Universidade de Brasília
Departamento de Estatística

Análise do desempenho acadêmico dos alunos na disciplina de Estatística Aplicada da Universidade de Brasília

Helena Santos Brandão

Brasília
2021

Helena Santos Brandão

Análise do desempenho acadêmico dos alunos na disciplina de Estatística Aplicada da Universidade de Brasília

Orientadora: Profª. Maria Teresa Leão Costa

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

Brasília
2021

Agradecimentos

Gostaria de agradecer a minha família - especialmente aos meu pais por todo o apoio emocional e acadêmico.

Agradeço também a meu namorado João por todo seu suporte; às minha amigas desde o começo do curso de Estatística na Universidade: Carolina, Juliana, Yasmin, Luisa e Maria Eduarda; ao meu amigo Thiago que também ficou presente comigo durante o curso ao longo de várias horas de estudo.

À minha tia Neca por te sido tão carinhosa e gentil comigo.

Aos meus filhotinhos Atena e Monet que ficaram perto de mim dando sempre muito amor.

Ao meu supervisor de estágio Thiago que me deu a oportunidade de aprender muito sobre programação e que também me deu suporte durante todo o processo de aprendizagem no estágio.

Agradeço também a todos que ajudaram de alguma forma em minha trajetória acadêmica e pessoal até o presente momento.

Por fim, quero agradecer a todos os professores que contribuiram com meu processo de aprendizagem até o presente momento e em especial a minha orientadora Maria Teresa por ter aceitado me auxiliar nesse projeto e por ter me ensinado tanto desde meu segundo semestre na Universidade.

Resumo

O presente estudo refere-se a análise de fatores que influenciam no resultado final de estudantes na disciplina Estatística Aplicada da Universidade de Brasília durante sua primeira tentativa completa na disciplina. Análises exploratórias referentes a dados de estudantes de 1994 até 2019 foram feitas; foram filtrados ainda dados relativos a 2018 e 2019 a fim de criar modelos para explicar o comportamento dos dados. Verificou-se também - como um estudo exploratório - a influência de variáveis sociodemográficas nos resultados dos discentes. Técnicas de Regressão Logística Multinível foram utilizadas no processo de modelagem, considerando-se que a variável resposta é binária (aprovação ou reprovação do aluno) e os diferentes níveis hierárquicos no banco de dados utilizado. Diferenças entre resultados segundo a modalidade na qual o estudante cursa a disciplina, o tipo de professor, a porcentagem de faltas do aluno na matéria e o tempo que o aluno passou na Universidade até cursar a matéria apresentaram resultados significativos; observou-se ainda que algumas variáveis socio-econômicas também aparentam influenciar na variável resposta.

Palavras-chave: Estatística, Regressão Logística, Regressão Multinível, Estatística Aplicada, Universidade de Brasília, Rendimento Acadêmico.

Abstract

The present study refers to the analysis of factors that influence the final result of students in the Applied Statistics course at the University of Brasília during their first full attempt in the course. Exploratory analyzes referring to student data from 1994 to 2019 were carried out; data for 2018 and 2019 were also filtered in order to create models to explain the behavior of the data. It was also verified - as an exploratory study - the influence of sociodemographic variables on the students' results. Multilevel Logistic Regression techniques were used in the modeling process - considering that the response variable is binary (student approval or failure) and the different hierarchical levels in the database used. Differences between results according to the modality in which the student attends the subject, the type of teacher, the percentage of absences by the student in the subject and the time that the student spent at the University until taking the subject presented significant results; it was also observed that some socioeconomic variables also seem to influence the response variable.

Keywords: Statistics, Logistic Regression, Multilevel Regression, Applied Statistics, University of Brasília, Academic Performance.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O conhecimento sobre conceitos e técnicas estatísticas mostra-se cada vez mais relevante na atualidade para profissionais de diversas áreas - tanto pelo avanço tecnológico, quanto pela emergência de novas análises de dados em áreas e assuntos variados. Neste contexto, o currículo de vários cursos de graduação contêm uma disciplina de Estatística como obrigatória ou optativa. Atendendo tal demanda, o departamento de Estatística da Universidade de Brasília (UnB) - além de ofertar as disciplinas pertinentes ao curso de bacharelado em Estatística - oferta para os demais cursos da Universidade as disciplinas: Estatística Aplicada, Probabilidade e Estatística e Bioestatística.

O departamento oferta a disciplina Estatística Aplicada desde 1974 a diversos cursos da Universidade (principalmente das áreas de Humanas e Ciências Sociais) - sendo atualmente parte obrigatória da componente curricular dos cursos de Administração, Biblioteconomia, Ciência Política, Ciências Ambientais, Ciências Contábeis, Ciências Sociais, Geografia, Gestão do Agronegócio, Psicologia e Relações Internacionais. Alguns cursos possuem ainda a matéria como componente optativo de sua base curricular e os demais alunos da universidade podem fazer a disciplina como Módulo Livre [1]. Atualmente são ofertadas cerca de 11 turmas - com mais de 750 vagas para os estudantes da universidade - a cada semestre.

A disciplina não possui outras disciplinas como pré-requisito para ser cursada - sendo apenas necessário os conhecimentos do Ensino Médio dos estudantes - e tem como objetivo principal introduzir conceitos estatísticos à alunos de determinados cursos de modo a familiariza-los com essas ideias para possíveis situações futuras em sua atuação profissional (contando para isso com uma carga horária de 90 horas de aulas presenciais). De modo geral, essa disciplina representa uma das poucas matérias na qual os estudantes terão contato com conceitos estatísticos e matemáticos durante sua formação acadêmica. Analisar, então, os fatores que podem contribuir para a aprovação dos alunos nessa disciplina - verificando, ainda, se há diferenças entre seus rendimentos - mostra-se importante para contribuir com o processo de ensino-aprendizagem da matéria e, consequentemente, para melhorar o desenvolvimento de formação profissional desses estudantes.

De fato, um estudo sobre o rendimento acadêmico desses alunos apresenta-se apropriado para satisfazer necessidades e objetivos da própria instituição, visto que a ampliação do conhecimento da Universidade sobre si mesma e seus estudantes revela-se necessária a fim de garantir suas funções científicas e sociais (VENDRAMINI et al, 2004).

[1] As disciplinas de módulo livre de um curso são todas as disciplinas de graduação que não são de abrangência restrita e que não constam no currículo do curso. Os créditos a integralizar em módulo livre são referentes às disciplinas ou atividades que não estão na lista de disciplinas obrigatórias nem de optativas do seu curso, porém estão previstas e oferecidas pela UnB - podendo somar ao total de créditos exigidos para o curso, desde que estejam no limite máximo permitido pelo respectivo currículo.

Além disso, é essencial que a Universidade reveja seus métodos, objetivos e metodologias de aprendizagem, de modo a acompanhar as demandas emergentes da sociedade (CUNHA e CARRILHO, 2005).

Tendo isso em consideração, a importância da análise do rendimento dos estudantes nessa disciplina mostra-se notória - observando-se, ainda, que a Universidade de Brasília exerce um papel relevante na formação de indivíduos tanto no Distrito Federal, quanto em outros estados, de modo a ser uma referência de ensino em diversas regiões do país.

O presente estudo visa, portanto, analisar os fatores que podem estar associados a aprovação ou reprovação do estudante na disciplina Estatística Aplicada com base em um banco de dados com informações acerca de seu resultado e algumas informações a respeito do estudante e sobre a turma em que cursou a disciplina. Técnicas de Regressão Logística Multinível foram empregadas para auxiliar no desenvolvimento de tais análises.

2 Objetivos

2.1 Objetivo Geral

Analisar o desempenho acadêmico dos alunos na disciplina de Estatística Aplicada da Universidade de Brasília - identificando fatores associados a sua aprovação na disciplina.

2.2 Objetivos Específicos

- Descrever o perfil dos estudantes de Estatística Aplicada;
- Analisar o rendimento dos estudantes de Estatística Aplicada ao longo do tempo e segundo características dos estudantes e oferta da disciplina;
- Construir um modelo de regressão para descrever como se dá o processo de aprovação dos alunos - identificando fatores associados a tais resultados - na disciplina de Estatística Aplicada em sua primeira tentativa na matéria.

3 Regressão Logística

Em diversos estudos e pesquisas, deseja-se explicar o comportamento de uma variável nominal com apenas duas categorias - por exemplo ao se analisar a aprovação ou reprovação de alunos em determinada disciplina a partir de um conjunto de variáveis explicativas. Nesse caso, a variável resposta é considerada binária e são atribuídos as codificações 0 e 1 a cada uma das categorias em questão (a escolha da categoria a ser considerada como 1 depende de qual característica é a mais importante no estudo em questão).

3.1 Função de regressão quando a variável resposta é binária

Seja:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

onde $Y_i = 0, 1$; ϵ_i representa o erro aleatório associado a tal medida e β_0 e β_1 representam os parâmetros da função de regressão.

Uma vez que $E(\epsilon_i) = 0$, pode-se notar que:

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

Assumindo que Y_i é uma variável aleatória Bernoulli (com $P(Y_i = 1) = \pi_i$, isto é, a probabilidade de ocorrer o evento de sucesso):

$$E(Y_i) = \pi_i + 0(1 - \pi_i) = \pi_i = \beta_0 + \beta_1 X_i.$$

A Figura 1 representa a propabilidade da variável resposta Y ser classificada como (1) com base na variável explicativa X.

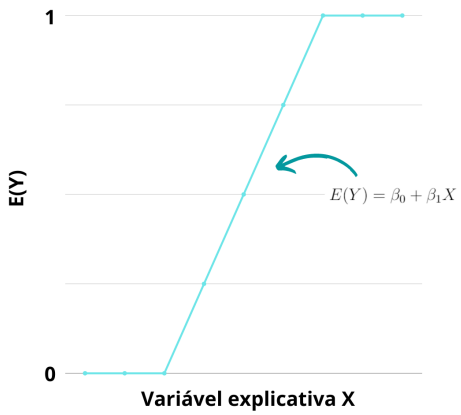
3.2 Problemas quando a variável resposta é binária

Os três principais problemas verificados quando a variável resposta é indicadora podem ser descritos como:

Problema 1: Os erros não possuem distribuição normal;

Problema 2: A variância dos erros não é constante;

Figura 1: Ilustração de uma função linear simples para uma variável resposta indicadora.



Problema 3: Restrições na função de resposta:

Devido ao fato de que a função de resposta representar as probabilidades quando a variável resposta é indicadora, deve-se observar a seguinte restrição no modelo:

$$0 \leq E(Y) = \pi \leq 1.$$

O problema encontra-se no fato de que muitas funções de resposta não possuem automaticamente essa restrição.

3.3 Modelo de Regressão Logística Simples

Para lidar com os problemas descritos no tópico anterior, optou-se por utilizar o modelo de Regressão Logística, o qual pode ser descrito da seguinte forma:

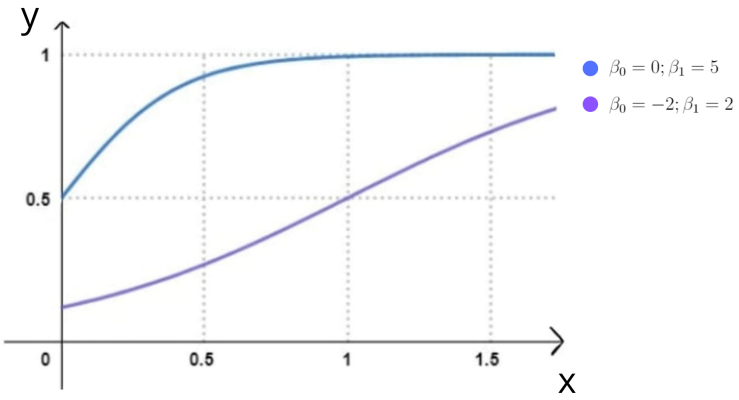
Sejam Y_i são variáveis aleatórias de Bernoulli independentes com $E(Y_i) = \pi_i$, onde,

$$\pi_i = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)}, \tag{i}$$

ou, de outra forma,

$$logit(\pi_i) = log(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 X_i.$$

Figura 2: Exemplos de funções em (i) quando $\beta_0 = 0; \beta_1 = 5$ e $\beta_0 = -2; \beta_1 = 2$.



3.4 Estimativa de Máxima Verossimilhança

As estimativas de máxima verossimilhança de β_0 e β_1 no modelo de regressão logística simples são os valores de β_0 e β_1 que maximizam a função de log-verossimilhança:

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

Não existe uma expressão analítica bem definida - até o presente momento - para os valores de β_0 e β_1 que maximizam tal função; são utilizados, portanto, técnicas computacionais para encontrar boas aproximações para tais resultados.

Utilizando as estimativas de β_0 e β_1 dadas por b_0 e b_1 respectivamente, tem-se que:

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

e consequentemente:

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

Em alguns casos pode-se verificar muitas observações repetidas com relação a variável resposta X. Nessa situação, a função de log-verossimilhança pode ser declarada da seguinte forma:

$$ln(L(\beta_0, \beta_1)) = \sum_{j=1}^{c} \{ln\binom{n_j}{Y_{.j}} + Y_{.j}(\beta_0 + \beta_1 X_j) - n_j ln[1 + exp(\beta_0 + \beta_1 X_j)]\},$$

onde $X_1, ..., X_c$ denota os níveis X em que as observações repetidas são obtidas e assumindo que há n_j respostas binárias no nível X_c.

3.5 Interpretação de b_1

A interpretação desse coeficiente é diferente da observada na Regressão Simples, uma vez que no caso em questão o coeficente angular - como pode ser observado pela Figura 2 - varia ao longo da curva; assim, a melhor interpretação a ser feita a cerca desse valor é feita ao analisar $exp(b_1)$: essa medida representa o aumento da razão de chances a cada aumento unitário na variável explicativa.

Considerando $\frac{\hat{\pi}_1}{1-\hat{\pi}_1} = odds_1 = exp(b_0+b_1X)$ e $\frac{\hat{\pi}_2}{1-\hat{\pi}_2} = odds_2 = exp(b_0+b_1(X+1))$, tem-se:

$$ln(odds_2) - ln(odds_1) = b_1 \Rightarrow ln(\frac{odds_2}{odds_1}) = b_1.$$

Logo,

$$exp(b_1) = \frac{odds_2}{odds_1}.$$

Pode-se notar, ainda, que o sinal de b_1 indica se a curva do modelo sobe ($b_1 > 0$) ou desce ($b_1 < 0$) - sendo que a taxa de variação da curva aumenta à medida que ($|b_1|$) aumenta. Quando $b_1 = 0$, a curva se achata para uma linha reta horizontal (nesse caso a variável resposta é independente da variável explicativa).

3.6 Regressão Logística Múltipla

O modelo de regressão logística simples pode ser facilmente estendido para mais de uma variável preditora:

$$E(Y) = \frac{exp(X'\beta)}{1 + exp(X'\beta)},$$
$$\pi' = X'\beta,$$

onde,

$$\beta'_{pX1} = [\beta_0, ..., \beta_p] \text{ e } X'_{pX1} = [X_0, ..., X_p]$$

Os processos para encontrar os parâmetros são análogos aos vistos no caso simples e o vetor de estimativas é dado, então, por:

$$b'_{pX1} = [b_0, ..., b_p].$$

3.7 Inferências sobre parâmetros de regressão

Para grandes amostras, estimadores de máxima verossimilhança para regressão logística geralmente possuem distribuição aproximadamente normal:

$$\frac{b_k - \beta_k}{s(b_k)} \sim Z \quad \text{para} \quad k = 0, 1, ..., p-1,$$

onde Z é uma variável aleatória normal padrão e $s_(b_k)$ é a estimativa aproximada para o desvio padrão de b_k.

Teste de Wald

Considere as seguintes hipósteses:

$$H_0 : \beta_k = 0;$$

$$H_1 : \beta_k \neq 0.$$

Neste caso, uma estatística de teste apropriada é:

$$Z^* = \frac{b_k}{s(b_k)}.$$

Eventualmente, o quadrado de Z^* é usado como uma estatística alternativa - o teste é então baseado em uma distribuição qui-quadrado com 1 grau de liberdade.

Teste da Razão de Verossimilhança

Considere as seguintes hipósteses:

$$H_0 : \beta_q = 0, \beta_{q+1} = 0, ..., \beta_p = 0;$$

$$H_1 : \text{pelo menos um dos} \quad \beta_k \quad \text{em} \quad H_0 \quad \text{não é nulo.}$$

Neste caso, uma estatística de teste apropriada é:

$$G^2 = -2ln(L(R)/L(F)),$$

onde $L(R)$ representa o valor da função de verossimilhança considerando-se o modelo reduzido - isto é, quando a hipótese nula é verdadeira - e $L(F)$ é o valor desta função considerando-se as estimativas de máxima verossimilhança quando utiliza-se o modelo completo; quando n é grande, $G^2 \sim \chi^2_{p-q}$ quando H_0 é verdadeira. A regra de decisão apropriada é então dada por:

$$\text{Se} \quad G^2 \leq \chi^2_{p-q}, \quad \text{não rejeitamos} \quad H_0;$$
$$\text{Se} \quad G^2 > \chi^2_{p-q}, \quad \text{rejeitamos} \quad H_0.$$

3.8 Testes de adequação

3.8.1 Teste de adequação qui-quadrado de Pearson

O teste de adequação qui-quadrado de Pearson assume apenas que as observações Y_{ij} são independentes e que o tamanho da amostra é razoável para os subgrupos analisados. As hipóteses são:

$$H_0 : E(Y) = [1 + exp(-X'\beta')]^{-1}$$
$$H_1 : E(Y) \neq [1 + exp(-X'\beta')]^{-1}$$

Considere os seguintes valores:

$$O_{j1} = Y_{.j} \quad \text{e} \quad O_{j0} = n_j - Y_{.j} \quad \text{para} \quad j = 1, ..., c;$$
$$\hat{\pi}_j = [1 + exp(-X'_j b)]^{-1};$$
$$E_{j1} = n_j\hat{\pi}_j \quad \text{e} \quad E_{j0} = n_j - n_j\hat{\pi}_j \quad \text{para} \quad j = 1, ..., c,$$

onde c representa o número de combinações distintas das variáveis preditoras e n_j representa o número de casos na classe j.

Nesse caso, tem-se que a estatística do teste será dada por:

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

Se a função de resposta logística for apropriada, $X^2 \sim \chi^2_{c-p}$, sendo p o número de parâmetros do modelo.

3.8.2 Teste de adequação de Hosmer–Lemeshow

Esse teste é realizado para casos nos quais os conjuntos de dados possuem pouca ou nenhuma repetição; o agrupamento de casos é feito então com base nos valores das estimativas das probabilidades (com aproximadamente o mesmo número de casos em cada classe).

Após formar os grupos, a estatística do teste é a mesma utilizada no teste anterior. Essa estatística de teste é bem aproximada pela distribuição qui-quadrada com $c-2$ graus de liberdade.

