações, tem-se que os estudantes dessas turmas tendem a apresentar um desempenho semelhante.

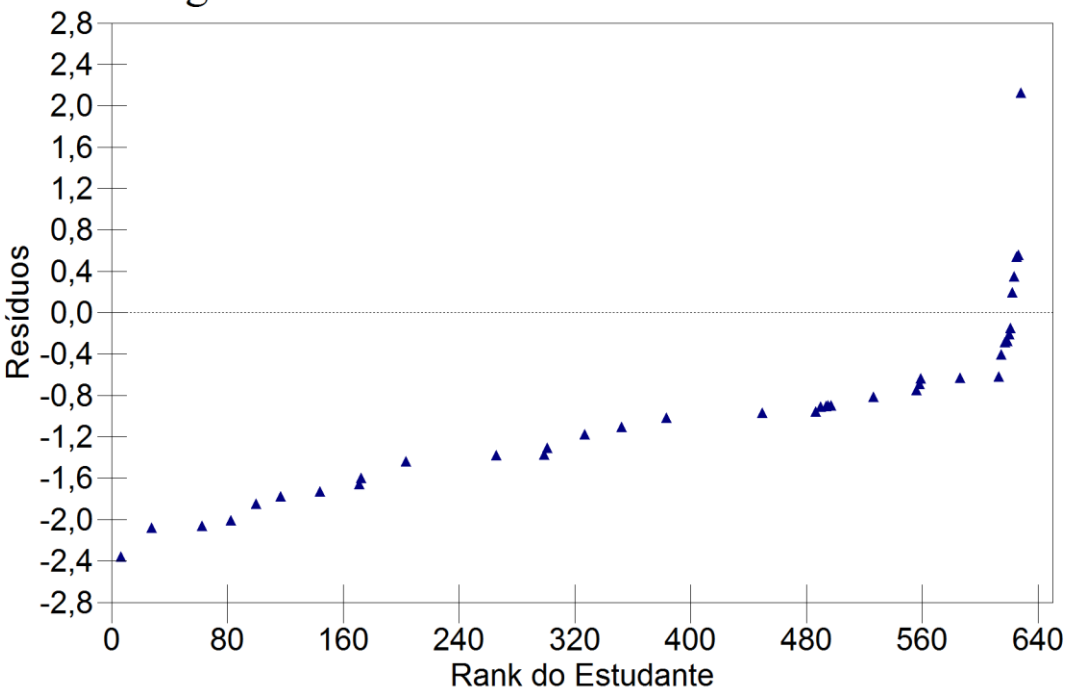
Figura 11 - Resíduos Padronizados do Nível da Turma em relação aos respectivos Valores da Distribuição NOrmal



A Figura 11 traz os resíduos padronizados plotados em relação aos valores da distribuição Normal, para se verificar o pressuposto da

normalidade no nível da turma, os resíduos deveriam estar distribuídos ao longo de uma linha reta. Entretanto, apesar de isso não ocorrer, é possível observar que a violação não chega a inviabilizar a utilização do modelo escolhido anteriormente.

Figura 12 - Resíduos do Nível do Estudante



Os pontos cruciais a serem verificados no diagnóstico do modelo considerando o nível do aluno são observar se a função estimada é monotônica e se tem forma sigmoidal. A partir da Figura 12, é possível analisar que esses pressupostos são atendidos.

Interpretação dos Resultados

Após o diagnóstico do modelo, é possível interpretar os resultados do modelo escolhido:

$$\text{logito}(\pi_{ik}) = 1{,}134 - 0{,}069\ \text{outras formas de saída}_{ik} + 0{,}956\ \text{ambos os turnos}_{k} + e_{ik} + u_{0k} \quad (54)$$

A média geral de aprovação na escala logito é igual a 1,134, convertendo esse valor para probabilidade, tem-se que 75,7% dos alunos foram aprovados na disciplina de Probabilidade e Estatística no conjunto das turmas, considerando um intervalo com 95% de confiança tem-se que o percentual de estudantes aprovados na disciplina está entre 68,0% e 82,0%.

