UNIVERSIDADE FEDERAL DO PARANÁ

Bruno Geronymo
Hermann Mogiz Delgado
Maria Helena de Oliveira
Vinicius César Pedroso

ESTUDO DO EFEITO DE VARIÁVEIS SÓCIOECONÔMICAS NO DESEMPENHO ESTUDANTIL

CURITIBA
24 de outubro de 2017

Resumo

A tarefa de coeducar jovens com diferentes realidades e perfis em um mesmo sistema educacional permeia o desafio de compreender por quê os desempenhos são distintos. Diante deste cenário, este estudo tem por finalidade avaliar a influência que as variáveis demográficas, sociais e relacionadas a escola têm sobre a aprovação de um estudante de ensino secundário considerando a disciplina de língua portuguesa. Perante a natureza binária da variável resposta, ajustou-se quatro modelos considerando as funções de ligação *Logito*, *Probito*, *Complemento Log-Log* e *Cauchy*. Ao considerar a análise de alguns apontadores de qualidade de modelos, foi verificado que a função de ligação *logito* proporcionou maiores benefícios entre os quatro modelos propostos. Após a verificação dos pressupostos o modelo foi validado com dados não utilizados para sua estimação e concluiu-se que os fatores são suficientemente satisfatórios para predizer a aprovação dos alunos deste estudo.

Palavras-chave: Desempenho estudantil, distribuição binomial, modelo de regressão generalizado.

Sumário

1 Introdução

A educação é um tema inquietante e presente de forma atuante na sociedade, principalmente no que diz respeito ao acesso ou a insuficiência de alguns ativos que condicionam um bom desempenho escolar dos jovens. Alunos que obtém notas altas, regulares ou insatisfatórias nas avaliações escolares têm entre si diferenças e semelhanças que ultrapassam o limite das salas de aula. Segundo uma pesquisa realizada por pesquisadores do Instituto Positivo, a educação num ambiente familiar estável, a convivência com atividades culturais, autocontrole no uso da internet são algumas situações que influenciam diretamente o desempenho escolar dos jovens enquanto que a falta dessas experiências e hábitos podem trazer resultados não satisfatórios para o aluno.

Neste contexto, o presente estudo transversal tem como objetivo avaliar a influência que as variáveis demográficas, sociais e relacionadas a escola têm sobre a aprovação de um estudante de ensino secundário na disciplina de língua portuguesa. Para tanto, utilizou-se a base de dados pública disponível na plataforma digital *Kaggle* contendo 649 instâncias.

2 Material e Métodos

2.1 Material

O estudo foi realizado em duas escolas portuguesas, tendo 33 atributos relativos as notas dos alunos na avaliação, características demográficas, sociais e relacionadas à escola. A coleta dos dados procedeu-se por meio da utilização de relatórios e questionários obtendo-se 649 observações. Abaixo são apresentados as variáveis que englobam a análise.

- **escola:** escola do aluno (GP - Gabriel Pereira; MS - Mousinho da Silveira)
- **sexo:** sexo do aluno (F - feminino; M - masculino)
- **idade:** idade do aluno (de 15 à 22 anos)
- **endereço:** tipo de endereço residencial do aluno (U - urbano ou R - rural)
- **famsize:** tamanho da família (LE3 - ≤ 3 ou GT3 - > 3)
- **Pstatus:** estado de coabitação dos pais (T - convivência; A - separados)
- **Medu:** educação da mãe (0 - nenhum, 1 - ensino primário (4º ano), 2 - 5º a 9º ano, 3 - ensino secundário ou 4 - ensino superior)
- **Fedu:** educação do pai (0 - nenhum, 1 - ensino primário (4º ano), 2 - 5º a 9º ano, 3 - ensino secundário ou 4 - ensino superior)
- **Mjob:** trabalho de mãe (1 - professora, 2 - área da saúde, 3 - serviços civis, 4 - em casa ou 5 - outro)
- **Fjob:** trabalho do pai (1 - professor, 2 - área da saúde, 3 - serviços civis, 4 - em casa ou 5 - outro)
- **razões:** motivo para escolher esta escola (1 - perto de casa, 2 - reputação da escola, 3 - preferência curso ou 4 - outra)

- **guardião:** responsável pelo aluno ('mãe', 'pai' ou 'outro')
- **horas de viagem:** tempo de viagem de casa à escola (1 - $<$ 15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora ou 4 - mais que 1 hora)
- **horário de estudo:** tempo de estudo semanal (1 - $<$ 2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 - $>$ 10 horas)
- **reprovação:** número de reprovações de classe passadas
- **supExt:** suporte educacional extra (S - sim; N - não)
- **famsup:** apoio escolar familiar (S - sim; N - não)
- **pay:** aulas extra pago no curso (matemática ou portugues) (S - sim; N - não)
- **atividades:** atividades extracurriculares (S - sim; N - não)
- **viveiro:** escola maternal atendida (S - sim; N - não)
- **enSup** quer fazer o ensino superior (S - sim; N - não)
- **internet:** - acesso à internet em casa (S - sim; N - não)
- **romântico:** com um relacionamento romântico (S - sim; N - não)
- **famrel:** qualidade das relações familiares (de 1 - muito ruim para 5 - excelente)
- **tempo livre:** tempo livre após a escola (de 1 - muito baixo para 5 - muito alto)
- **goout:** sair com amigos (de 1 - muito baixo a 5 - muito alto)
- **Dalc:** consumo de álcool no dia útil (de 1 - muito baixo a 5 - muito alto)
- **Walc:** consumo de álcool no fim de semana (de 1 - muito baixo para 5 - muito alto)
- **saúde:** estado de saúde atual (de 1 - muito ruim a 5 - muito bom)
- **ausências:** número de ausências escolares (de 0 a 93)