3.9 Resíduos de regressão logística

Uma vez que a variável resposta seja binária, o i-ésimo residual ϵ_i assumirá apenas os valores: $1-\hat{\pi}_i$ (se $Y_i = 1$) ou $-\hat{\pi}_i$ (se $Y_i = 0$). Assim, sua distribuição não pode ser considerada normal e é desconhecida. Nesse caso, utilizam-se outras medidas para analisar esses resíduos, tais como:

Resíduo de Pearson

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

Resíduo de Pearson estudantizado

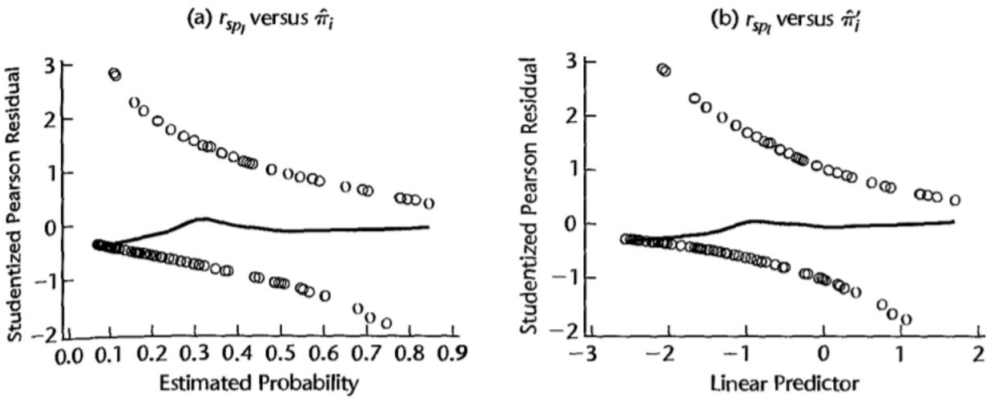
$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}},$$

onde h_{ii} é o i-ésimo valor na diagonal da matriz $\hat{W}^{1/2}X(X'\hat{W}X)^{-1}X'\hat{W}^{-1/2}$, sendo $\hat{W}$ a matriz (n x n) cujos elementos são dados por $\hat{\pi}_i(1-\hat{\pi}_i)$.

Gráfico de resíduos *versus* probabilidades previstas

Se o modelo de regressão logística estiver correto, uma suavização leve do gráfico dos resíduos em relação à probabilidade estimada $\hat{\pi}_i$ deve resultar aproximadamente em uma linha horizontal com intercepto no zero. Qualquer desvio significativo desse comportamento sugere que o modelo pode ser inadequado. Na prática, pode-se ainda aplicar essa análise considerando-se o resíduo de Pearson, o resíduo de Pearson estudantizado e os valores dos resíduos por si só.

Figura 3: Exemplos de gráficos de resíduos com suavização leve.



Fonte: Neter et al. Applied Linear Statistical Models (p. 595).

3.10 Detecção de observações influentes

Influência na estatística qui-quadrado de Pearson

Seja X^2 a estatística de Pearson com base no conjunto de dados completo, e considere $X^2_{(i)}$ o valor dessa estatística quando o i-ésimo caso é desconsiderado; a estatística delta qui-quadrado é então definida:

$$\Delta X_i^2 = X^2 - X^2_{(i)}.$$

Esse valor fornece medidas de influência do i-ésimo caso nessas estatísticas - a decisão de quando a observação é de fato influente ou não é relativamente subjetiva e pode ser feita com o auxílio de gráficos.

3.11 Previsão de uma nova observação

3.11.1 Escolha da regra de previsão

Algumas abordagens podem ser feitas na escolha da regra para realizar a previsão de uma nova obsrvação nos dados a partir da estimativa $\hat{\pi}_h$:

1 - Usar 0.5 como ponto de corte

Essa é uma boa escolha quando as probabilidades entre ocorrer 0 e 1 são próximas ou quando o custo de se fazer uma observação errada em ambos os sentidos for próximo.

2 - Escolher o melhor ponto de corte

Nesse caso, será escolhido o ponto de corte com o qual a proporção de previsões incorretas é mais baixa. Essa é uma boa abordagem quando a amostra utilizada no processo de modelagem é aleatória e representativa da população de interesse.

3 - Usar probabilidades *a priori* e custos de previsões incorretas para determinar o corte

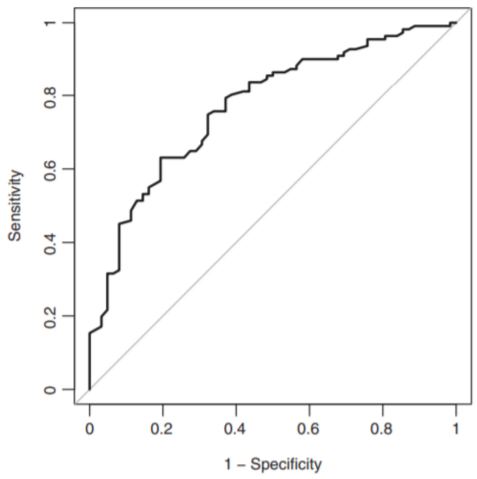
3.11.2 Estimativa da taxa de erro de previsão

Essa medida é realizada ao aplicar a regra de predição escolhida a um conjunto de dados de validação. Se os resultados ao analisar os dados de validação e os dados originais forem relativamente semelhantes, então, há uma indicação de que a capacidade preditiva do modelo criado é de fato confiável; caso contrário, o modelo ajustado não preve bem novas observações.

3.11.3 Curva de ROC

A curva de ROC (Curva de característica de operação do receptor) é um gráfico que mostra a sensibilidade e a especificidade das previsões para todos os cortes possíveis π_0. Quanto melhor o poder preditivo, maior é a área sob a curva. [2]

Figura 4: Exemplo de gráfico de ROC.



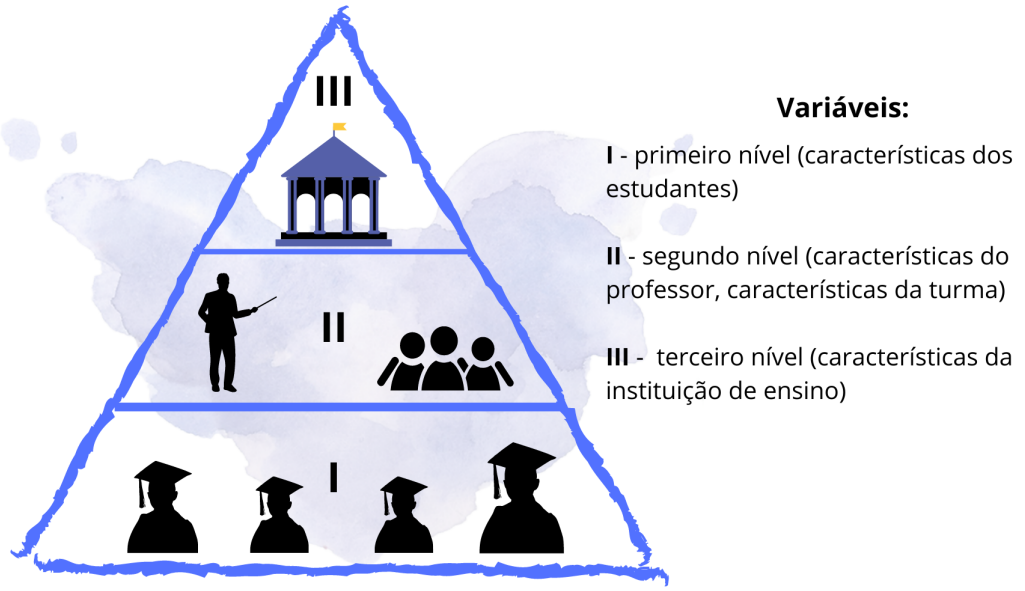
Fonte: Agresti, An Introduction to Categorical Data Analysis (p. 112)

[2] A sensibilidade representa a probabilidade se prever $\hat{Y} = 1$ quando Y = 1; já a especificidade mostra a probabilidade de $\hat{Y} = 0$ quando Y = 0.

4 Regressão Multinível

A Regressão multinível analisa dados organizados de forma hierárquica - em diferentes níveis; por exemplo, ao analisar observações referentes a estudantes em uma escola com diferentes turmas. Essa técnica leva em consideração que esse tipo de dado - no nível de observação primária (no exemplo dado seria o nível do estudante) - não possuem observações independentes entre si, sendo correlacionadas de acordo com o agrupamento dos demais níveis (no exemplo seria pelo nível da turma). A Figura 5 exemplifica esse cenário:

Figura 5: Exemplo da organização dos dados utilizados em técnicas de Regressão Multinível com base em um contexto acadêmico.



4.1 Modelo com dois níveis

O modelo de Regressão Multinível pressupõe ainda que haja apenas uma variável resposta (medida no nível mais baixo) e que haja variáveis explicativas em todos os níveis.

Suponha que deseja-se estimar a variável resposta Y_{ij} referente a uma i-ésima observação do primeiro nível e ao j-ésimo grupo do segundo nível. Considere ainda, sem perda de generalidade, que haja 2 variáveis explicativas X_1 e X_2. Nesse sentido, o modelo pode ser escrito da seguinte forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + e_{ij},$$

em que o índice $j = 1, ..., J$ indica o grupo - referente ao segundo nível - e o índice $i = 1, ..., n_j$ o a i-ésima observação do grupo j. Os erros residuais e_{ij} são assumidos como tendo uma média de zero, e uma variância a ser estimada. Os coeficientes β_{0j}, β_{1j} e β_{2j} são assumidos como possuindo distribuição normal multivariada.

Considere, agora, a variável explicativa Z; tem-se que:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} Z_j + u_{kj}, k = 0, 1, 2, \tag{i}$$

em que u_{0j}, u_{1j} e u_{2j} indicam os erros residuais no nível do grupo - considerados com média zero e sendo independentes dos erros do primeiro nível.

Nesse caso, o erro será diferente para diferentes valores da variável explicativa X_{ij}: heterocedasticidade. Para verificar a homocedasticidade ou heterocedasticidade dos dados deve-se analisar o valor da correlação intraclasse ρ:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij},$$

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}.$$

O valor deρ indica a proporção da variância total explicada pela estrutura de agrupamento na população. Segundo (Hox, 2018), valores de ρ perto de 36% são considerados altos para dados de ciências sociais.

4.2 Interpretação dos parâmetros

Os coeficientes β_{1j} e β_{2j} representam o quanto a média da variável resposta muda - para o grupo j - com o aumento de uma unidade das variáveis X_1 e X_2 respectivamente. O intercepto β_{0j} representa o valor médio de Y caso as variáveis explicativas sejam nulas.

Uma vez que na regressão multinível os coeficientes de interceptação e inclinação variam entre as classes, eles são freqüentemente chamados de coeficientes aleatórios. Espera-se, contudo, que esta variação não seja totalmente aleatória.

Observando-se a equação (i), se γ_{01} for positivo, então o valor médio da variável resposta tende a ser maior em grupos onde a variável Z admite valores maiores. Analogamente, caso γ_{01} seja negativo.

Se $k = 1$ ou $k = 2$ na equação (i), então a equação indica que a relação entre a variável resposta e as variáveis explicativas dependem da variável Z. Assim, se nesse caso γ_{k1} for positivo, o efeito de $\mathbf{X}$ sobre Y aumenta em grupos onde Z admite valores mais altos - analogamente no caso em que γ_{k1} seja negativo. Dessa forma, a variável Z age como uma variável moderadora entre $\mathbf{X}$ e Y.

4.3 Passos para a construção do modelo

Passo 1

Analisa-se o modelo somente de intercepto ou modelo nulo:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}.$$

Assim, obtem-se o valor de ρ:

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}.$$

Passo 2

Elabora-se um modelo com todas as variáveis explicativas fixas do nível mais baixo:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij},$$

onde X_{pij} representa as p variáveis explicativas do primeiro nível. Neste passo, estima-se a contribuição de cada variável explicativa deste nível.

Passo 3

Acrescentam-se as variáveis explicativas do segundo nível:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{0j} + e_{ij},$$

onde Z_{qj} representa as q variáveis explicativas do segundo nível.

Passo 4

Avalia-se se algum dos coeficientes das variáveis explicativas do nível mais baixo influencia a variância no segundo nível. Este modelo, chamado modelo de coeficientes randômicos, é dado por:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + u_{pj}X_{pij} + u_{0j} + e_{ij},$$

onde os u_{pj} são os resíduos do primeiro nível dos coeficientes das variáveis explicativas X_{pij} do primeiro nível.

Passo 5

Adicionam-se as interações entre-níveis entre variáveis explicativas do segundo nível e aquelas variáveis explicativas do primeiro nível que tiveram variância significativa de coeficientes no passo 4:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{pq}z_{qj}X_{pij} + \gamma_{0q}Z_{qj} + u_{pj}X_{pij} + u_{0j} + e_{ij}.$$

4.4 Estimação dos parâmetros

A estimativa dos parâmetros é feita principalmente pelo método de máxima verossimilhança. Duas funções de verossimilhança diferentes são usadas na modelagem de regressão multinível. Uma é chamada de *full maximum likelihood* (FML); neste método, tanto os coeficientes de regressão quanto o os componentes de variância são incluídos na função de verossimilhança. A outra chama-se *restricted maximum likelihood* (RML); aqui, apenas os componentes de variância estão incluídos na função de verossimilhança e os coeficientes de regressão são estimados em uma segunda estimativa. Na prática, ambos os métodos produzem resultados semelhantes.

Teste de Wald

Esse teste é válido para grandes amostras e suas hipóteses são:

$$H_0 : \theta = 0;$$

$$H_1 : \theta \neq 0.$$

Sendo,

$$Z = \frac{\hat{\theta}}{\hat{\sigma}_\theta},$$

a estatística do teste (Z possuindo distribuição normal padrão).

4.5 Comparação de modelos

Para avaliar o ajustamento do modelo aos dados, pode-se calcular uma estatística de *deviance*:

$$deviance = (-2ln(\mathcal{L})),$$

onde $\mathcal{L}$ representa o valor de covergência da função de verossimilhança. Em geral, os modelos com um menor *deviance* se encaixam melhor do que os modelos com maior *deviance*.

A diferença da *deviance* para dois modelos aninhados - isto é, quando um pode ser derivado do outro pela remoção ou inclusão de parâmetros - tem uma distribuição qui-quadrado, com graus de liberdade iguais à diferença no número dos parâmetros estimados nos dois modelos. Essa pode ser uma forma de verificar qual dos modelos é o mais adequado para explicar o comportamento dos dados.

Se os modelos a serem comparados não forem aninhados, pode-se utilizar ou o modelo mais simples ou comparar o valor do Critério de Informação de Akaike (AIC) calculado a partir *deviance* d e o número de parâmetros estimados q:

$$AIC = d + 2q.$$

5 Metodologia

No desenvolvimento do presente trabalho foram utilizados dados extraídos do Sistema de Informações Acadêmicas (SIGRA) contendo informações sobre cada estudante que cursou a disciplina Estatística Aplicada no período de 1994 até 2019. Os passos comentados a seguir foram realizados utilizando-se principalmente a linguagem de programação R a partir do RStudio.

Limpeza do banco de dados

O processo de limpeza do banco de dados deu-se em várias etapas as quais foram revisadas diversas vezes. As variáveis presentes no banco de dados inicial eram: id do estudante (codificação aleatória), Curso do estudante, Código da disciplina, Nome da disciplina, Período no qual a disciplina foi cursada, Tipo do professor, Turma, Menção, Porcentagem de faltas do estudante, Horário de início da aula, Horário de termino da aula e ano de ingresso do estudante na Universidade.

Após filtrar apenas os casos da disciplina de Estatística Aplicada verificou-se que haviam inconsistências gerando observações repetidas de alunos devido a turmas com horários diferentes das aulas na semana; turmas que durante o semestre tiveram professores diferentes; estudantes que cursaram mais de um curso ao longo do período observado, entre outros problemas. Para contornar o problema relativo ao horário da disciplina, substituiu-se as variáveis referentes aos horários das aulas pelo turno na qual elas foram realizadas (Manhã, Tarde ou Noite).

Posteriormente, foram selecionados apenas os casos nos quais os alunos fizeram a disciplina até o final - isto é, retirou-se casos de trancamento e crédito concedido. Com isso, foram filtradas apenas as primeiras tentativas completas de cada estudante na matéria. Nesses pontos algumas adaptações foram feitas nos dados a fim de corrigir algumas inconsistências observadas. Por fim, criou-se uma variável referente a modalidade da disciplina para o estudante (Módulo Livre, optativa ou obrigatória).

Análise Exploratória

Nessa parte, foram realizadas inicialmente análises univariadas e bivariadas das principais variáveis presentes no banco de dados após todo seu processo de limpeza; para isso, foram criados gráficos e medidas para auxiliar tal processo. Posteriormente foram feitas análises bivariadas referentes às variáveis explicativas (consideradas mais relevantes) - nessa parte foram analisados apenas os dados relativos aos dois últimos anos presentes no banco de dados: 2018 e 2019 - relizando-se, ainda, alguns testes qui-quadrado. Por fim, foram feitas algumas das análises mencionadas considerando-se apenas os casos em que a menção final do estudante na disciplina foi diferente de SR (o qual pode ser considerado

como uma forma de abandono do estudante uma vez que essa menção é obtida quando o aluno possui mais de 25% de faltas ou quando este tira nota zero em todas as avaliações da matéria). Ainda nessa parte foi criada a categoria "Outros"referente a alguns cursos cuja frequência no banco de dados foi relativamente pequena.[3]

Modelagem dos dados

Para a modelagem dos dados foi utilizado técnicas de Regressão Logística Multinível, porquanto a variável resposta em questão pode assumir apenas valores binários (1 - aprovação; 0 - reprovação) e observando-se o fato de que os dados mostram-se agrupados em diferentes níveis: estudante e turma (não sendo portanto possível assumir independência entre as observações). Esse processo foi feito para cada um dos casos: o conjunto de dados completo e o conjunto de dados desconsiderando-se os casos nos quais os estudantes obtiveram menção SR no seu resultado final na disciplina. Os principais cálculos utilizados no processo foram feitos pelo R a partir do aplicativo RStudio. Inicialmente obteve-se os coeficientes do modelo a partir de variáveis explicativas consideradas pertinentes. Em seguida, foram calculados os coefecientes gerados a partir de amostras do banco de dados original (divididas entre dados de teste e de validação - cada amostra contém metade das informações dos dados originais). Foram, então, realizadas observações referentes às razões de chances do modelo final e, por fim, foram feitas considerações a cerca do diagnóstico desse modelo.

Análise de dados sobre perfil do estudante

Nessa parte, foram realizadas análises referente ao pareamento do banco de dados utilizado no estudo até então com dados de pesquisa do perfil do estudante da Universidade de Brasília. Durante esse processo, optou-se por analisar dados relativos a estudantes ingressantes em 2017 e 2018 na Universidade; como os dados da pesquisa sobre o perfil do aluno foram obtidos de forma voluntária, não foi possível encontrar informações relativas a todos os estudantes no banco de dados originais - foi possível obter 1165 observações pareadas (cerca de 67% do que havia no banco de dados original). Os resultados relativos a tais informações foram analisados de forma análoga aos tópicos anteriores e - como não foram provenientes de uma amostra representativa - não tem o intuito de inferir sobre a população em questão, servindo então como um estudo exploratório a cerca de tais informações.