Em relação à variável forma de saída, tem-se que a razão de chances é igual a 0,933, ou seja, a chance de aprovação de quem saiu da graduação (formatura, desligamento por rendimento, desligamento voluntário, novo vestibular e mudança de curso) é 0,933 vezes de quem ainda está cursando a graduação considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,890 a 0,978. Assim, tem-se que a aprovação entre os estudantes que ainda estão cursando a graduação é 6,7% maior do que a aprovação entre os alunos que saíram da graduação, mas não necessariamente se formaram.

Em relação à variável turno da turma, tem-se que a razão de chances é igual a 2,601, ou seja, a chance de aprovação dos estudantes que cursam a disciplina em uma turma destinada a ambos os turnos (diurno e noturno) é 2,601 vezes a chance de aprovação dos estudantes que cursam a disciplina em uma turma do turno diurno, considerando um intervalo com 95% de

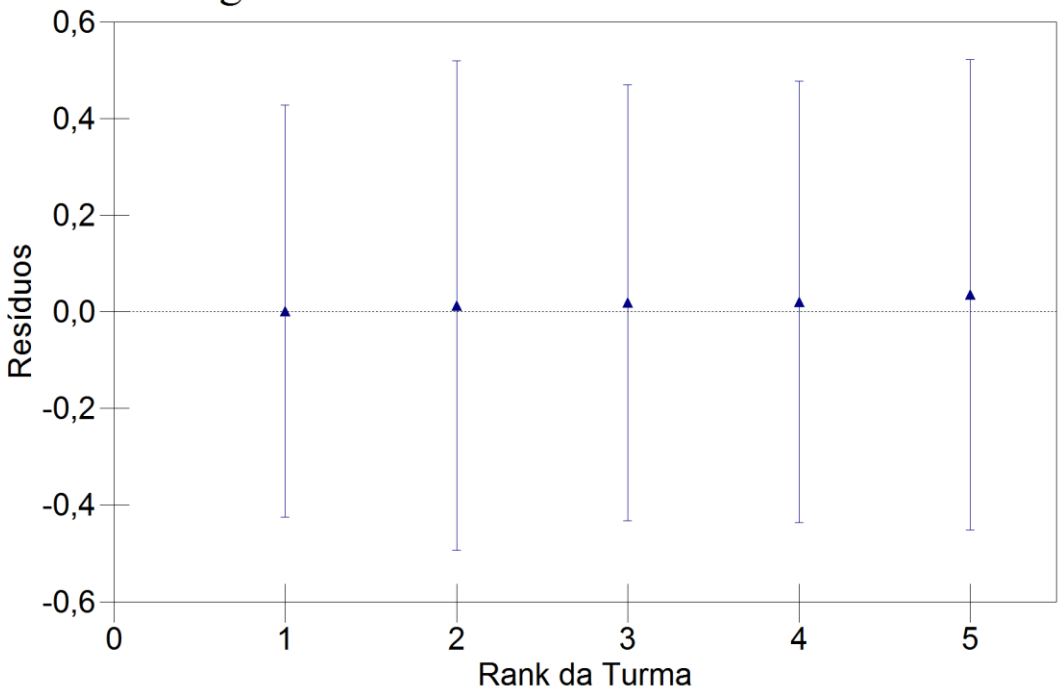
confiança tem-se que a razão de chances varia de 1,141 a 5,930. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina em uma turma destinada a ambos os turnos (diurno e noturno) é 160,1% maior do que a aprovação entre os alunos que cursam a disciplina em uma turma do turno diurno.

Bioestatística

Tabela 12 - Modelagem Estatística - Bioestatística
Coeficiente de Correlação Intraclasse = 0,583
Variável do Nível do Estudante: Forma de Saída
Variável do Nível da Turma: Professor
Modelo Final: logito (π_{ik}) = 1,829 – 0,179 outras formas de saída$_{ik}$ – 1,525 professor8$_k$ + 0,295 professo11$_k$ + 1,340 professor16$_k$ + e$_{ik}$ + u0$_k$
Média geral de aprovação = 86,2% (73,8% a 93,2%)
Razão de chances (saiu da graduação) = 0,836 (0,732 a 0,955)
Razão de chances (professor 8) = 0,218 (0,052 a 0,910)
Razão de chances (professor 11) = 1,343 (0,229 a 7,862)
Razão de chances (professor 16) = 3,819 (0,518 a 28,135)

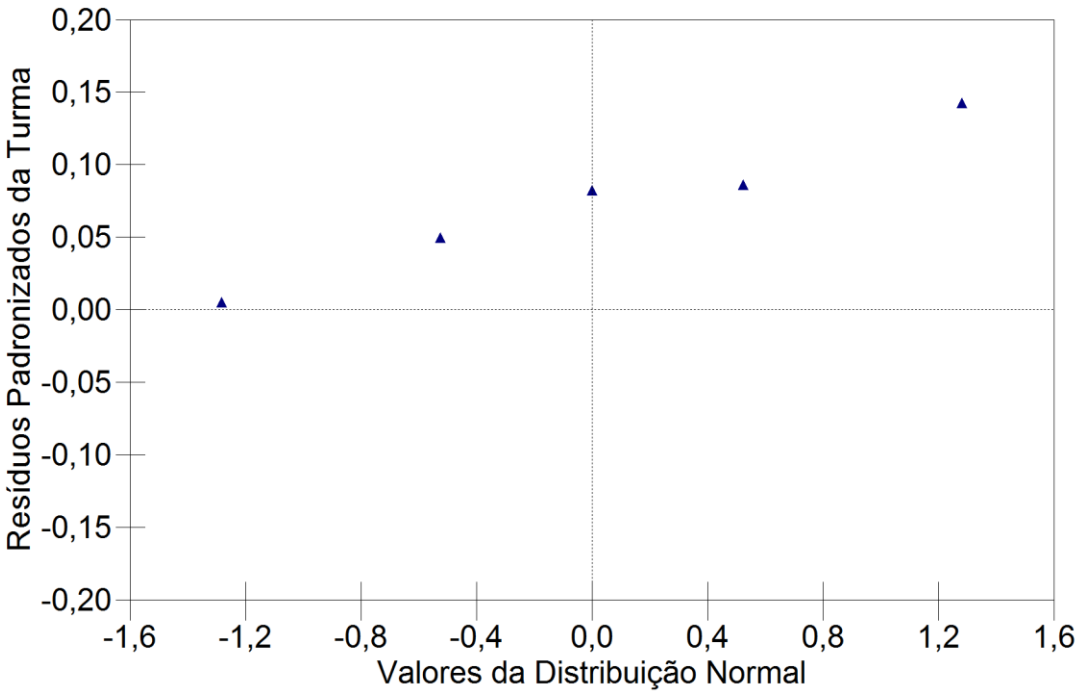
Análise de Resíduos

Figura 13 - Resíduos do Nível da Turma



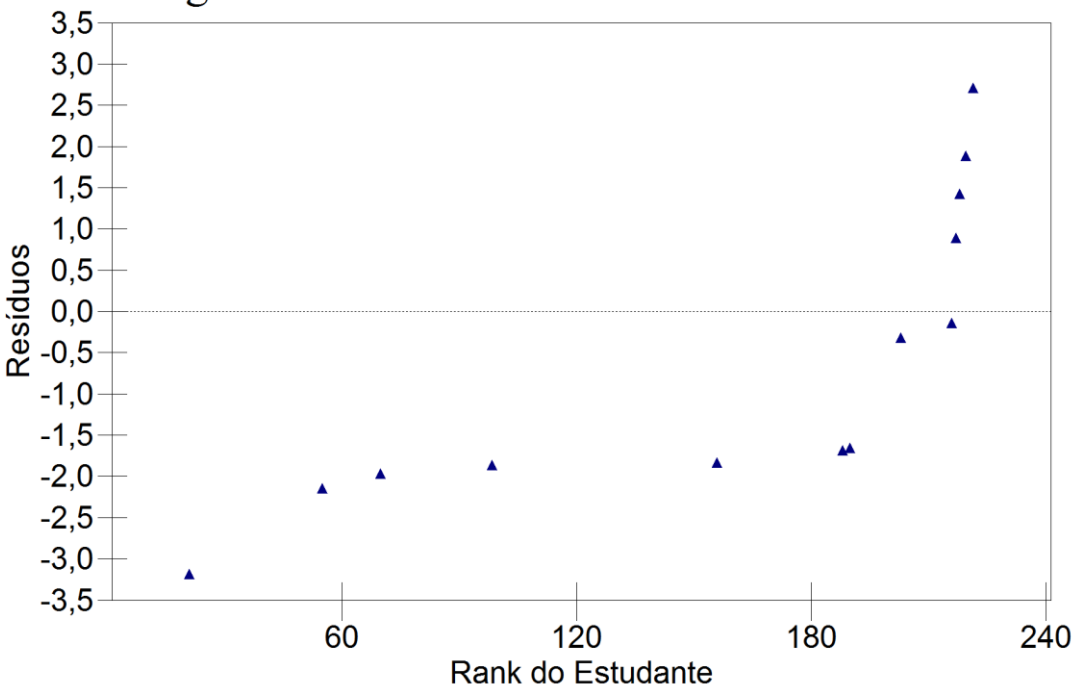
A Figura 13 apresenta os resíduos plotados em ordem crescente de magnitude com seus respectivos intervalos de confiança. O intervalo que intercepta o zero, mostra que o desempenho daquela turma não é significativamente diferente do desempenho global das turmas. Se o intervalo de confiança é inteiramente abaixo da linha pontilhada, a aprovação dos estudantes é significativamente menor para essa turma; já se o intervalo de confiança é totalmente acima da linha pontilhada, a aprovação dos alunos é significativamente maior para aquela turma. Como não ocorre nenhuma dessas duas situações, tem-se que os estudantes dessas turmas tendem a apresentar um desempenho semelhante.