[3]Tais cursos foram referenciados no apêndice.

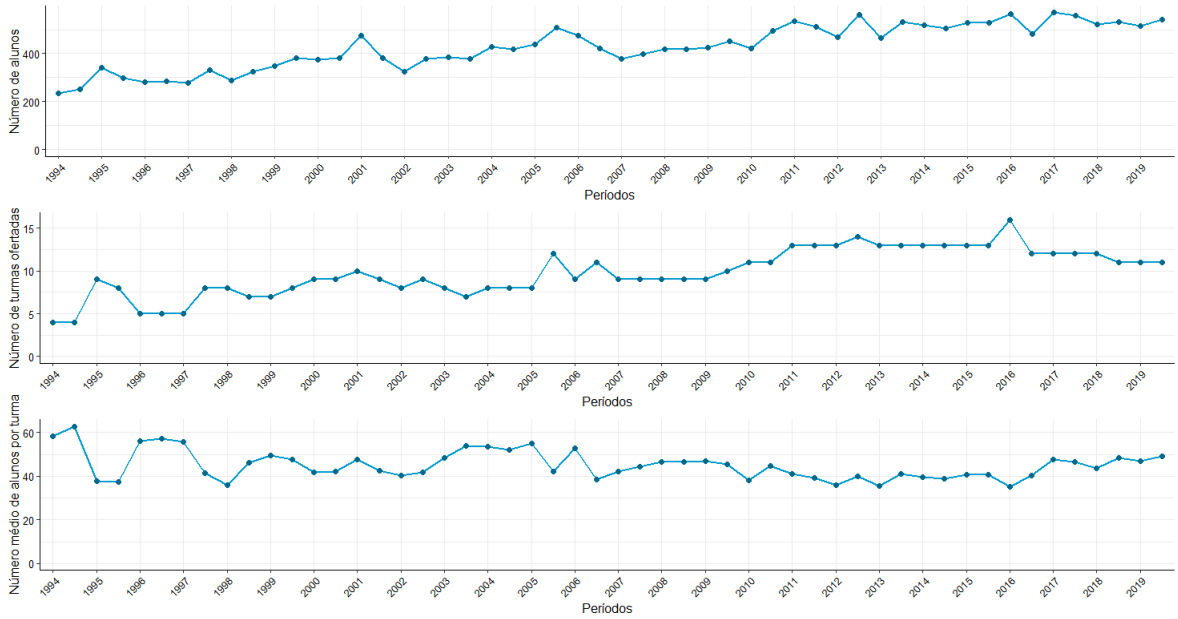
6 Análise exploratória

6.1 Análises Gerais

Nessa parte do estudo, pretende-se analisar de modo geral os comportamentos das variáveis presentes no banco de dados final - contendo observações referentes a 22245 estudantes ao longo dos anos de 1994 até 2019.

6.1.1 Visão Geral

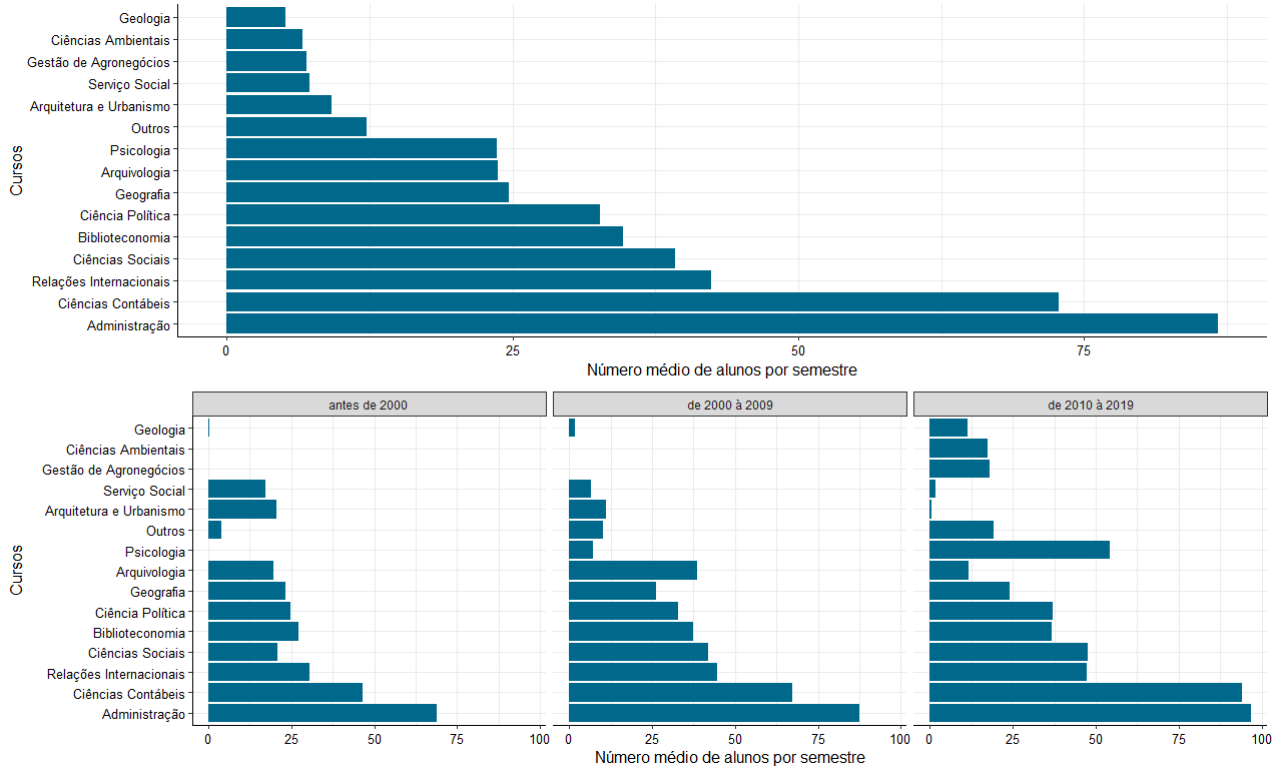
Figura 6: Nº de alunos por semestre; Nº de turmas por semestre; Nº médio de alunos por turma a cada semestre respectivamente.



A figura 6 indica que o número de estudantes que tentaram pela primeira vez a disciplina, foi subindo até 2010 - possuindo certa variação nesses períodos. Aparentemente, a partir de 2011, os números de alunos novos na matéria se estabilizou - possuindo picos em 2013, 2016 e 2017. O número de turmas ofertadas, nas quais estes alunos foram matriculados, apresenta uma tendência crescente até 2016 - onde ocorreu um pico - e após esse período os números de turmas aparentaram ter um pequeno decrescimento. Já o número médio de alunos novos na disciplina por turma aparenta oscilar bastante ao longo do intervalo de tempo analisado - possuindo um pico em meados de 1994.

6.1.2 Cursos dos estudantes

Figura 7: Cursos os estudantes de Estatística Aplicada em sua primeira vez na disciplina (1994 a 2019).

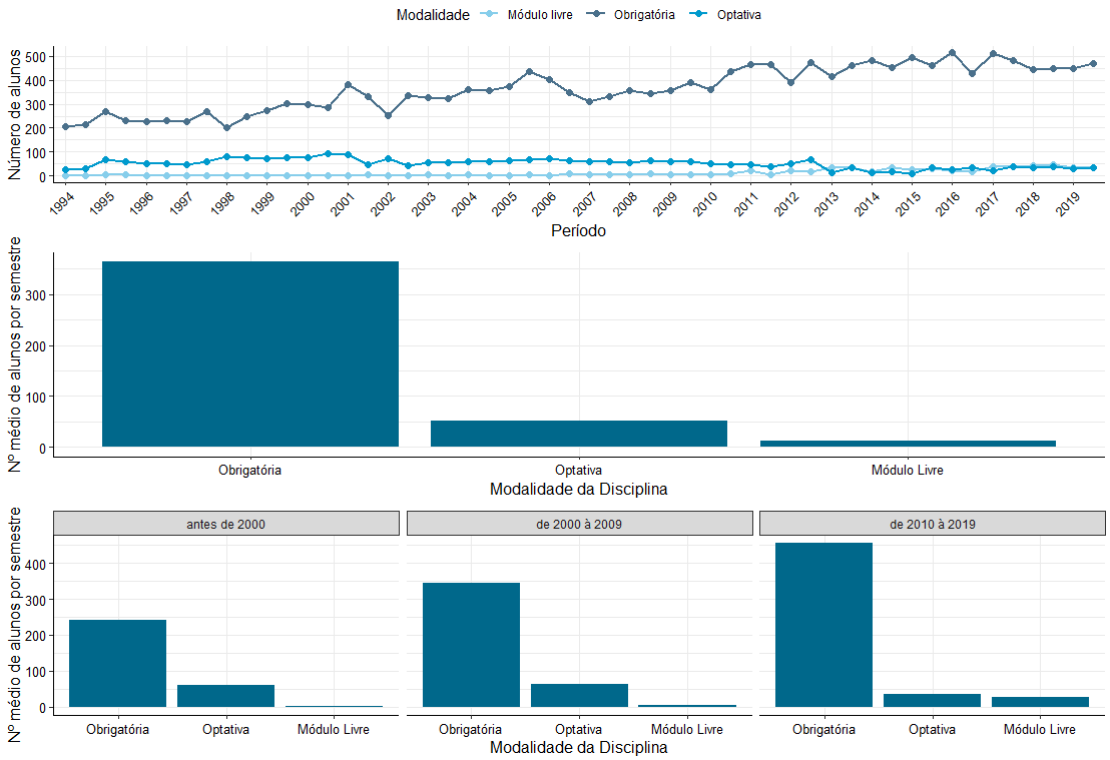


Pela Figura 7, nota-se que os principais cursos cujos alunos fizeram Estatística Aplicada pela primeira vez no intervalo de tempo completo considerado são Administração e Ciências Contábeis respectivamente - com mais de 70 alunos, em média, por semestre. A categoria "Outros"[4] no eixo dos cursos na imagem acima representam estudantes de cursos que apareceram com frequência muito baixa (menos de 100 observações para cada um desses cursos). No total foram verificados estudantes de 76 cursos diferentes no banco de dados. Nota-se ainda, que o número médio de alunos de cada curso por semestre aparenta aumentar em períodos mais recentes - sendo que durante o intervalo de 2010 a 2019, os números médios de alunos por semestre dos cursos de Administração e de Ciências Contábeis chegam a quase 100 estudantes por curso.

[4]Ver apêndice.

6.1.3 Modalidade da disciplina para os diferentes cursos dos estudantes

Figura 8: Nº de alunos segundo modalidade da disciplina: por semestre, nº médio e por período respectivamente.



Analisando a Figura 8, pôde-se notar que o número de estudantes - em sua primeira tentativa completa na disciplina - cursando a matéria com a modalidade obrigatória teve uma tendência crescente ao longo dos anos. A partir de 2013, notou-se ainda que o número de estudantes cursando a disciplina como Módulo Livre [5] e como optativa assemelharam-se bastante a partir desse ano.

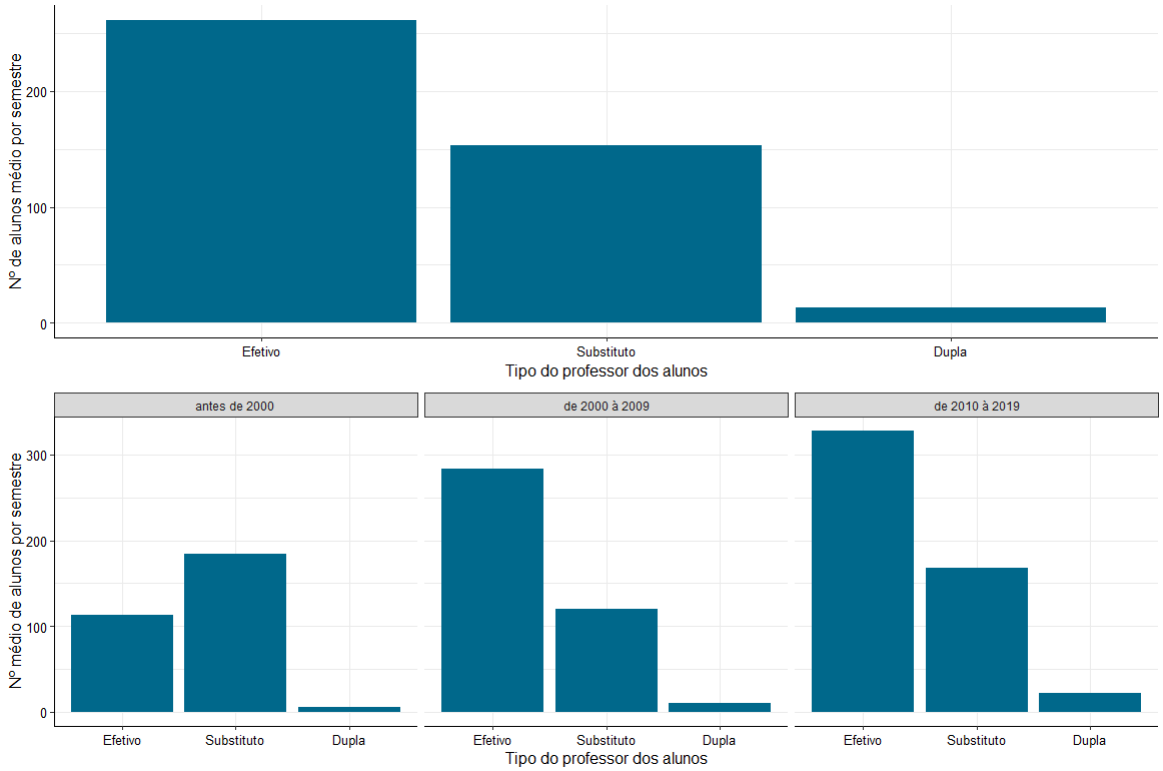
Percebe-se ainda que - em média - mais de 350 alunos por semestre cursam a disciplina com a modalidade obrigatória em sua primeira tentativa; cerca de 50, em média, cursam como optativa em cada semestre; e uma quantidade bem reduzida cursa como Módulo Livre. Percebeu-se ainda que o número médio de estudantes cursando como optativa a disciplina aparentou diminuir em períodos mais recentes - enquanto o comportamento contrário foi observado em estudantes cursando a matéria como Módulo Livre.

[5] Lembrando que - assim como dito na introdução - as disciplinas de módulo livre de um curso são todas as disciplinas de graduação que não são de abrangência restrita e que não constam no currículo do curso.

6.1.4 Professores

Pela Figura 9, nota-se que em média mais de 250 alunos por semestre que tentam pela primeira vez a disciplina a cursaram com professores efetivos do quadro do departamento; cerca de 150 fazem com professores substitutos; e uma pequena quantidade faz com dupla de professores. [6] Percebe-se ainda que, de 1994 até 1999 em média a maioria dos alunos no semestre cursavam a matéria com professores substitutos - o que mudou a partir de 2000 até 2019.

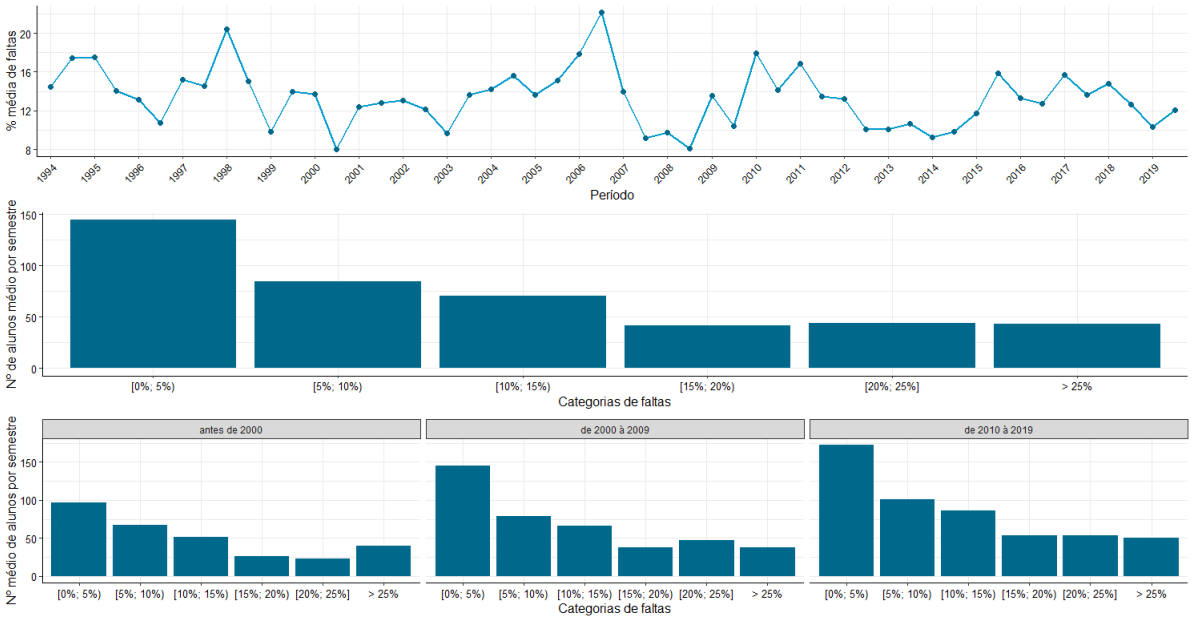
Figura 9: Professores da disciplina Estatística Aplicada que oferecem a disciplina para alunos que estão em sua primeira tenativa na matéria (1994 a 2019).



[6] Dupla de professores, nesse caso, caracteriza-se pelos casos nos quais os estudantes acabaram cursando a matéria com dois professores diferentes.

6.1.5 Porcentagem de faltas dos estudantes na disciplina

Figura 10: Porcentagem de faltas dos estudantes: por semestre, nº médio e por período respectivamente.