Figura 14 - Resíduos Padronizados do Nível da Turma em relação aos respectivos Valores da Distribuição Normal



A Figura 14 traz os resíduos padronizados plotados em relação aos valores da distribuição Normal, para se verificar o pressuposto da normalidade no nível da turma, os resíduos deveriam estar distribuídos ao longo de uma linha reta. Entretanto, isso não ocorre, o que certamente se deve à existência de poucas observações, o fato de haver apenas cinco turmas de Bioestatística. Essa limitação no número de turmas, entretanto, não inviabilizará o uso do modelo definido anteriormente.

Figura 15 - Resíduos do Nível do Estudante



Os pontos cruciais a serem verificados no diagnóstico do modelo considerando o nível do aluno são observar se a função estimada é monotônica e se tem forma sigmoidal. A partir da Figura 15, é possível analisar que esses pressupostos são atendidos.

Interpretação dos Resultados

Após o diagnóstico do modelo, é possível interpretar os resultados do modelo escolhido:

$$\text{logito}(\pi_{ik}) = 1{,}829 - 0{,}179\ \text{outras formas de saída}_{ik} - 1{,}525\ \text{professor8}_k + 0{,}295\ \text{professor11}_k + 1{,}340\ \text{professor16}_k + e_{ik} + u_{0k} \quad (46)$$

A média geral de aprovação na escala logito é igual a 1,829, convertendo esse valor para probabilidade, tem-se que 86,2% dos alunos foram aprovados na disciplina de Bioestatística no conjunto das turmas, considerando um intervalo com 95% de confiança tem-se que o percentual de estudantes aprovados na disciplina está entre 73,8% e 93,2%.

Em relação à variável forma de saída, tem-se que a razão de chances é igual a 0,836, ou seja, a chance de aprovação de quem saiu da graduação (formatura, desligamento por rendimento e novo vestibular) é 0,836 vezes a chance de aprovação de quem ainda está cursando a graduação, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,732 a 0,955. Assim, tem-se que a aprovação entre os estudantes que ainda estão cursando a graduação é 16,4% maior do que a aprovação entre os alunos que saíram da graduação, mas não necessariamente se formaram.

Considerando o professor 8, em relação à variável professor, tem-se que a razão de chances é igual a 0,218, ou seja, a chance de aprovação tendo aula com o professor 8 é 0,218 vezes a chance de aprovação tendo aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,052 a 0,910. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 3 é 78,2% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 8.

Considerando o professor 11, em relação à variável professor, tem-se que a razão de chances é igual a 1,343, ou seja, a chance de aprovação tendo aula com o professor 11 é 1,343 vezes a chance de aprovação tendo

aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,229 a 7,862. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 11 é 34,3% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 3.

Considerando o professor 16, em relação à variável professor, tem-se que a razão de chances é igual a 3,819, ou seja, a chance de aprovação tendo aula com o professor 16 é 3,819 vezes a chance de aprovação tendo aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,518 a 28,135. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 16 é 281,9% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 3.

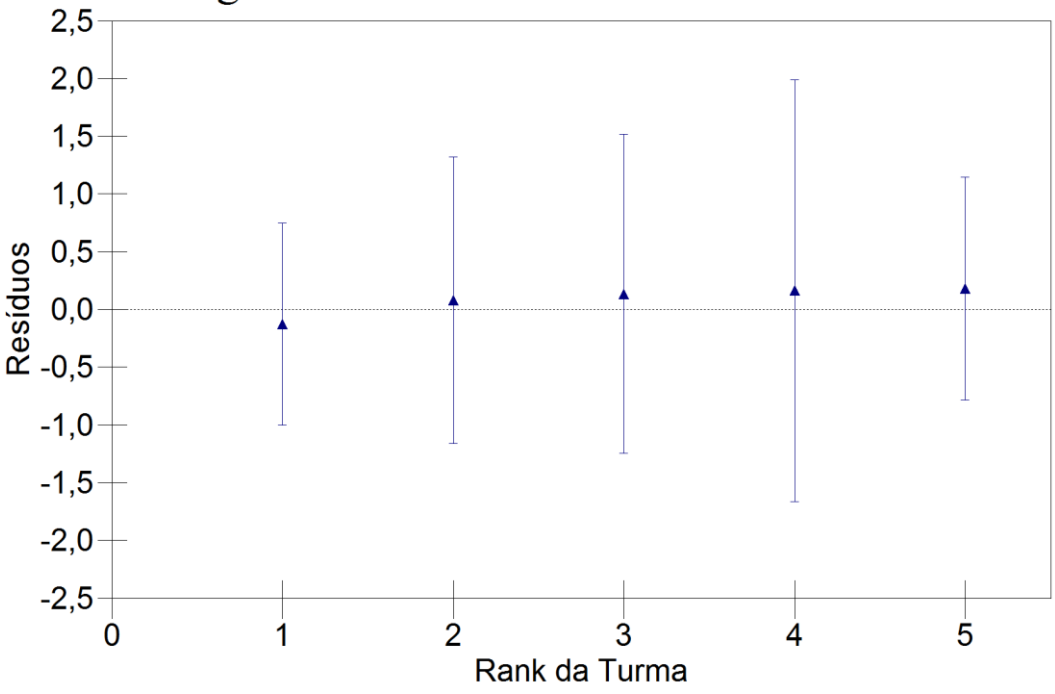
Bioestatística (desconsiderando os alunos com menção SR)

Com o objetivo de fazer a modelagem da aprovação para os estudantes que não obtiveram menção SR, reduziu-se de 222 para 210 o universo dos alunos que foram considerados no modelo, o que mostra que no 1º e no 2º semestres de 2008, 11 estudantes ficaram com SR na disciplina de Bioestatística.

Tabela 13 - Modelagem Estatística - Bioestatística - Sem SR's
Coeficiente de Correlação Intraclasse = 0,633
Variável do Nível do Estudante: Forma de Saída
Variável do Nível da Turma: Professor
Modelo Final: logito (π_{ik}) = 2,255 – 0,149 outras formas de saída$_{ik}$ – 1,607 professor 8$_k$ – 0,194 professor 11$_k$ + 1,815 professor 16$_k$ + e$_{ik}$ + u$_{0k}$
Média geral de aprovação = 90,5% (78,6% a 96,1%)
Razão de chances (saiu da graduação) = 0,862 (0,743 a 0,999)
Razão de chances (professor 8) = 0,200 (0,033 a 1,225)
Razão de chances (professor 11) = 0,824 (0,104 a 6,521)
Razão de chances (professor 16) = 6,141 (0,204 a 184,565)