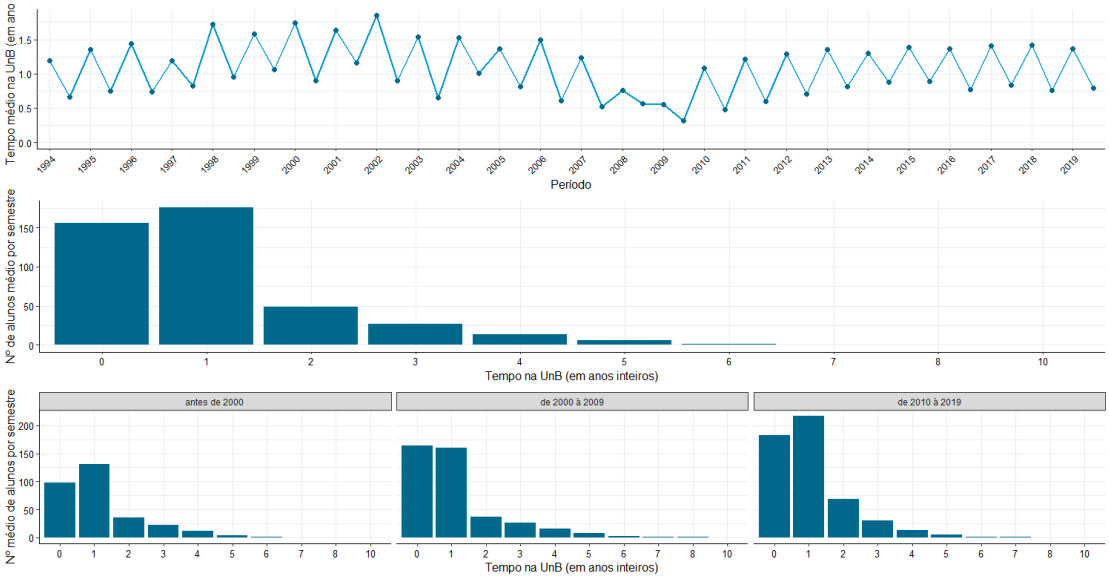


Pela Figura 10, pode-se verificar a frequência registrada de faltas dos estudantes que cursaram a primeira vez a matéria. A maioria dos registros indica uma porcentagem de faltas menor que 25% - sendo que ao longo dos semestres a porcentagem média de faltas dos estudantes aparenta variar consideravelmente. A maioria dos estudantes - em média - por semestre aparentam possuir porcentagens menores de faltas. Nota-se ainda que uma quantidade significativa de estudantes teve 100% de registros de falta.

6.1.6 Tempo do estudante da Universidade

Pela Figura 11, nota-se que a maioria dos estudantes que tentam a disciplina pela primeira vez possui menos de 2 anos na Universidade; no entando, é possível verificar que ao longo dos anos essa tendência aparentou se modificar bastante - sendo que em torno dos anos 2000 o tempo médio dos estudantes em sua primeira tentativa na disciplina foi de quase 2 anos completos na universidade; esses números foram diminuindo até meados de 2009 a partir de quando eles voltaram a subir - chegando ao tempo médio de pouco mais de 1 ano inteiro na universidade até tentar pela primeira vez a disciplina.

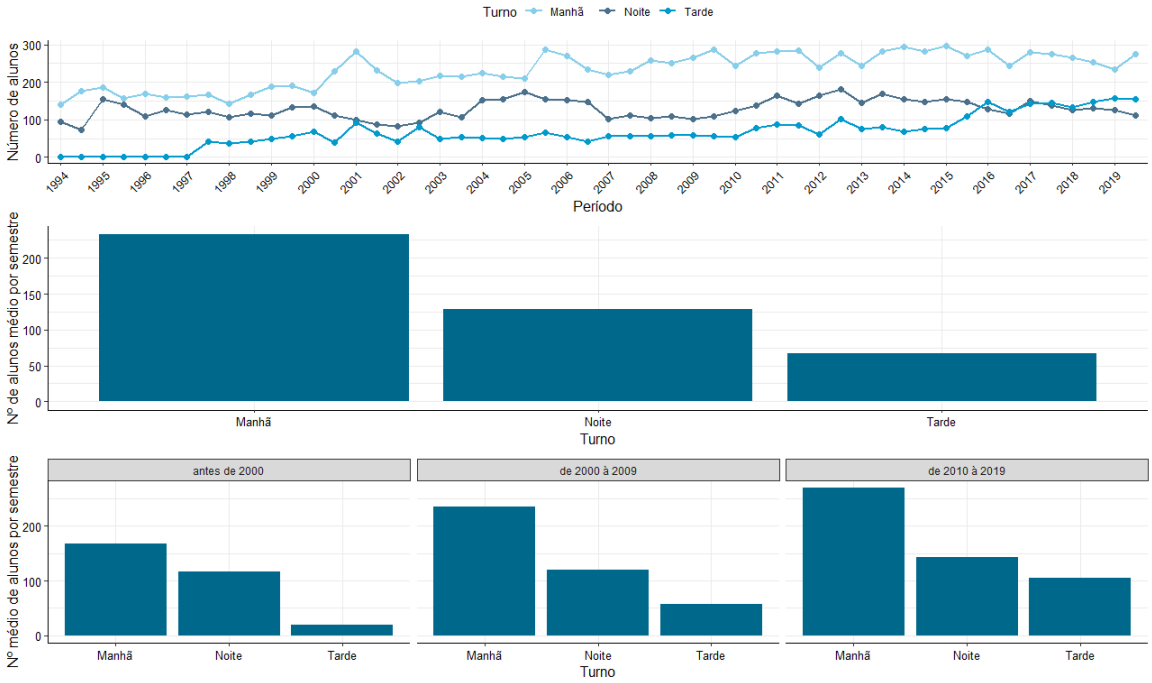
Figura 11: Tempo em anos dos estudantes na Universidade durante sua primeira tentativa na disciplina (1994 a 2019).



6.1.7 Turno de oferta da disciplina

A partir da Figura 12, verificou-se que a maioria dos estudantes - na sua primeira tentativa na matéria - fizeram a discplina no período noturno; seguidos respectivamente pelos turnos da manhã e da tarde - essa tendência aparenta se manter ao longo de todo o período analisado.

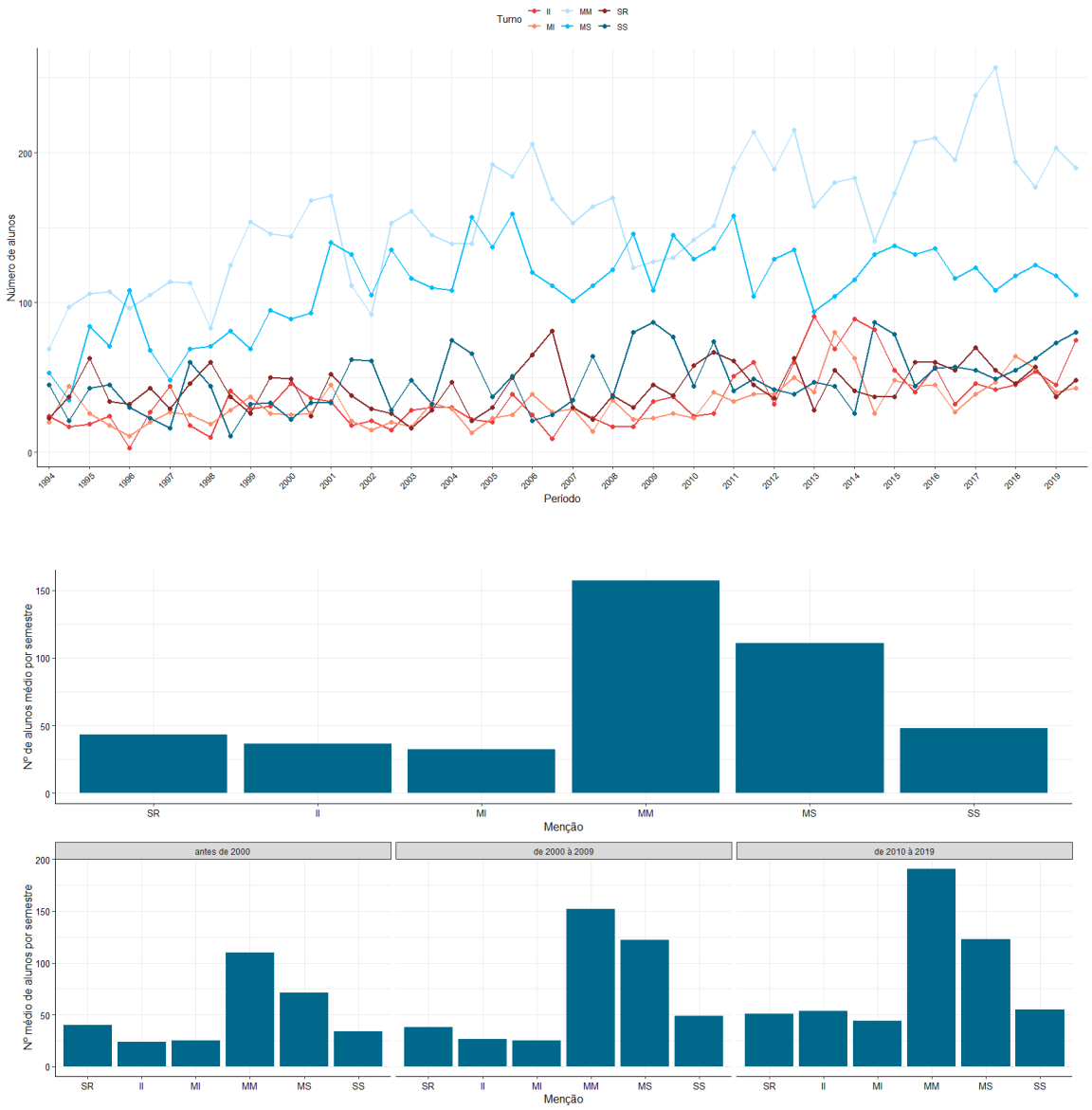
Figura 12: Turno no qual foi cursada a matéria na primeira tentativa dos estudantes (1994 a 2019).



6.1.8 Menção dos estudantes

Pelos gráficos na Figura **??**, nota-se que a menção [7] mais comum dentre os alunos que tentam pela primeira vez a disciplina foi MM (menção mínima para aprovação), seguida - respectivamente - por MS e SS. Dentre os alunos reprovados na matéria, a maioria ficou com menção SR, seguido respectivamente por II e MI. Assim, de forma resumida, percebe-se que - dentre cada categoria (Aprovado ou Reprovado) as menções mais comuns são as mais baixas da respectiva categoria.

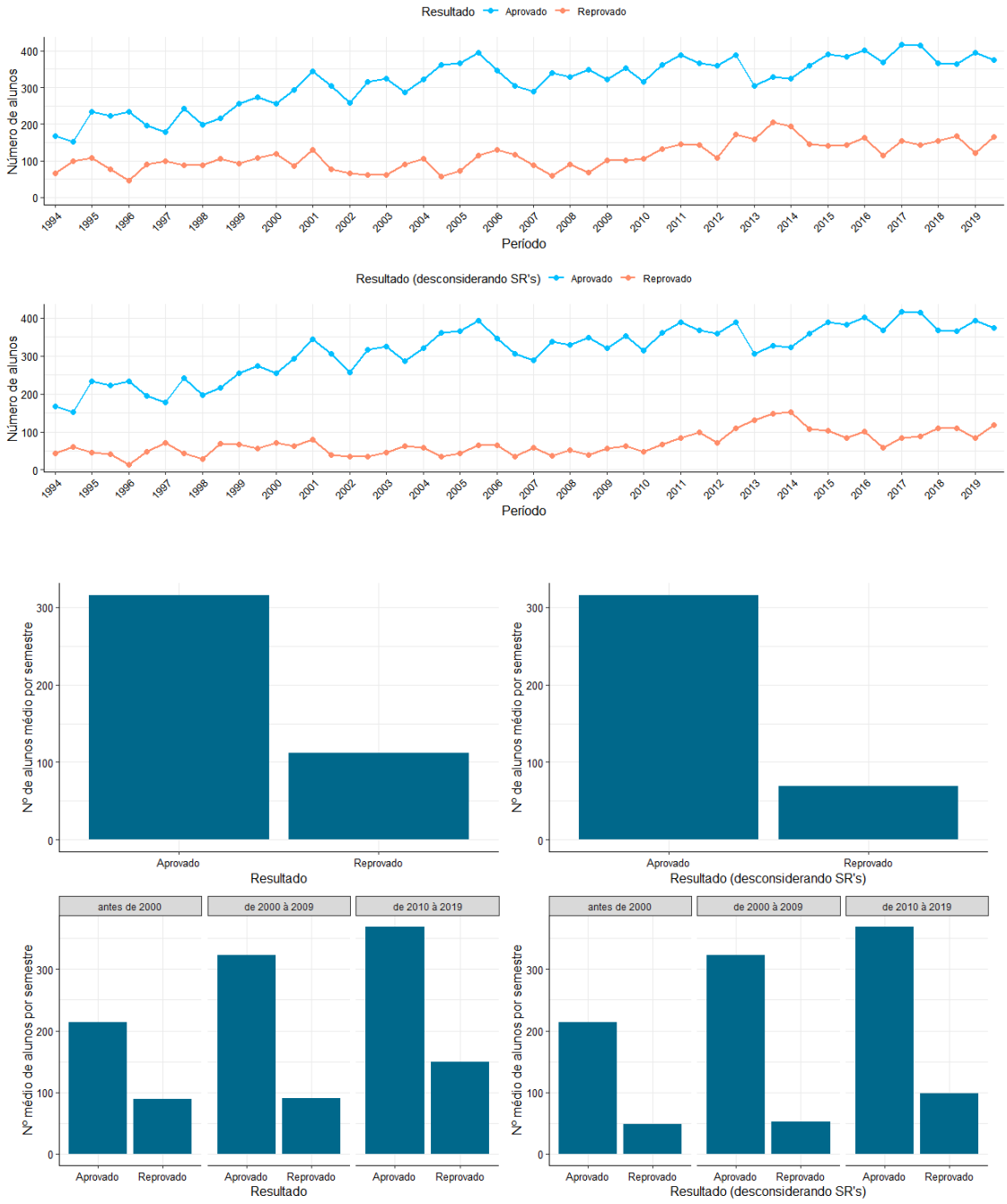
Figura 13: Menção final dos estudantes na sua primeira tentativa na disciplina (1994 a 2019).



[7] As menções consideradas no presente estudo foram (em ordem crescente) SR, II, MI, MM, MS e SS - sendo as 3 primeiras referentes a reprovações e as 3 últimas referentes a menções de aprovação dos estudantes.

6.1.9 Resultado final dos estudantes

Figura 14: Resultado final dos estudantes na sua primeira tentativa na disciplina (1994 a 2019).



Observou-se que em média - dentre os alunos que fizeram a matéria pela primeira vez - mais de 300 alunos são aprovados por semestre - enquanto cerca de 100 alunos reprovam a cada semestre em sua primeira tentativa na matéria - sendo que de modo geral a probabilidade de aprovação vale cerca de 74%. Analisando-se os períodos de 1994 até 1999 e 2010 à 2019, verificou-se que em cada um deles as probabilidade de aprovação foram em torno de 70%. Já ao considerar o período de 2000 à 2009 verificou-se que esse

valor foi aproximadamente 78%.

Já ao analisar as taxas de reprovação dos estudantes desconsiderando os casos de SR - entendendo nesse caso o SR como uma forma de abandono da disciplina - pôde-se verificar que a probabilidade de aprovação geral na disciplina vale aproximadamente 82% e essas probabilidades valem nos diferentes períodos considerados cerca de 81%, 86% e 79% respectivamente. [8]

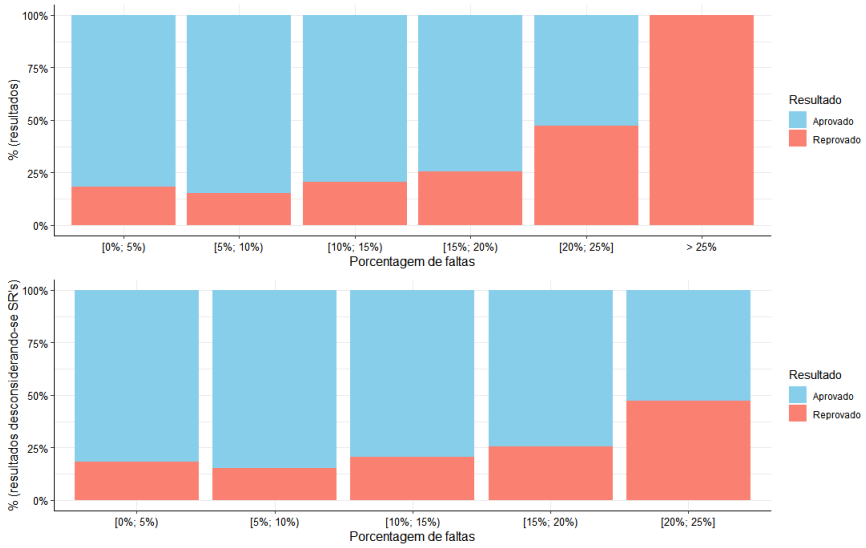
[8]O SR pode ser considerado como uma forma de abandono do estudante uma vez que essa menção é obtida quando o aluno possui mais de 25% de faltas ou quando este tira nota zero em todas as avaliações da matéria.

6.2 Análises entre a variável resposta e as explicativas

No desenvolvimento da modelagem, optou-se por trabalhar apenas com os resultados relativos aos anos de 2018 e 2019 (os quais representam os dados mais recentes no banco de dados analisado). Essa escolha deve-se ao fato de que muitos fatores aparentaram mudar ao longo dos anos da Universidade, sendo então interessante analisar quais são as variáveis que infuenciam atualmente no rendimento dos estudantes em Estatística Aplicada. Além disso, foi considerado dois casos: dados completos e dados sem SR's (uma vez que o SR pode ser interpretado como uma forma de abandono da matéria pelo estudante). A seguir, então, foram feitas algumas análises bivariadas, relacionando variáveis candidatas a participarem do modelo de acordo com o resultado final obtido pelos estudantes na sua primeira tentativa na disciplina.

6.2.1 Faltas por resultado

Figura 15: Resultado final dos estudantes segundo porcentagem de faltas na disciplina (2018-2019).



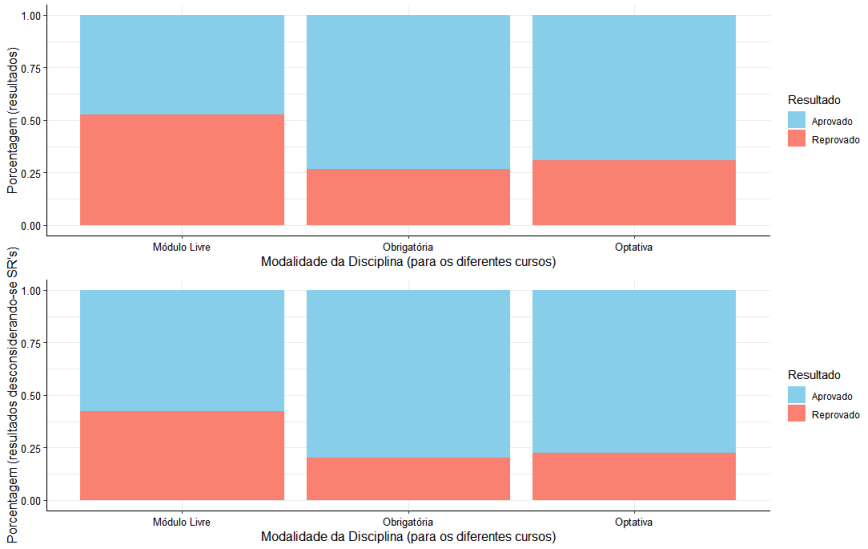
Pela Figura 15, pode-se perceber que - para ambos os casos considerados - quanto maior a porcentagem de faltas os estudantes na matéria, maior a proporção de alunos reprovados.