Análise de Resíduos

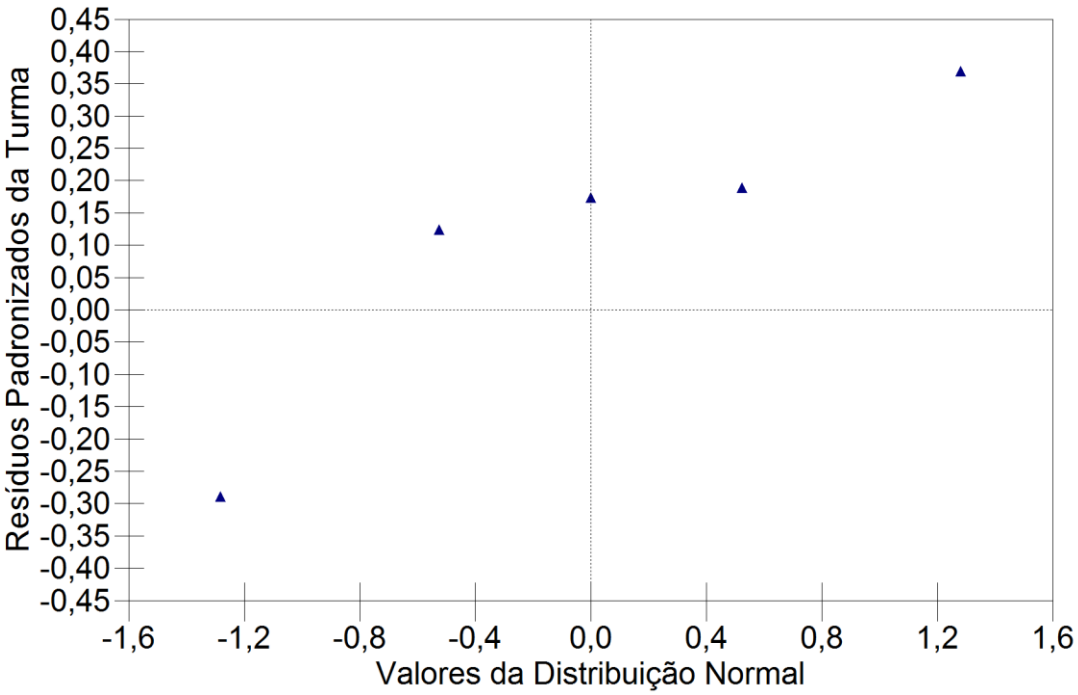
Figura 16 - Resíduos do Nível da Turma



A Figura 16 apresenta os resíduos plotados em ordem crescente de magnitude com seus respectivos intervalos de confiança. O intervalo que intercepta o zero, mostra que o desempenho daquela turma não é

significativamente diferente do desempenho global das turmas. Se o intervalo de confiança é inteiramente abaixo da linha pontilhada, a aprovação dos estudantes é significativamente menor para essa turma; já se o intervalo de confiança é totalmente acima da linha pontilhada a aprovação dos alunos é significativamente maior para aquela turma. Como não ocorre nenhuma dessas duas situações, tem-se que os estudantes dessas turmas tendem a apresentar um desempenho semelhante.

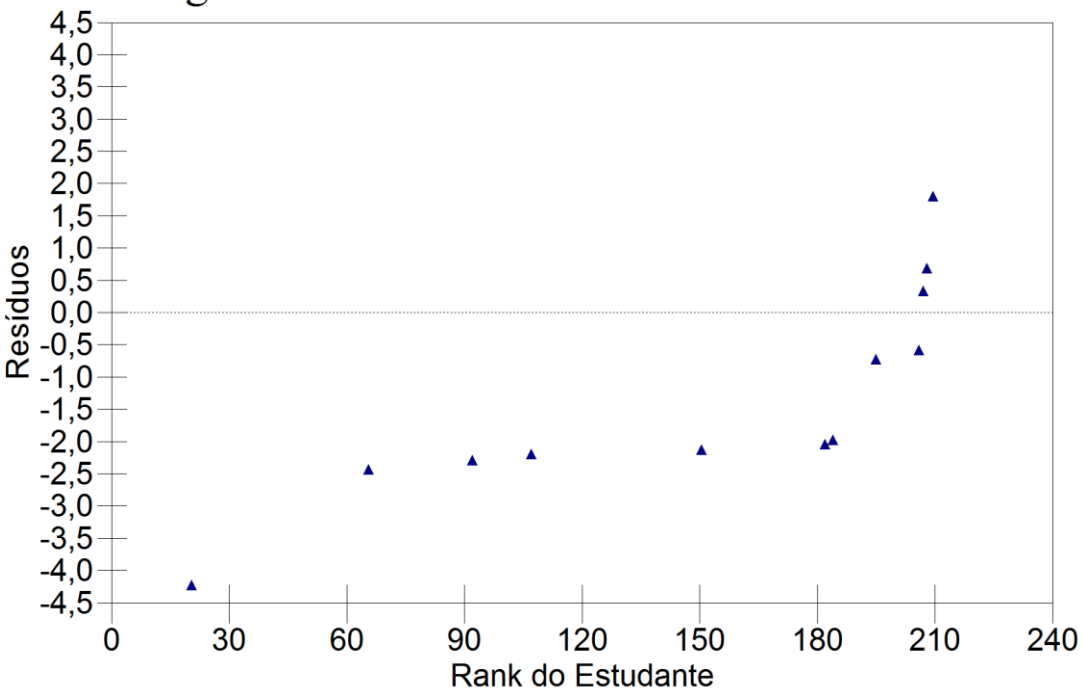
Figura 17 - Resíduos Padronizados do Nível da Turma em relação aos respectivos Valores da Distribuição Normal