6.2.2 Modalidade da disciplina por resultado

Pela Figura 16, cosiderando-se todas as menções, mais da metade dos estudantes que cursaram a disciplina - nos últimos dois anos analisados - como uma matéria de módulo livre em sua primeira tenatativa reprovaram, enquanto a maior proporção de

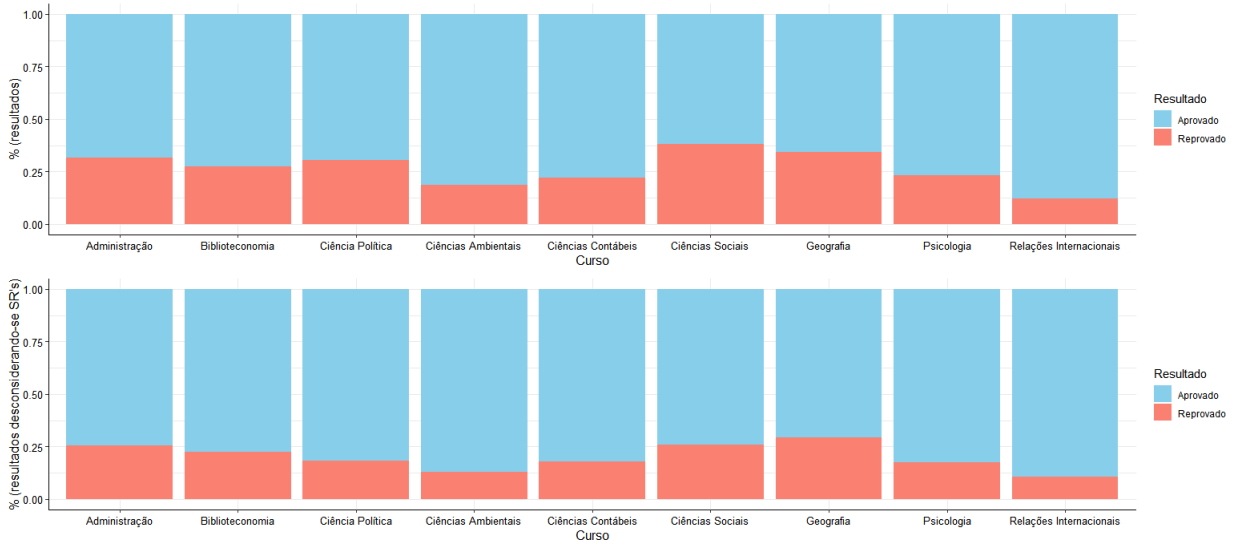
aprovação foi verificada em alunos que cursaram a maéria como disciplina optativa. Já ao descosniderar casos de SR, notou-se que os resultados aparentam não diferir muito quanto as modalidades optativa e obrigatória - contudo, as maiores proporções de reprovação ainda mostram-se presentes em estudantes que cursam a disciplina como Módulo Livre.

Figura 16: Resultado final dos estudantes segundo modalidade da disciplina (2018 - 2019).



6.2.3 Cursos para os quais a disciplina é obrigatória

Figura 17: Resultado final dos estudantes segundo curso: casos apenas de cursos para os quais a matéria é obrigatória (2018-2019).



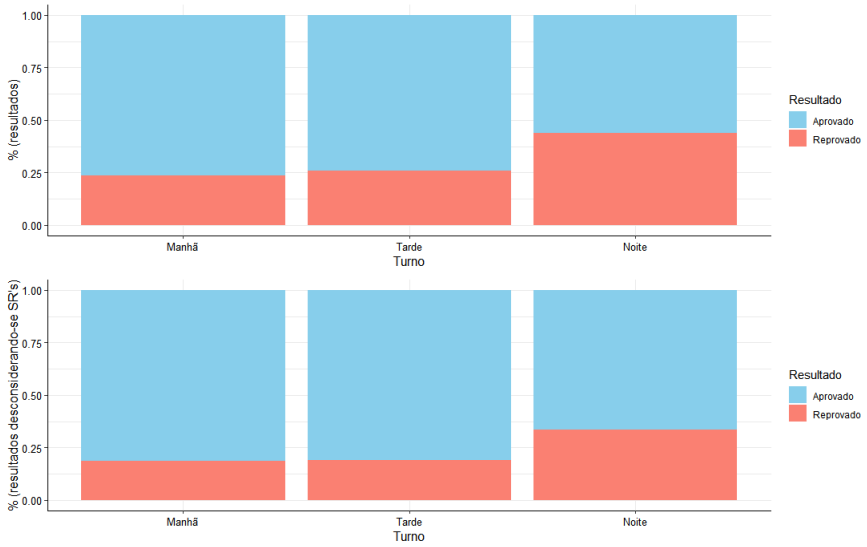
Pela Figura 17, nota-se que as proporções de reprovação dos estudantes, em ambos os casos considerados, aparentam divergir bastante de acordo com o curso que eles estão fazendo durante a disciplina; os melhores resultados foram relativos a alunos

de Relações Internacionais e os piores foram - de modo geral - relativos a Administração, Blibliotecономia, Ciência Política, Ciências Sociais e Gegrafia.

6.2.4 Turno de oferta da disciplina

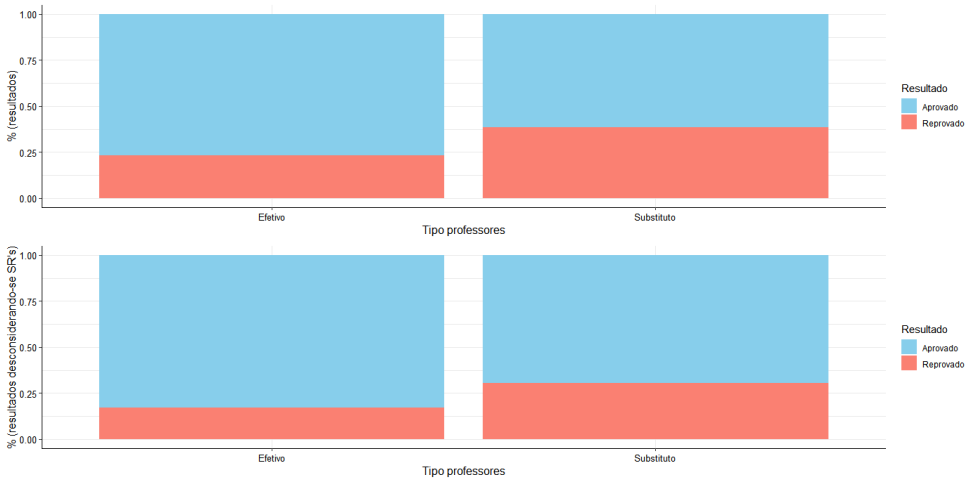
Nota-se, pela Figura 18, que - em ambos os casos considerados - os alunos do turno da noite obtiveram a maior proporção de reprovação de estudantes na sua primeira tentativa.

Figura 18: Resultado final dos estudantes segundo o turno de oferta da disciplina (2018-2019).



6.2.5 Tipo de professor

Figura 19: Resultado final do estudante segundo tipo do professor (2018-2019).

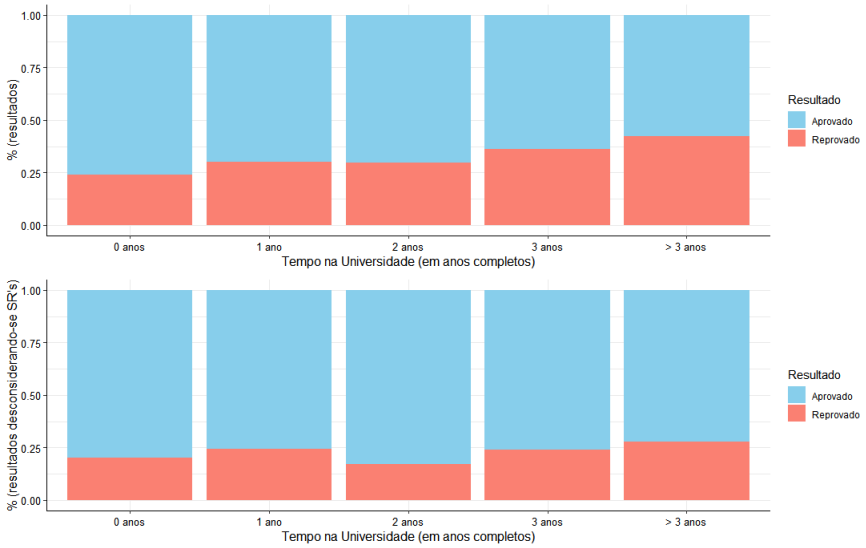


Pela Figura 19, nota-se que, em ambos os casos, a proporção de reprovação dos

alunos que fazem pela primeira vez a disciplina aparenta ser maior quando o professor que ministra a matéria foi um professor substituto.

6.2.6 Tempo do estudante na Universidade por resultado

Figura 20: Resultado final do estudante segundo a quantidade de anos que este está na Universidade (2018-2019).



Nessa parte, pode-se verificar que - de modo geral - quanto maior o tempo do estudante que tenta pela primeira vez cursar a matéria até final na Universidade, maior a proporção de reprovação desses. Já ao se desconsiderar casos de SR, nota-se que as proporções de reprovação não aparentam mudar muito de acordo com o tempo do estudante na Universidade.

6.3 Testes qui-quadrado

A tabela abaixo indica os resultados dos testes qui-quadrado entre a variável resposta e as explicativas sob as hipóteses (considerando-se apenas as observações relativas a 2018 e 2019)[9]:

$$H_0 : \text{não há associação entre as variáveis}$$

$$H_1 : \text{existe associação entre as variáveis}$$

Em ambos os casos, os p-valores obtidos foram muito pequenos - indicando que

[9]Nesse caso não foram considerados os cursos dos estudantes a fim de simplificar o processo de modelagem dos dados

Tabela 1: Resultados dos testes qui-quadrado entre a variável resposta e as explicativas

Testes qui-quadrado entre a variável resposta e as variáveis explicativas				
Caso Geral				
Variável explicativa	**% de aprovação**	**Estatística do teste**	**df**	**p-valor**
Turno				
manhã	77%	70.146	2	<0.001
tarde	74%			
noite	56%			
Tipo professor				
Efetivo	77%	55.482	1	<0.001
Substituto	61%			
Modalidade da disciplina				
Obrigatória	73%	47.978	2	<0.001
Optativa	69%			
Módulo Livre	48%			
Faltas				
[0%; 5%)	82%	578.8	5	<0.001
[5%; 10%)	85%			
[10%; 15%)	79%			
[15%; 20%)	75%			
[20%; 25%]	53%			
> 25%	0%			
Tempo na UnB				
0 anos	76%	23.54	4	<0.001
1 ano	70%			
2 anos	70%			
3 anos	64%			
mais que 3 anos	58%			
Desconsiderando SR's				
Variável explicativa	**% de aprovação**	**Estatística do teste**	**df**	**p-valor**
Turno				
manhã	82%	42.569	2	<0.001
tarde	81%			
noite	66%			
Tipo professor				
Efetivo	83%	47.328	1	<0.001
Substituto	69%			
Modalidade da disciplina				
Obrigatória	78%	35.005	2	<0.001
Optativa	78%			
Módulo Livre	58%			
Faltas				
[0%; 5%)	82%	85.133	4	<0.001
[5%; 10%)	85%			
[10%; 15%)	79%			
[15%; 20%)	75%			
[20%; 25%]	53%			
Tempo na UnB				
0 anos	80%	9.6984	4	0.04583
1 ano	76%			
2 anos	83%			
3 anos	76%			
mais que 3 anos	72%			

pode-se considerar a existência de evidências estatísticas suficientes para rejeitar a hipótese nula de idenpendência entre as variáveis, ou seja, pode-se considerar que existe associação entre essas variáveis - justificando, dessa fora, sua implementação nos respectivos modelos.

7 Modelagem: caso geral

Nesse tópico do estudo foram colocados os resultados referentes ao processo de modelagem dos dados dos estudantes (relativos aos anos 2018 e 2019) considerando-se as seguintes variáveis explicativas: turno no qual o estudante cursou a disciplina (turno de referência: Manhã); tempo - em anos completos - que o estudante estava na Universidade até cursar a matéria pela primeira vez; classificação dos professores (referência: Professores efetivos); modalidade da disciplina (referência: disciplina obrigatória) - lembrando que na variável resposta o sucesso é indicado pela aprovação do estudante. Nesse caso não foi considerado a variável porcentagem de faltas dos alunos na matéria porquanto existe um efeito de autocorelação entre esta e a variável resposta: alunos com mais de 25% de faltas automáticamente serão reprovados na disciplina.

Para construção do modelo foram utilizadas técnicas de Regressão Logística Multinível dado que a variável resposta é binária e a estrutura hierárquica dos dados - sendo que existem dois principais níveis no estudo: o nível do estudante e o nível das turmas. Os resultados para os coeficientes do modelo seguem:

Tabela 2: Resultados do modelo geral testado de acordo com os valores de referência.

Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	1.2397	0.14267	8.689	<0.001	
Tempo na UnB	-0.2028	0.04439	-4.569	<0.001	
Modalidade (optativa)	0.1815	0.21174	0.857	0.391	
Modalidade (módulo livre)	-0.7784	0.18644	-4.175	<0.001	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.2890	0.5376			
Turno (tarde)	0.0451	0.2124	-0.10		
Turno (noite)	0.2880	0.5367	-1.00	0.10	
Professor substituto	0.6333	0.7958	-0.95	-0.20	0.94

Pelos resultados acima, nota-se que as diferenças de resultado entre as modalidades obrigatória e optativa aparentam não ser significativas, porém a diferença entre modalidade Módulo Livre e obrigatória mostraram-se significantes - enquanto os demais resultados relativos aos efeitos fixos mostraram significativos. Com relação aos efeitos aleatórios, pôde-se verificar que as correlações calculadas foram relativamente altas entre as variáveis (dados com 2178 observações e 46 turmas).

Calculando-se os coeficientes de determinação R^2 marginal e condicional (referentes respectivamente ao modelo apenas com efeitos fixos e ao modelo completo) obteve-se os seguintes valores: 0.03 e 0.108 - indicando que o modelo completo explica cerca de 11% da variância dos dados - o modelo desconsiderando-se os efeitos aleatórios explica cerca de 3% dessa variância.

7.1 Validação do modelo

Tabela 3: Coeficientes dos modelos de teste e de validação de acordo com valores de referência para o caso geral.

Amostra teste					
Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	1.2671	0.15919	7.96	<0.001	
Tempo na UnB	-0.2040	0.06255	-3.261	0.001112	
Modalidade (optativa)	0.2229	0.31694	0.703	0.481928	
Modalidade (módulo livre)	0.2229	0.26195	-3.505	<0.001	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.2971	0.5451			
Turno (tarde)	0.3039	0.5513	0.22		
Turno (noite)	0.2972	0.5452	-1.00	-0.22	
Professor substituto	0.7355	0.8576	-0.82	-0.43	0.82
Amostra Validação					
Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	1.2981	0.13968	9.294	<0.001	
Tempo na UnB	-0.2009	0.06059	-3.315	<0.001	
Modalidade (optativa)	0.0621	0.28306	0.219	0.826435	
Modalidade (módulo livre)	-0.7432	0.27336	-2.719	0.006556	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.2725	0.522			
Turno (tarde)	0.2795	0.5287	-0.97		
Turno (noite)	0.4708	0.6861	-1.00	0.97	
Professor substituto	0.3506	0.5921	-1.00	0.97	1.00

Para verificar se os resultados obtidos de fato podem ser considerados representativos, realizou-se o processo de modelagem para amostras do banco de dados separando-as em dados de teste e de validação do modelo (cada uma contendo 1089 observações); os resultados dos coeficientes obtidos foram representados na Tabela 3.

Verificou-se que os resultados gerados pelos dados de teste e de validação aparentam serem relativamente semelhantes entre si. Posteriormente, utilizou-se os coeficientes gerados pelo modelo de teste para verificar quais os resultados obtidos a partir das observações referentes à amostra de validação - e comparar tais resultados com os verdadeiros valores observados na amostra de validação[10]:

Tabela 4: Resultados sobre a validação do modelo gerado pela amostra de teste para o caso geral.

Validação do modelo de teste			
Resultado/previsão	**Aprovação**	**Reprovação**	**Total**
Aprovação	446	311	757
Reprovação	121	178	299
Total	567	489	1056

[10]O valor 0.722 foi utilizado como ponto de corte para regra de decisão nesse caso - tal valor foi obtido através da curva ROC do modelo.

A partir dos resultados da tabela acima, verificou-se que taxa de acerto foi de aproximadamente 59% - indicando, assim, que o modelo gerado aparenta ser relativamente satisfatório. Dessa forma, optou-se por utilizar o banco de dados completo a fim de fazer a modelagem final dos dados.

7.2 Razões de chances

Tabela 5: Estimativas das Razões de Chances com base nos valores de referência para o caso geral.

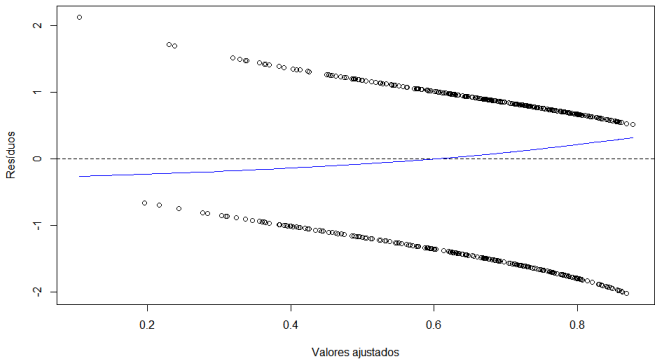
Preditores	Razão de chances	IC(95%)	p-valor
Intercepto	3.45	2.61 – 4.57	<0.001
Tempo na UnB	0.82	0.75 – 0.89	<0.001
Disciplina Optativa	1.20	0.79 – 1.82	0.391
Disciplina Módulo Livre	0.46	0.32 – 0.66	<0.001

A partir dos resultados da Tabela 5, pôde-se notar que o aumento em um ano completo na Universidade acarreta na diminuição em 0.82 vezes na chance de aprovação na primeira tentativa dos estudantes na disciplina Estatística Aplicada; se o estudante cursar a matéria como Módulo Livre as chances de aprovação são 0.46 vezes a chance de aprovação de estudantes que cursam como obrigatória (54% menor).

7.3 Diagnóstico do modelo

Resíduos

Figura 21: Gráfico de resíduos e valores ajustados para o caso geral.