A Figura 17 traz os resíduos padronizados plotados em relação aos valores da distribuição Normal, para se verificar o pressuposto da normalidade no nível da turma, os resíduos deveriam estar distribuídos ao longo de uma linha reta. Entretanto, isso não ocorre, o que certamente se deve à existência de poucas observações, o fato de haver apenas cinco

turmas de Bioestatística. Essa limitação no número de turmas, entretanto, não inviabilizará o uso do modelo definido anteriormente.

Figura 18 - Resíduos do Nível do Estudante



Os pontos cruciais a serem verificados no diagnóstico do modelo considerando o nível do aluno são observar se a função estimada é monotônica e se tem forma sigmoidal. A partir da Figura 18, é possível analisar que esses pressupostos são atendidos.

Interpretação dos Resultados

Após o diagnóstico do modelo, é possível interpretar os resultados do modelo escolhido:

$$\text{logito}(\pi_{ik}) = 2{,}255 - 0{,}149\ \text{outras formas de saída}_{ik} - 1{,}607\ \text{professor } 8_k - 0{,}194\ \text{professor } 11_k + 1{,}815\ \text{professor } 16_k + e_{ik} + u_{0k} \quad (58)$$

A média geral de aprovação na escala logito é igual a 1,829, convertendo esse valor para probabilidade, tem-se que 90,5% dos alunos foram aprovados na disciplina de Bioestatística no conjunto das turmas, considerando um intervalo com 95% de confiança tem-se que o percentual de estudantes aprovados na disciplina está entre 78,6% e 96,1%.

Em relação à variável forma de saída, tem-se que a razão de chances é igual a 0,862, ou seja, a chance de aprovação de quem saiu da graduação (formatura, desligamento por rendimento e novo vestibular) é 0,862 vezes a chance de aprovação de quem ainda está cursando a graduação, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,743 a 0,999. Assim, tem-se que a aprovação entre os estudantes que ainda estão cursando a graduação é 13,8% maior do que a aprovação entre os alunos que saíram da graduação, mas não necessariamente se formaram.

Considerando o professor 8, em relação à variável professor, tem-se que a razão de chances é igual a 0,200, ou seja, a chance de aprovação tendo aula com o professor 8 é 0,200 vezes a chance de aprovação tendo aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,033 a 1,225. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 3 é 80,0% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 8.

Considerando o professor 11, em relação à variável professor, tem-se que a razão de chances é igual a 0,824, ou seja, a chance de aprovação tendo aula com o professor 11 é 0,824 vezes a chance de aprovação tendo aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,104 a 6,521. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 3 é 17,6% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 11.

Considerando o professor 16, em relação à variável professor, tem-se que a razão de chances é igual a 6,141, ou seja, a chance de aprovação tendo aula com o professor 16 é 6,141 vezes a chance de aprovação tendo aula com o professor 3, considerando um intervalo com 95% de confiança tem-se que a razão de chances varia de 0,204 a 184,565. Assim, tem-se que a aprovação entre os estudantes que cursam a disciplina com o professor 16 é 514,1% maior do que a aprovação entre os alunos que cursam a disciplina como o professor 3.