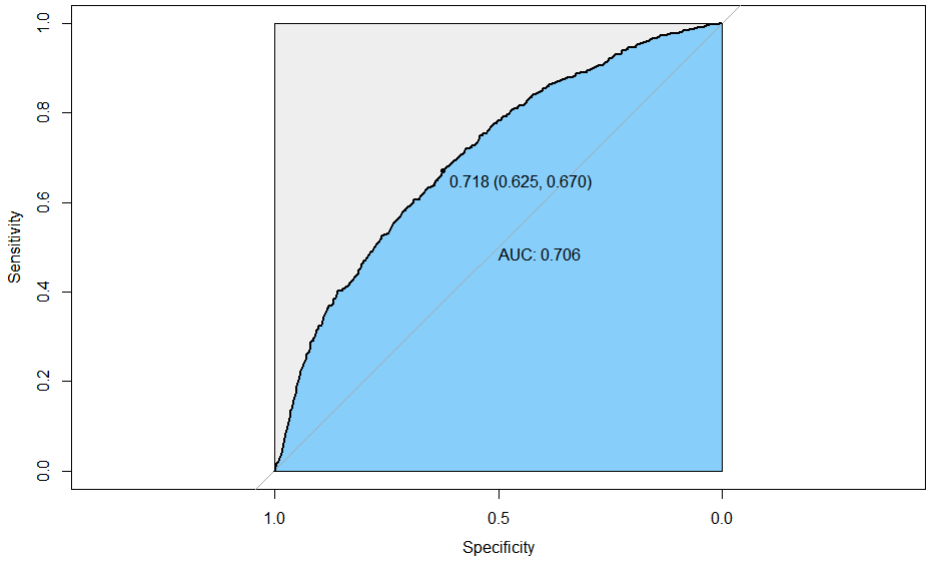


Pela Figura 21, notou-se que uma suavização leve do gráfico dos resíduos em relação aos valores predidos resultou aproximadamente em uma linha horizontal com intercepto no zero - indicando então que o modelo aparenta estar ajustado de forma adequada aos dados.

Curva ROC

A partir da Figura 22, notou-se que - dado que a área sobre a curva parece relativamente grande - o poder preditivo do modelo parece ser de fato satisfatório; além disso, considerou-se 0.706 como sendo um ponto de corte adequado na escolha da regra de previsão.

Figura 22: Gráfico da curva ROC para o caso geral.



Analisando, agora, a matriz de confusão e os resultados gerados pela previsão obteve-se:

Tabela 6: Resultados obtidos pela Matriz de Confusão para o caso geral.

	Valor observado	
Valor estimado	**Y = 1**	**Y = 0**
1	1101	275
0	400	335

Medidas	**Valores**
Acurácia	0.6802
IC (95%) para acurácia	(0.6599; 0.7001)
Sensibilidade	0.7335
Especificidade	0.5492

A matriz de confusão retoma uma boa acurácia total do modelo em 68% - sendo que a sensibilidade do modelo é consideravelmente elevada (73%) enquanto sua especificidade é de aproximadamente (55%).

8 Modelagem: caso sem SR's

Assim como na seção anterior, este tópico contempla o processo de modelagem - utilizando técnicas de regressão logística multinível - dos dados relativos aos anos 2018 e 2019 - desconsiderando, no entanto, os casos de SR's no banco de dados - entendendo então que estes representam uma forma de abandono da disciplina por parte do estudante. Foram consideradas as seguintes variáveis para o modelo: turno no qual o estudante cursou a disciplina (turno de referência: Manhã); porcentagem de faltas do estudante na disciplina; classificação dos professores (referência: Professores efetivos); modalidade da disciplina (referência: disciplina obrigatória)[11]. A imagem a seguir mostra os coeficientes calculados para o modelo[12]:

Tabela 7: Coeficientes do modelo sem SR's testado de acordo com os valores de referência.

Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	2.4916	0.2314	10.767	<0.001	
Modalidade (optativa)	0.2889	0.2519	1.147	0.2515	
Modalidade (módulo livre)	-0.9860	0.2219	-4.444	<0.001	
Faltas [5%; 10%)	-0.5236	0.264	-1.994	0.0462	
Faltas [10%; 15%)	-1.0136	0.2559	-3.961	<0.001	
Faltas [15%; 20%)	-1.4158	0.2867	-4.938	<0.001	
Faltas [20%; 25%]	-2.5396	0.2809	-9.040	<0.001	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.5174	0.7193			
Turno (tarde)	0.0853	0.292	-1.00		
Turno (noite)	0.6764	0.8224	-0.99	0.99	
Professor substituto	1.6429	1.2817	-0.72	0.73	0.8

Calculando-se os coeficientes de determinação R^2 marginal e condicional (referentes respectivamente ao modelo apenas com efeitos fixos e ao modelo completo) obteve-se os seguintes valores: 0.153 e 0.268 - indicando que o modelo completo explica cerca de 15.3% da variância dos dados - o modelo desconsiderando-se os efeitos aleatórios explica cerca de 26.8% dessa variância.

8.1 Validação do modelo

Analogamente como feito no modelo geral, realizou-se um processo de subamostragem do banco de dados dividindo-o ao meio a fim de testar a validação do modelo sem SR's. Os coeficientes gerados pelos dois modelos acima foram relativamente semelhantes

[11] Nesse caso, assim como na modelagem anterior, o sucesso é representado pela aprovação do estudante na disciplina.

[12] Dados considerando 1983 observações com 46 turmas.

entre si - os resultados podem ser observados na Tabela 8.

Tabela 8: Coeficientes dos modelos de teste e de validação de acordo com valores de referência para o caso sem SR's.

Amostra teste					
Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	2.5026	0.293	8.542	<0.001	
Modalidade (optativa)	0.2347	0.3716	0.632	0.5276	
Modalidade (módulo livre)	-1.0613	0.3081	-3.445	<0.001	
Faltas [5%; 10%)	-0.5508	0.3501	-1.573	0.1157	
Faltas [10%; 15%)	-0.8700	0.345	-2.52	0.0117	
Faltas [15%; 20%)	-1.4802	0.3891	-3.80	<0.001	
Faltas [20%; 25%]	-2.5004	0.3885	-6.437	<0.001	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.3129	0.5594			
Turno (tarde)	0.4107	0.6408	-1.00		
Turno (noite)	1.1555	1.0749	-1.00	1.00	
Professor substituto	1.2414	1.1142	-1.00	1.00	1.00
Amostra Validação					
Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**	
Intercepto	2.1071	0.23136	9.107	<0.001	
Modalidade (optativa)	0.0352	0.34624	0.102	0.9190	
Modalidade (módulo livre)	-0.9819	0.32772	-2.996	0.0027	
Faltas [5%; 10%)	-0.0373	0.33423	-0.112	0.9111	
Faltas [10%; 15%)	-0.7051	0.29465	-2.393	0.0167	
Faltas [15%; 20%)	-0.8217	0.34842	-2.358	0.0184	
Faltas [20%; 25%]	-2.1537	0.35172	-6.123	<0.001	
Efeitos Aleatórios (id turma)	**Variância**	**Desvio padrão**	**Correlação**		
Intercepto	0.4799	0.6928			
Turno (tarde)	0.1714	0.414	-1.00		
Turno (noite)	3.1749	1.7818	-0.17	0.17	
Professor substituto	0.6248	0.7904	-0.22	0.22	-0.92

Utilizando o modelo de teste para verificar quais os resultados obtidos a partir das observações referentes à amostra de validação - e compararando tais resultados com os verdadeiros valores observados na amostra de validação - observou-se que a taxa de acerto foi de aproximadamente 66% - indicando que aparentemente o modelo gerado responde corretamente em pouco mais da metade das vezes (resultados mostrados na tabela 9). Dessa forma, optou-se por utilizar o banco de dados completo a fim de fazer a modelagem final dos casos sem SR's.[13]:

Tabela 9: Resultados sobre a validação do modelo gerado pela amostra de teste para o caso sem SR's.

Validação do modelo de teste			
Resultado/previsão	**Aprovação**	**Reprovação**	**Total**
Aprovação	508	249	757
Reprovação	74	131	205
Total	582	380	962

[13]O valor 0.8 foi utilizado como ponto de corte para regra de decisão nesse caso - tal valor foi obtido através da curva ROC do modelo.

8.2 Razões de chances

A partir das estimativas acerca das razões de chances geradas pelo modelo sem SR's, verificou-se que os resultados mais siginificativos acerca da porcentagem de faltas do estudante na disciplina mostram que as chances de aprovação deste diminui com o aumento na quantidade de faltas (a partir do valor de referência de [0%, 5%) de faltas; se o estudante cursar a matéria como Módulo Livre as chances de aprovação são de 0.37 vezes a de estudantes que cursam como obrigatória (63% menor). A tabela abaixo ilustra tal cenário:

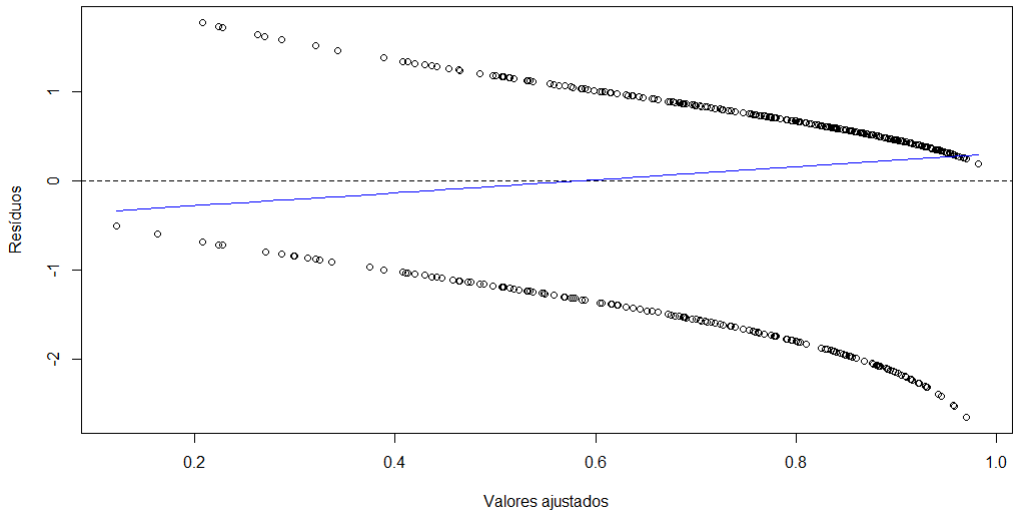
Tabela 10: Estimativas das Razões de Chances com base nos valores de referência para o caso sem SR's.

Preditores	Razão de chances	IC(95%)	p-valor
Intercepto	12.08	7.68 – 19.01	<0.001
Disciplina Optativa	1.33	0.81 – 2.19	0.252
Disciplina Módulo Livre	0.37	0.24 – 0.58	<0.001
Faltas [5%; 10%)	0.59	0.35 – 0.99	0.046
Faltas [10%; 15%)	0.36	0.22 – 0.60	<0.001
Faltas [15%; 20%)	0.24	0.67 – 1.76	<0.001
Faltas [20%; 25%]	0.08	0.26 – 0.60	<0.001

8.3 Diagnóstico do modelo

Resíduos

Figura 23: Gráfico de resíduos e valores ajustados para o caso sem SR's.



Pela Figura 23, notou-se que uma suavização leve do gráfico dos resíduos em

relação aos valores predidos resultou aproximadamente em uma linha horizontal com intercepto no zero - indicando então que o modelo aparenta estar ajustado de forma adequada aos dados.

Curva ROC

A partir da Figura 24, notou-se que - dado que a área sobre a curva parece relativamente grande - o poder preditivo do modelo parece ser de fato satisfatório; além disso, considerou-se 0.783 como sendo um ponto de corte adequado na escolha da regra de previsão. Analisando a matriz de confusão e os resultados gerados pela previsão obteve-se uma boa acurácia total do modelo em 78% - sendo que a sensibilidade do modelo é consideravelmente elevada (86%) enquanto sua especificidade mostrou-se relativamente baixa (50%) os resultados estão presentes na Tabela 11.

Figura 24: Gráfico da curva ROC para o caso sem SR's.

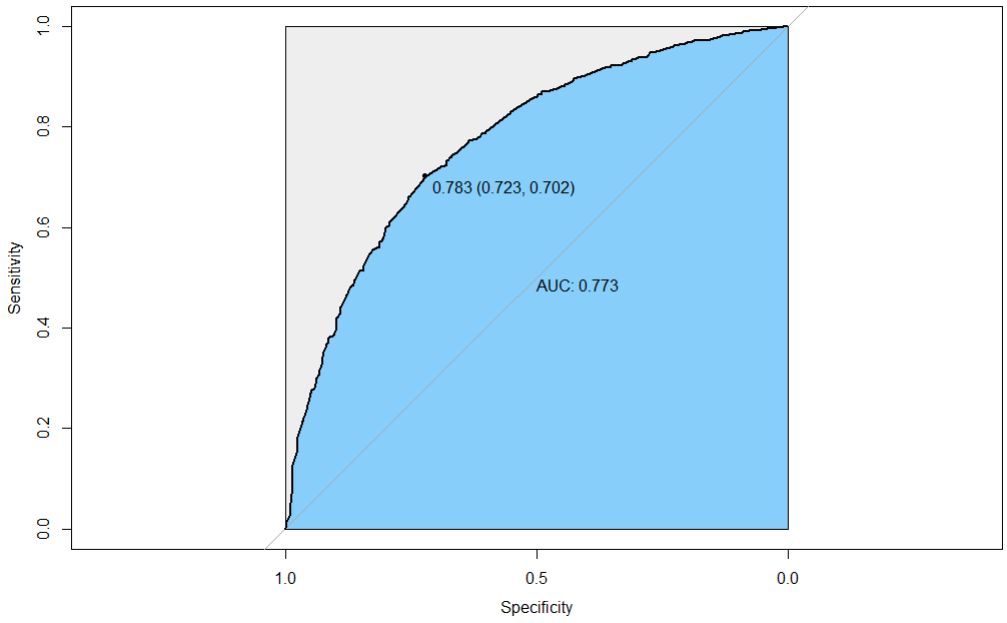


Tabela 11: Resultados obtidos pela Matriz de Confusão para o caso sem SR's.

	Valor observado	
Valor estimado	**Y = 1**	**Y = 0**
1	1291	211
0	210	211

Medidas	**Valores**
Acurácia	0.7811
IC (95%) para acurácia	(0.7619; 0.7994)
Sensibilidade	0.8601
Especificidade	0.5

9 Análise de dados sobre perfil do estudante

Nesse tópico, foi elaborado um estudo exploratório a fim de verificar possíveis características sociodemograficas - e de tragetória pré-uniersitária - que podem estar associadas a probabilidade de aprovação dos estudantes em Estatística Aplicada. [14]

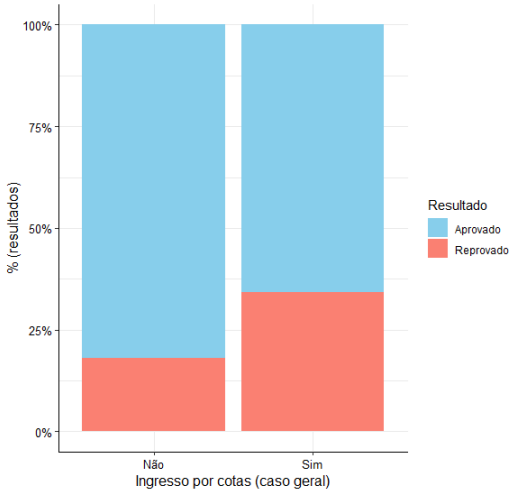
9.1 Análise exploratória

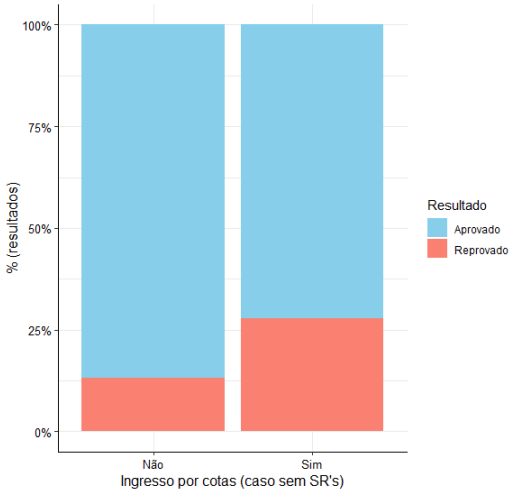
9.1.1 Forma de ingresso

Figura 25: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo sua forma de ingresso na Universidade.

Resultado por forma de ingresso (caso geral)			
	Cotas		
Resultado	**Sim**	**Não**	**Total**
Aprovado	439	414	853
Reprovado	97	215	312
Total	536	629	1165

Resultado por forma de ingresso (caso sem SR's)			
	Cotas		
Resultado	**Sim**	**Não**	**Total**
Aprovado	439	414	853
Reprovado	67	160	227
Total	506	574	1080





Pelas imagens acima, pôde-se notar que em ambos os casos considerados - com e sem SR na análise - a proporção de reprovação dos estudantes aparenta ser consideravelmente maior em estudantes que entraram na Universidade por algum sistema de cotas do que em alunos que entraram pelo sistema universal. No caso geral as chances de aprovação para cotistas e não cotistas são respectivamente 52% e 22%; no caso sem os SR's esses valores são respectivamente 39% e 15%.

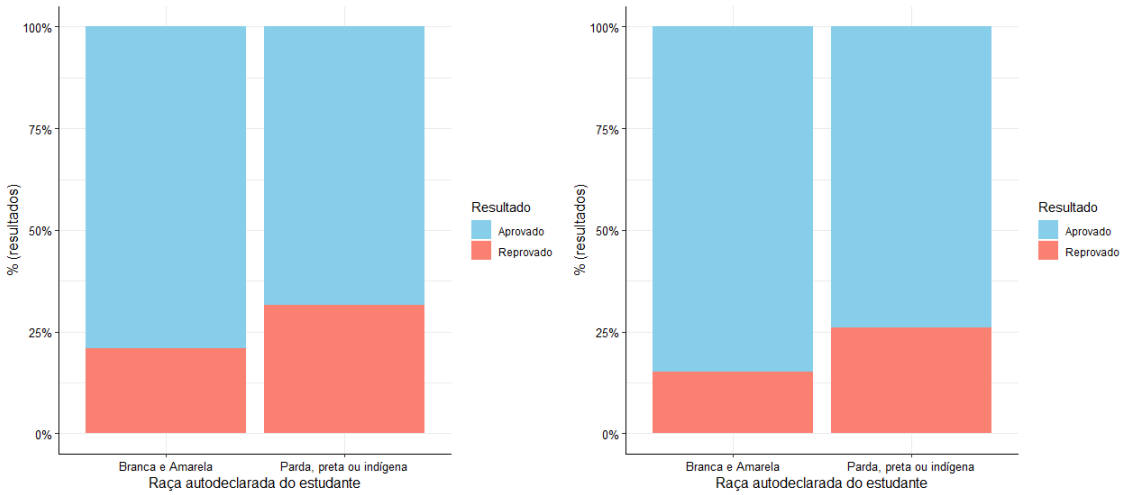
[14] Dados relativos a estudantes ingressantes em 2017 e 2018 na Universidade.

9.1.2 Raça autodeclarada do estudante

Figura 26: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo sua raça autodeclarada.

Resultado por raça autodeclarada (caso geral)			
	Raça		
Resultado	**brancos ou amerelos**	**pretos, pardos ou indígeneas**	**Total**
Aprovado	380	399	779
Reprovado	100	184	284
Total	480	583	1063

Resultado por raça autodeclarada (caso sem SR's)			
	Raça		
Resultado	**brancos ou amerelos**	**pretos, pardos ou indígeneas**	**Total**
Aprovado	380	399	779
Reprovado	68	141	209
Total	448	540	988



Os gráficos e tabelas acima indicam que a proporção de reprovação é maior em alunos que se autodeclararam pretos, pardos ou indígeneas (em ambos os casos) do que os que se declararam brancos ou amarelos. As chances de reprovação no caso geral para pessoas brancas e amarelas é de cerca de 26%, enquanto para alunos pretos, pardos ou indígenas vale aproximadamente 46%; no caso desconiserando-se os SR's, essas chances valem respectivamente 18% e 35%. [15]

[15] Com relação a essa variável foram desconsiderados 102 e 92 casos devido a valores faltantes com relação ao caso geral e ao caso sem SR's respectivamente.

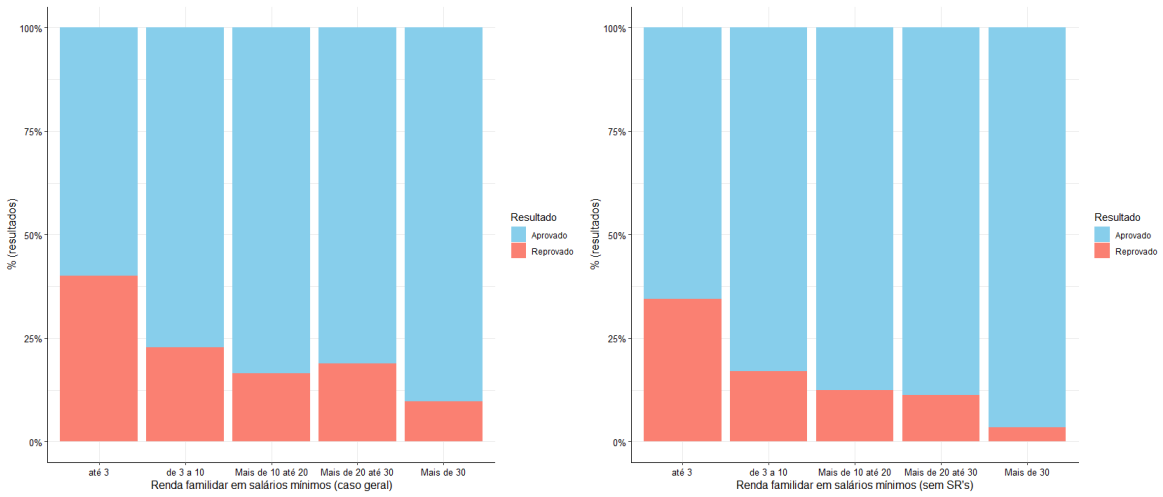
9.1.3 Renda familiar do estudante

Com relação a renda familiar informada pelo estudante, notou-se que no caso geral aparenta existir uma relação entre tal informção e a variável explicativa - sendo que quanto menor a renda familiar dos estudantes, aparentemente maior é a proporção de reprovação dos alunos em sua primeira tentativa na disciplina. Já ao considerar apenas os casos sem SR's, notou-se que esse resultado mostrou-se de forma ainda mais evidente nos dados. [16]

Figura 27: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo sua renda familiar.

Resultado por renda familiar (caso geral)						
	Renda em salários mínimos					
Resultado	**Até 3**	**de 3 a 10**	**Mais de 10 até 20**	**Mais de 20 até 30**	**Mais de 30**	**Total**
Aprovado	222	306	147	47	56	778
Reprovado	148	90	29	11	6	284
Total	**370**	**396**	**176**	**58**	**62**	**1062**

Resultado por renda familiar (caso sem SR's)						
	Renda em salários mínimos					
Resultado	**Até 3**	**de 3 a 10**	**Mais de 10 até 20**	**Mais de 20 até 30**	**Mais de 30**	**Total**
Aprovado	222	306	147	47	56	778
Reprovado	117	63	21	6	2	209
Total	**339**	**369**	**168**	**53**	**58**	**987**



[16] Com relação a essa variável foram desconsiderados 103 e 93 casos devido a valores faltantes com relação ao caso geral e ao caso sem SR's respectivamente.

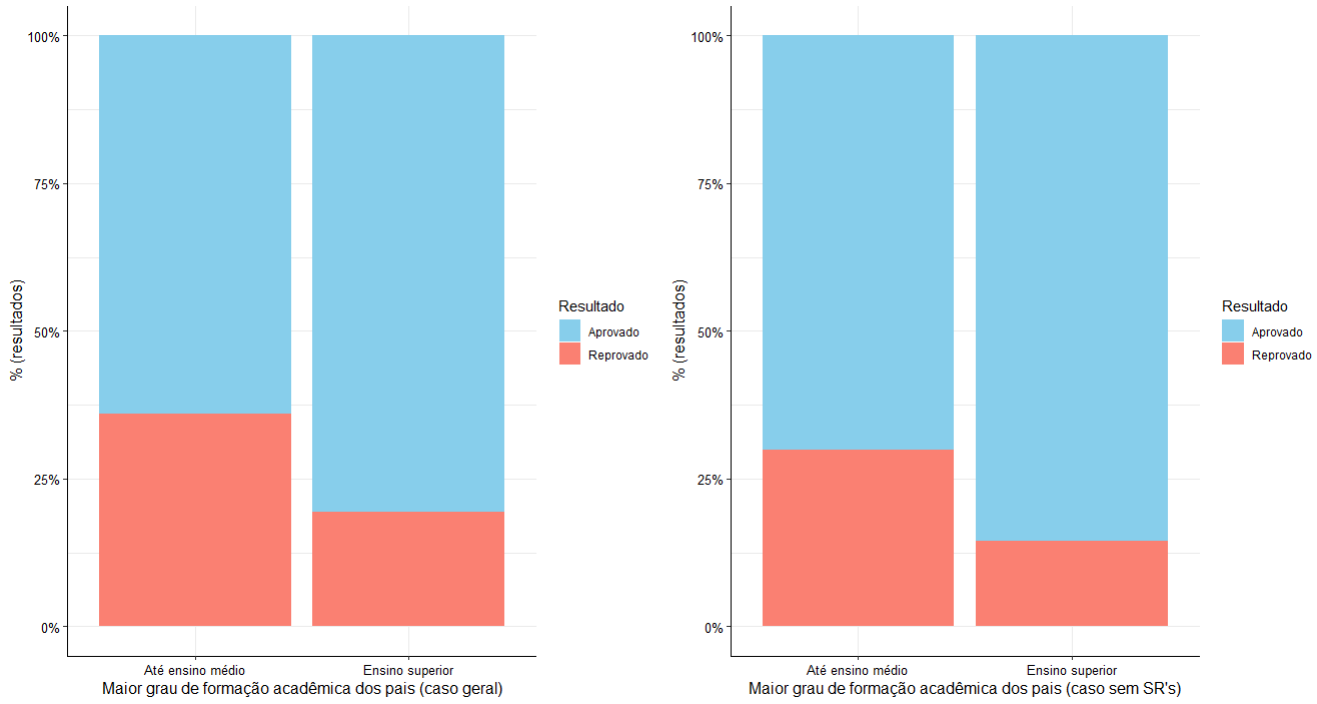
9.1.4 Formação acadêmica dos pais

Nesse caso, pôde-se notar que estudantes cujos ambos os pais não possuem formação superior apresentam maior proporção de reprovação. As chances de reprovação no caso geral para alunos cujos pais não tem formação superior vale cerca de 56%; para alunos os quais pelo menos um dos pais tem formação superior esse valor é de cerca de 24%. Para os casos sem SR's essas chances valem respectivamente 43% e 17%.[17]

Figura 28: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo a maior formação acadêmica de um dos pais do estudante.

Resultado por formação dos pais			
	Maior formação de um dos pais		
Resultado	**Até o ensino médio**	**Superior**	**Total**
Aprovado	269	476	745
Reprovado	151	115	266
Total	420	591	1011

Resultado por formação dos pais (caso sem SR's)			
	Maior formação de um dos pais		
Resultado	**Até o ensino médio**	**Superior**	**Total**
Aprovado	269	476	745
Reprovado	115	81	196
Total	384	557	941



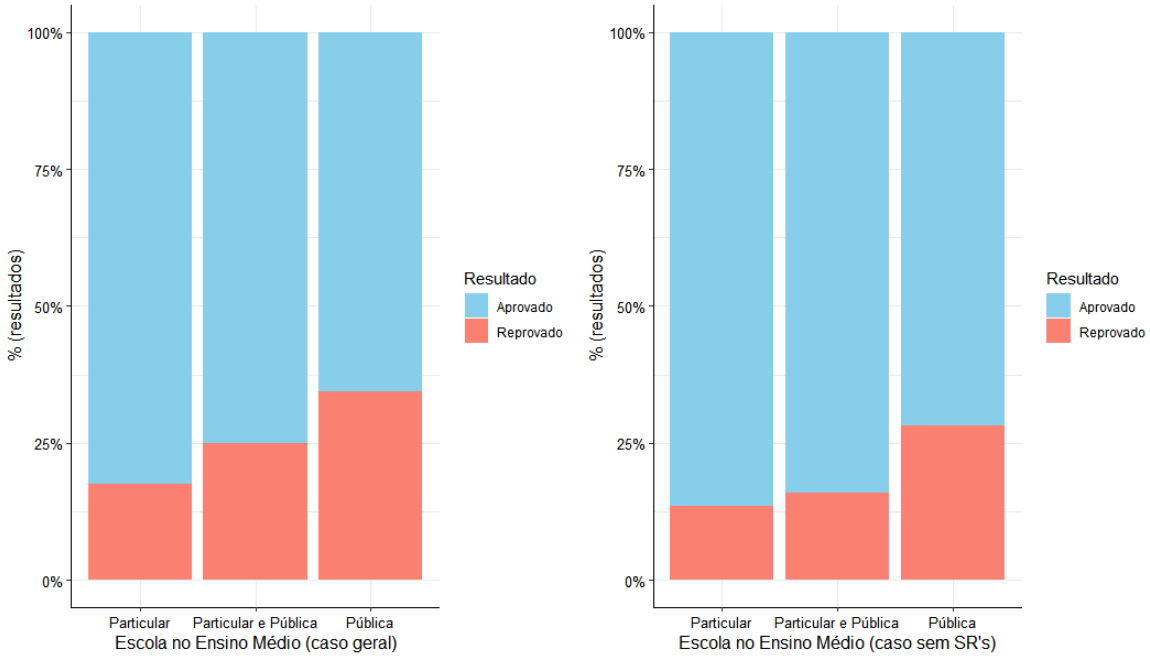
[17] Com relação a essa variável foram desconsiderados 154 e 139 casos devido a valores faltantes com relação ao caso geral e ao caso sem SR's respectivamente.

9.1.5 Ensino médio do estudante

Figura 29: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo o tipo de escola durante o Ensino Médio.

Resultado por tipo de escola no Ensino Médio (caso geral)				
	Escola			
Resultado	**Particular**	**Particular e pública**	**Pública**	**Total**
Aprovado	373	42	363	778
Reprovado	79	14	191	284
Total	**452**	**56**	**554**	**1062**

Resultado por tipo de escola no Ensino Médio (caso sem SR's)				
	Escola			
Resultado	**Particular**	**Particular e pública**	**Pública**	**Total**
Aprovado	373	42	363	778
Reprovado	58	8	143	209
Total	**431**	**50**	**506**	**987**



Pelos resultados acima, no caso geral verificou-se que as menores proporções de reprovação foram em estudantes que fizeram o ensino médio apenas em escolas particulares, seguido por quem fez tanto em escolas particulares quanto públicas. Já ao analisar os casos sem SR's, pôde-se notar que os resultados referentes as duas categorias citadas anteriormente aparentam ser bastante semelhantes entre si - diferindo apenas ao se comparar com os resultados de estudantes de escolas públicas (os quais tiveram maior proporção de reprovação na matéria durante sua primeira tentativa).[18]

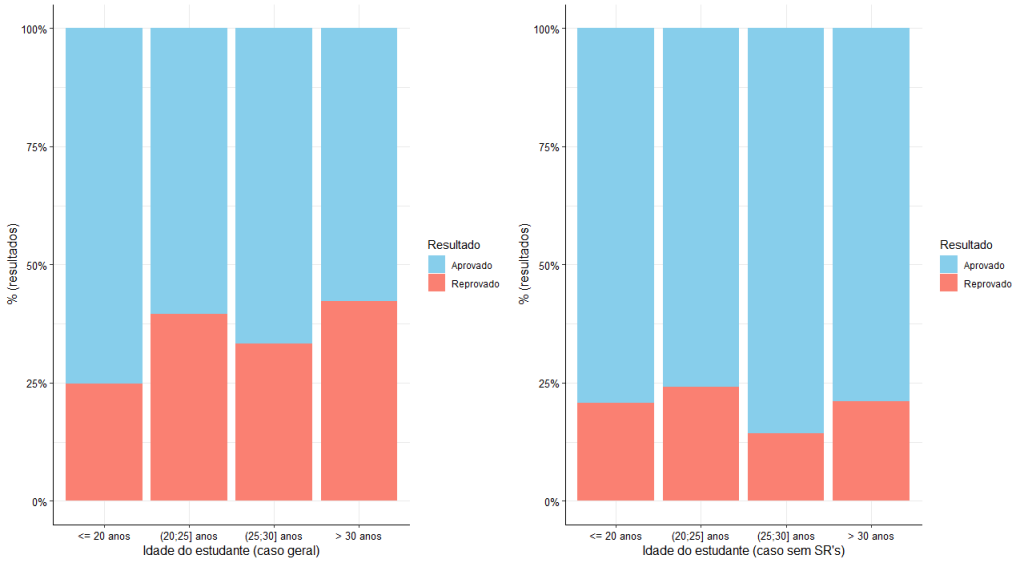
[18]Com relação a essa variável foram desconsiderados 103 e 93 casos devido a valores faltantes com

9.1.6 Idade do estudante

Figura 30: Resultado final dos estudantes na sua primeira tentativa na disciplina segundo o faixa etária do estudante.

Resultado por idade do estudante (caso geral)					
	Idade				
Resultado	**Até 20 anos**	**de 21 a 25 anos**	**de 26 a 30 anos**	**mais de 30 anos**	**Total**
Aprovado	754	72	12	15	841
Reprovado	248	47	6	11	306
Total	**1002**	**119**	**18**	**26**	**1147**

Resultado por idade do estudante (caso sem SR's)					
	Idade				
Resultado	**Até 20 anos**	**de 21 a 25 anos**	**de 26 a 30 anos**	**mais de 30 anos**	**Total**
Aprovado	754	72	12	15	841
Reprovado	198	23	2	4	225
Total	**952**	**95**	**14**	**19**	**1066**



Ano analisar essa variável, notou-se que - no caso geral - o aumento de idade dos estudantes na sua primeira tentativa na disciplina aparentava indicar em maiores proporções de reprovação; já ao se analisar os casos sem SR's esse comportamento não aparentou ser presente. [19]

A variável relativa ao sexo informado pelo estudante foi analisada e a diferença entre as diferentes categorias não mostro-se significativa - por isso, optou-se por não incluir tal variável na análise exploratória e no processo de modelagem.

relação ao caso geral e ao caso sem SR's respectivamente.

[19] Com relação a essa variável foram desconsiderados 18 e 14 casos devido a valores faltantes com relação ao caso geral e ao caso sem SR's respectivamente.

9.2 Modelagem

Para fins exploratórios apenas contruiram-se modelos (para o caso geral e para o caso sem SR's) para verificar quais variáveis em estudo aparentaram influenciar mais nos resultados finais dos estudantes em sua primeira tentativa na disciplina.

9.2.1 Caso Geral

Retirando-se efeitos de correlação fortes entre as variáveis em estudo - para o caso geral - as variáveis finais escolhidas para o modelo foram: variável indicadora de se o estudante entrou por algum sistema de cotas ou pelo sistema universal; renda familiar do estudante (categorizada); idade (em anos) dos estudantes; e modalidade da disciplina para o curso do estudante - além da variável Turno no nível das turmas dos estudantes. Essas variáveis foram controladas pelas turmas identificadas no banco de dados e os coeficientes encontrados para o modelo foram: [20]

Tabela 12: Modelo considerando-se todas as observações.

Efeitos fixos	**Estimativa**	**Erro padrão**	**Z**	**p-valor**
Intercepto	2.58	0.47	5.44	<0.001
Cotas (sim)	-0.74	0.21	-3.56	<0.001
Renda (3 a 10 salários mínimos)	0.70	0.19	3.61	<0.001
Renda (mais de 10 a 20 salários mínimos)	0.82	0.29	2.84	0.0045
Renda (mais de 20 a 30 salários mínimos)	0.62	0.42	1.46	0.1436
Renda (mais de 30 salários mínimos)	1.50	0.52	2.87	0.0041
Modalidade (optativa)	-0.06	0.02	-2.80	0.0050
Modalidade (Módulo Livre)	-1.35	0.37	-3.63	<0.001
Idade	1.55	0.32	-4.84	<0.001
Efeitos aleatórios (turma)	**Variância**	**Desvio padrão**		
Intercepto	0.220	0.469		
Turno (tarde)	0.621	0.788		
Turno (noite)	1.321	1.149		

Analisando as razões de chances calculadas para o presente modelo (tabela 13), notou-se que a chance de aprovação de um aluno cotista em sua primeira tentativa na disciplina aparenta ser aproximadamente metade da de um aluno que ingressou pelo Sistema Universal; quanto maior a renda familiar do estudante, maior sua chance de aprovação; a chances de aprovação de alunos que cursam a disciplina como módulo livre ou como optativa aparentam ser bem menores do que a chance de alunos que cursam como matéria obrigatória; e o aumento em um ano na idade do estudante aparenta diminuir suas chances de aprovação na primeira tentativa em Estatística Aplicada.

[20] Após retirar as observações faltantes, o banco de dados utilizado nesse modelo contou com 1011 observações.

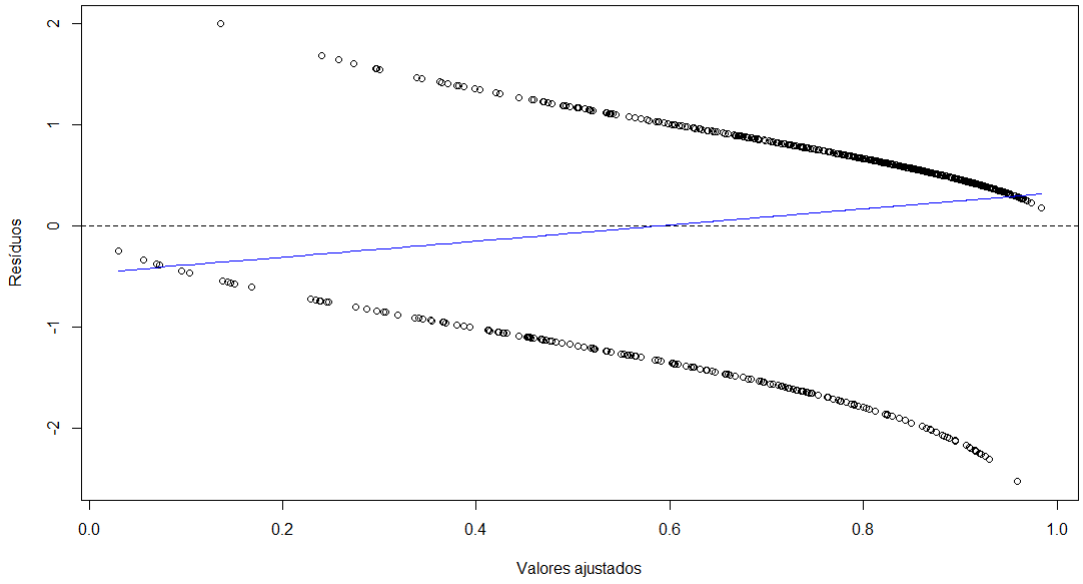
Tabela 13: Razão de chances considerando-se todas as observações.