Conclusão

Com o presente trabalho, é possível perceber que o modelo Multinível é adequado para modelar a aprovação dos estudantes de Estatística Aplicada e de Bioestatística, visto que, respectivamente, 40% e 60% da variância no desempenho desses alunos pode ser atribuída à turma a que eles pertencem. No caso dos estudantes de Probabilidade e Estatística, apenas 10% da variância no desempenho dos mesmos é devido à turma de que eles fazem parte. Logo, as características individuais dos estudantes é que tendem a definir o seu próprio desempenho. Por essa razão a abordagem Multinível não é adequada para este caso.

No caso do modelo da aprovação desconsiderando os alunos que obtiveram menção SR, é possível perceber que o modelo Multinível é adequado para modelar a aprovação dos estudantes de todas as disciplinas, visto que nas disciplinas de Estatística Aplicada, Probabilidade e Estatística e Bioestatística, respectivamente, 35%, 23% e 63% da variância no desempenho desses alunos pode ser atribuída à turma que eles pertencem.

No caso das disciplinas de Estatística Aplicada e Bioestatística, a modelagem da aprovação dos estudantes desconsiderando os alunos que obtiveram menção SR apresentou um resultado semelhante à modelagem considerando todos os estudantes.

No caso da disciplina de Probabilidade e Estatística, tem-se que a retirada dos alunos que obtiveram menção SR resultou em uma maior variabilidade no desempenho dos estudantes devido às diferenças existentes

entre as turmas, o que possibilitou o uso do Modelo Multinível para explicar a aprovação desses alunos.

A aprovação dos alunos de Estatística Aplicada é explicada, principalmente, pelas variáveis sexo, modalidade da disciplina e situação professor da turma (quadro/substituto).

A aprovação dos alunos de Probabilidade e Estatística é explicada, principalmente, pelas variáveis forma de saída e turno da turma.

A aprovação dos alunos de Bioestatística é explicada, principalmente, pelas variáveis forma de saída e professor da turma. Uma justificativa possível para o fato de a modelagem desta disciplina não ter ficado tão adequada se deve à ocorrência de situações distintas que ocorreram no 1º e no 2º semestres de 2008. Enquanto no 1º semestre de 2008, a turma foi significativa, havendo assim, um indicativo de que o desempenho dos estudantes era diferente ao se considerar cada uma delas, o mesmo não aconteceu no 2º semestre de 2008.

Uma sugestão é que seja dada continuidade a este trabalho utilizando dados mais atuais, pois, dessa forma, será possível avaliar o papel do Departamento de Estatística na formação dos estudantes de outras áreas e também refletir sobre a elaboração de ações que venham a contribuir de forma significativa para que os estudantes tenham um desempenho mais satisfatório nessas disciplinas.

A utilização do MLwiN limitou um pouco a modelagem, visto que não foi possível definir um modelo de coeficientes aleatórios, pois o método de estimação utilizado não aceitou o fato de a matriz de variância não ser

positiva definida. Assim, fica a sugestão de se utilizar o SAS também para a modelagem em estudos futuros.

Referências Bibliográficas

AGRESTI, Alan. **An introduction to categorical data analysis.** New York: John Wiley & Sons, 1996. 290 p.

FERRÃO, Maria Eugénia. **Introdução aos modelos de regressão multinível em educação.** Campinas, SP: Komedi, 2003. 106p. (Coleção Avaliação construindo o campo e a crítica)

GELMAN, Andrew; HILL, Jennifer. **Data analysis using regression and multilevel/hierarchical models.** Cambridge: Cambridge University Press, 2007. 625 p. (Analytical methods for social research)

HOX, J. J.. **Multilevel analysis: techniques and applications.** Mahwah, NJ: Lawrence Erlbaum Associates, 2002. 304 p.

KREFT, Ita; LEEUW, Jan de. **Introducing multilevel modeling.** London: Sage Publications, 1998. 149 p.

LAROS, J. A.; MARCIANO, João Luiz Pereira. Análise multinível aplicada a dados do NELS:88. **Estudos em Avaliação Educacional**, v. 19, p. 263-278, 2008.

NETER, John; WASSERMAN, William. **Applied linear statistical models: Regression, analysis of variance, and experimental designs.** Homewood:

R D Irwin 842 p.

REISE, Steven Paul; DUAN, Naihua (Coord.) **Multilevel modeling: methodological advances, issues, and applications.** Mahwah, NJ: Lawrence Erlbaum Associates, 2003. 314 p. (Multivariate applications book series)