Variáveis explicativas	Razão de chances	IC(95%)	p-valor
Intercepto	13.14	5.20 – 33.21	<0.001
Cotas (sim)	0.48	0.32 – 0.72	<0.001
Renda (3 a 10 salários mínimos)	2.01	1.38 – 2.94	<0.001
Renda (mais de 10 a 20 salários mínimos)	2.28	1.29 – 4.03	0.005
Renda (mais de 20 a 30 salários mínimos)	1.86	0.81 – 4.27	0.144
Renda (mais de 30 salários mínimos)	4.49	1.61 – 12.49	0.004
Modalidade (optativa)	0.26	0.12 – 0.54	0.005
Modalidade (Módulo Livre)	0.21	0.11 – 0.40	<0.001
Idade	0.94	0.90 – 0.98	<0.001

9.2.2 Diagnóstico do modelo

Resíduos

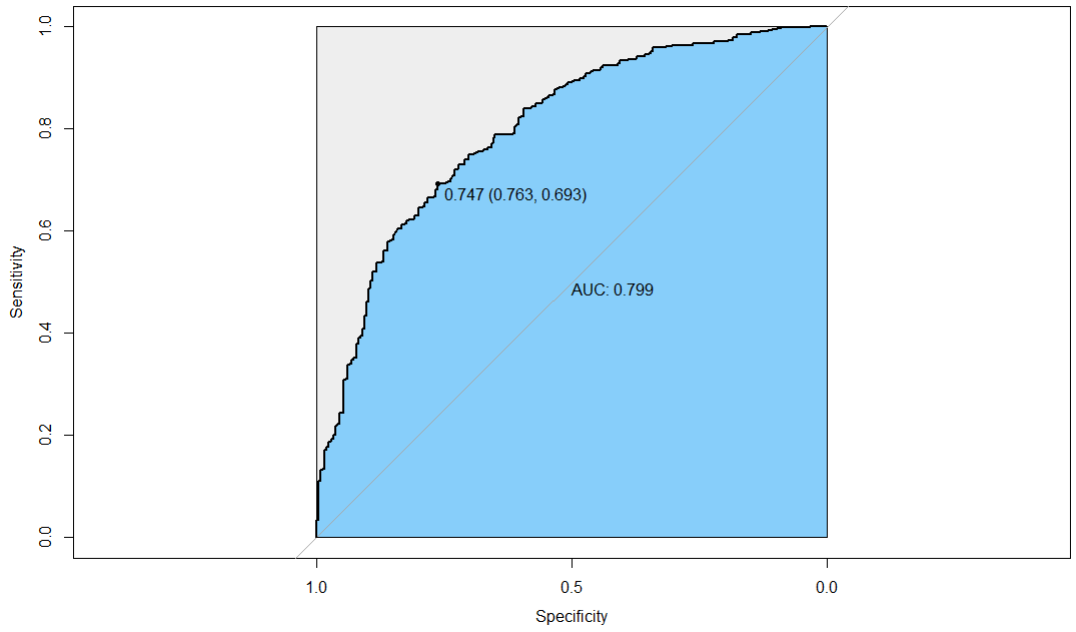
Figura 31: Gráfico de resíduos considerando-se todas as observações.



Pela Figura 31, notou-se que uma suavização leve do gráfico dos resíduos em relação aos valores predidos resultou aproximadamente em uma linha horizontal com intercepto no zero - indicando então que o modelo - caso os dados utilizados fossem representativos - aparenta estar ajustado de forma adequada aos dados.

Curva ROC

Figura 32: Gráfico da curva ROC considerando-se todas as observações.



A partir da Figura 32, notou-se que - dado que a área sobre a curva parece relativamente grande - o poder preditivo do modelo parece ser de fato satisfatório; além disso, considerou-se 0.747 como sendo um ponto de corte adequado na escolha da regra de previsão. Verificando-se os resultados gerados pela matriz de confusão:

Tabela 14: Matriz de confusão considerando-se todas as observações.

	Valor observado	
Valor estimado	**Y = 1**	**Y = 0**
1	516	63
0	229	203

Medidas	**Valores**
Acurácia	0.7112
IC (95%) para acurácia	(0.6822; 0.739)
Sensibilidade	0.6926
Especificidade	0.7632

A matriz de confusão retoma uma boa acurácia total do modelo em 71.12% -sendo que a sensibilidade do modelo vale (69.26%) enquanto sua especificidade e de aproximadamente (76.32%).

9.2.3 Desconsiderando-se SR's

Retirando-se efeitos de correlação fortes entre as variáveis em estudo - para o caso geral - as variáveis finais escolhidas para o modelo foram: variável indicadora de se o estudante entrou por algum sistema de cotas ou pelo sistema universal; renda familiar do estudante (categorizada); e modalidade da disciplina para o curso do estudante - além da variável Turno no nível das turmas dos estudantes. Essas variáveis foram controladas pelas turmas identificadas no banco de dados e os coeficientes encontrados para o modelo foram:[21]

Tabela 15: Modelo para o caso sem SR's.

Efeitos fixos	Estimativa	Erro padrão	Z	p-valor
Intercepto	1.81	0.29	6.31	<0.001
Cotas (sim)	-0.89	0.24	-3.72	<0.001
Renda (3 a 10 salários mínimos)	0.83	0.21	3.87	<0.001
Renda (mais de 10 a 20 salários mínimos)	1.00	0.33	3.03	0.0024
Renda (mais de 20 a 30 salários mínimos)	1.07	0.53	2.02	0.0439
Renda (mais de 30 salários mínimos)	2.30	0.79	2.91	0.0036
Modalidade (optativa)	-0.97	0.46	-2.12	0.0344
Modalidade (Módulo Livre)	-1.62	0.36	-4.55	<0.001
Efeitos aleatórios (turma)	**Variância**	**Desvio padrão**		
Intercepto	0.244	0.494		
Turno (tarde)	0.861	0.928		
Turno (noite)	1.969	1.403		

Tabela 16: Razões de chances para o caso sem SR's.

Variáveis explicativas	Razão de chances	IC(95%)	p-valor
Intercepto	6.10	3.48 – 10.71	<0.001
Cotas (sim)	0.41	0.26 – 0.66	<0.001
Renda (3 a 10 salários mínimos)	2.29	1.51 – 3.48	<0.001
Renda (mais de 10 a 20 salários mínimos)	2.71	1.42 – 5.18	0.002
Renda (mais de 20 a 30 salários mínimos)	2.90	1.03 – 8.19	0.044
Renda (mais de 30 salários mínimos)	10.00	2.13 – 47.04	0.004
Modalidade (optativa)	0.38	0.15 – 0.93	0.034
Modalidade (Módulo Livre)	0.20	0.10 – 0.40	<0.001

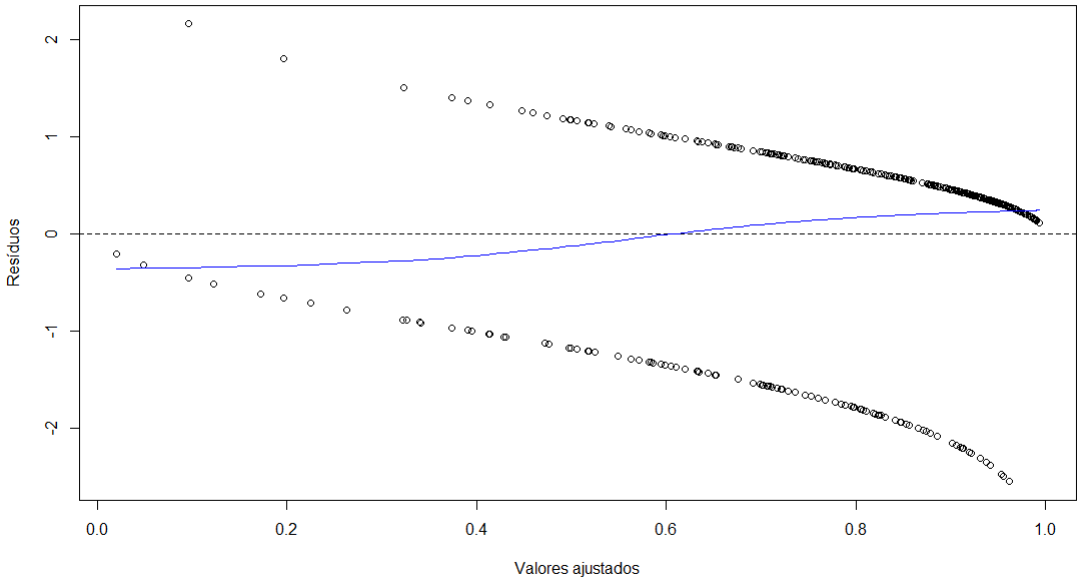
Analisando as razões de chances calculadas para o presente modelo (Tabela 16), notou-se que - assim como visto pelo modelo anterior - a chance de aprovação de um aluno cotista em sua primeira tentativa na disciplina aparenta ser aproximadamente 60% menor que a de um aluno que ingressou pelo Sistema Universal; quanto maior a renda familiar do estudante, maior sua chance de aprovação; a chances de aprovação de alunos

[21] Após retirar as observações faltantes, o banco de dados utilizado nesse modelo contou com 941 observações.

que cursam a disciplina como módulo livre ou como optativa aparentam ser bem menores do que a chance de alunos que cursam como matéria obrigatória - sendo que alunos de módulo livre apresentam ainda menores chances de aprovação.

Resíduos

Figura 33: Gráfico dos resíduos para o caso sem SR's.



Pelo gráfico acima, notou-se que uma suavização leve do gráfico dos resíduos em relação aos valores predidos resultou aproximadamente em uma linha horizontal com intercepto no zero - indicando então que o modelo - caso os dados utilizados fossem representativos - aparenta estar ajustado de forma adequada aos dados.

Curva ROC

Tabela 17: Tabelas sobre a matriz de confusão para o caso sem SR's.

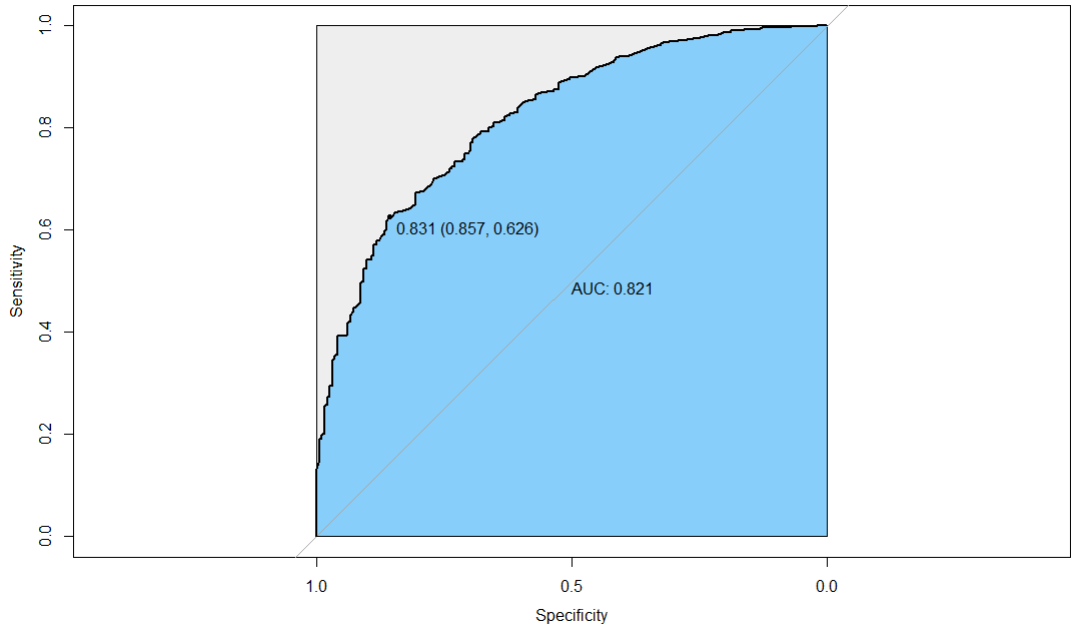
	Valor observado	
Valor estimado	**Y = 1**	**Y = 0**
1	461	27
0	284	169

Medidas	**Valores**
Acurácia	0.6695
IC (95%) para acurácia	(0.6384; 0.6995)
Sensibilidade	0.6188
Especificidade	0.8622

A matriz de confusão retoma uma boa acurácia total do modelo em 66.95% -sendo que a sensibilidade do modelo vale 61.88% enquanto sua especificidade vale aproximadamente 86.22%.

A partir da Figura 34 (curva ROC), notou-se que - dado que a área sobre a curva parece relativamente grande - o poder preditivo do modelo parece ser de fato satisfatório; além disso, considerou-se 0.831 como sendo um ponto de corte adequado na escolha da regra de previsão.

Figura 34: Gráfico da curva ROC para o caso sem SR's.



10 Conclusão

Com base nas análises realizadas no presente estudo, pôde-se verificar que o resultado final do estudante de Estatística Aplicada da Universidade de Brasília - em sua primeira tentativa completa de cursar a disciplina - aparenta ser explicado parcialmente pelas variáveis: turno no qual este cursa a disciplina principalmente ao se analisar os dados desconsiderando os SR's do estudantes - sendo que nesse caso os melhores desempenhos foram verificados nas turmas diurnas em relação às noturnas; tipo do professor - sendo que em ambos os casos considerados as taxas de aprovação aparentam ser significativamente menores quando a matéria é ofertada por professores substitutos; modalidade da discplina - nesse caso pôde-se notar que estudantes que cursam a matéria como módulo livre aparentam reprovar mais, enquanto os resultados entre alunos que cursam a matéria como optativa e obrigatória mostraram-se bastante semelhantes entre si; período no qual foi cursada a matéria - analisando principalmente os dados sem os SR's, verificou-se que atualmente as taxas de reprovação da disciplina são maiores que em momentos anteriores; e, por fim, a porcentagem de faltas dos estudantes na matéria - quanto maior esse percentual, maior aparenta ser a chance de reprovação do discente.

Com relação aos modelos construídos, os resultados obtidos através da modelagem desconsiderando-se os casos de SR's mostraram-se relativamente diferentes dos obtidos na modelagem geral dos dados, indicando que assumir que o SR na menção final do estudante indique um abandono deste com relação à disciplina possa produzir resultados mais razoáveis para analisar sua aprovação ou reprovação. Além disso, os resultados de ambos os modelos indicam que o tempo do estudante na Universidade aparenta ter uma influência negativa no resultado final deste na disciplina.

Conlusões iniciais induzem a novos questionamentos acerca dos motivos pelos quais as taxas de reprovação - na primeira tentativa dos estudantes em Estatística Aplicada - são maiores em alunos que fazem a matéria como módulo livre e com professores substitutos - sendo ainda interessante buscar entender os motivos pelos quais há tantos casos de SR (considerados nesse estudo como uma forma de evasão do aluno) - sendo interessante ainda notar a diferença de resultados entre os diferentes períodos considerados: será possível que tais diferenças possam ser explicadas por alguma mudança na Universidade? Ou seria outro o motivo principal desse fenômeno?

Finalmente, ao analisar os dados relativos a pesquisa sobre o perfil do estudante, pôde-se verificar que a desigualdade socio-econômica aparenta influenciar diretamente nos resultados acadêmicos dos estudantes na Universidade - especificamente, alunos de Estatística Aplicada. Sendo, então, importante analisar tais dados a partir de uma amostra representativa, a fim de testar se de fato as conclusões feitas são significativas.

Referências

AGRESTI, Alan. **An Introduction to Categorical Data Analysis** Third edition. University of Florida, Florida, United States: Wiley,2019

CUNHA, Simone Miguez; CARRILHO, Denise Madruga. O processo de adaptação ao ensino superior e o rendimento acadêmico. **Revista: Psicol. Esc. Educ.** (Impr.) vol.9 no.2 Campinas Dec. 2005. ⟨https://www.scielo.br/scielo.php?pid=S1413-85572005000200004&script=sci_arttext⟩ (acesso em 05/03/2021)

HOX, Joop J.; MOERBEEK, Mirjam; SCHOOT, Rens van de,. **Multilevel Analysis: Techniques and Applications** Third edition. New York, NY: Routledge,2017

KUTNER, Michael H.; NACHTSHEIM, Christopher J.; NETER, John; LI, William. **Applied Linear Statistical Models** Fifth edition. New York, NY: McGraw-Hill Irwin,2005

VENDRAMINI, Claudette Maria Medeiros; et al. Contrução e validação de uma escala sobre avaliação da vida acadêmica (EAVA). **Revista: Estud. psicol.** (Natal) vol.9 no.2 Natal May/Aug. 2004 ⟨https://www.scielo.br/scielo.php?pid=S1413-294X2004000200007&script=sci_arttext⟩ (acesso em 05/03/2021)

Apêndice

Cursos classificados como "Outros":

Agronomia	Artes Cênicas	Artes Visuais
Biotecnologia	Ciência da Computação	Ciências Biológicas
Ciências Econômicas	Ciências Farmacêuticas	Ciências Naturais
Computação	Comunicação	Comunicação Social
Desenho Industrial	Design	Direito
Educação Artística	Educação Física	Enfermagem
Enfermagem e Obstetrícia	Engenharia	Engenharia Ambiental
Engenharia Automotiva	Engenharia Civil	Engenharia de Computação
Engenharia de Energia	Engenharia de Produção	Engenharia de Redes de Comunicação
Engenharia de Software	Engenharia Elétrica	Engenharia Eletrônica
Engenharia Florestal	Engenharia Mecânica	Engenharia Mecatrônica
Estatística	Farmácia	Filosofia
Física	Fisioterapia	Geofísica
Gestão Ambiental	Gestão de Políticas Públicas	Gestão do Agronegócio
História	Informática	Jornalismo
Letras	Letras-Tradução	Línguas Estrangeiras Aplicadas - MSI
Matemática	Medicina	Medicina Veterinária
Museologia	Música	Nutrição
Odontologia	Pedagogia	Química
Química Tecnológica	Saúde Coletiva	Teoria
Crítica e História da Arte	Terapia Ocupacional e Turismo	