A base de dados apresenta três variáveis correspondentes as notas obtidas pelos 649 estudantes em três avaliações distintas na disciplina de língua portuguesa. Segundo informações apresentadas no site das escolas, a aprovação do estudante em uma determinada disciplina ocorre se a média aritmética das três avaliações for maior que 10 (escala de 0 a 20). Com isso, categorizou-se a nota final considerando o critério exposto, indicando a aprovação ou reprovação do educando conforme apresentado abaixo.

- **aprov:** média das três notas obtidas na disciplina de língua portuguesa indicando reprovação (0 - se a média for $<$ 10) ou aprovação (1 - se a média for $\geq$ 10)

O *software* R foi utilizado para a realização da análise do estudo. Os pacotes utilizados para isso foram: *car*, *MASS*, *statmod*, *Epi*.

2.2 Métodos

Dada a natureza binária da variável resposta *aprov* (0 e 1), ajustou-se quatro modelos considerando as funções de ligação *Logito*, *Probito*, *Complemento Log-Log* e *Cauchy*.

3 Resultados e Discussão

3.1 Modelagem

Em um primeiro momento foram ajustados quatro modelos lineares generalizados saturados, cada qual com uma função de ligação (*Logito*, *Probito*, *Complemento Log-Log* e *Cauchy*). A seleção das variáveis que constituiriam o modelo procedeu-se através do método de seleção de variáveis *stepwise*. A tabela 1 apresenta uma comparação entre os modelos considerando o AIC, os graus de liberdade residual (*df residual*), a estatística Q_{hl} e o seu respectivo *p-valor*.

Tabela 1: Comparação entre funções de ligação

	Logito	Probito	Clog-log	Cauchy
AIC	456.45	458.86	462.59	456.30
df residual	534	534	538	534
Q_{hl}	6.09	6.81	11.10	7.87
p-valor	0.6369	0.5571	0.1958	0.4459

Conforme as informações apresentadas na tabela 1, tem-se que o menor AIC é constatado no modelo utilizando a função de ligação *Cauchy*. Logo após, com um AIC próximo, o modelo *Logito*.

Optou-se em assumir o modelo logito pelo seu AIC ser muito próximo da *Cauchy*, bem como por a estatística Q_{hl} apresentar a maior probabilidade de bom ajuste do modelo e pela facilidade de interpretação dos indivíduos preditos. O modelo fica dado por:

$$g(\pi_i) = ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 2.064 - 1.805escola_i - 1.1484sexo_i - 0.147Mjobsaúde_{i31}$$

$$+0.1977Mjoboutro_{i32} + 1.2461Mjobserv_{i33} + 0.6803Mjobprof_{i34} - 0.9777guardianM_{i41}$$

$$-0.0523guardianOutro_{i42} - 1.4492reprovação_i - 1.4492supExt_i + 0.5507atividadesE_i$$

$$+2.05enSup_i - 0.204tempolivre_i - 0.085ausência_i$$

3.2 Análise de diagnóstico

A análise de diagnóstico permite verificar a adequação do modelo aos dados, bem como verificar se todos os pressupostos estão sendo atendidos e se há a presença de pontos influentes. Por meio dos gráicos abaixo, torna-se possível a verificação dos pressupostos.

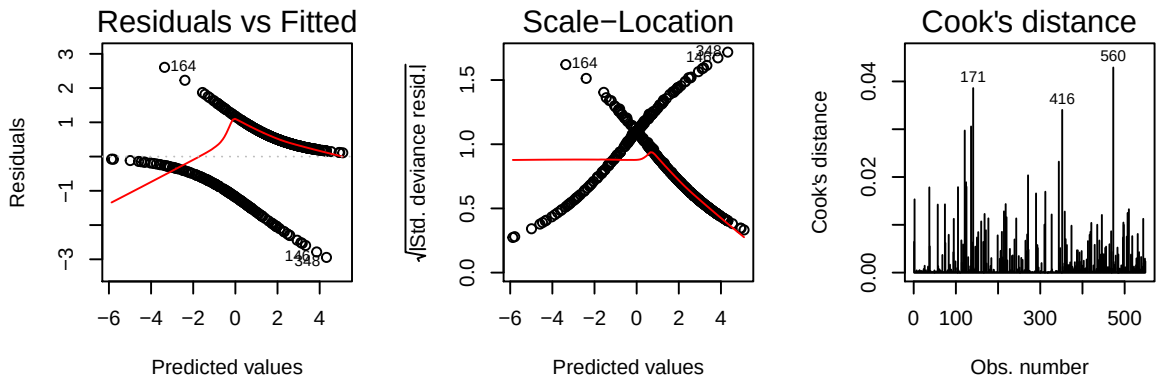

Figura 1: Análise de Diagnóstico do Modelo Proposto - Pressupostos

De acordo com a Figura 1, constata-se que não há fuga dos pressupostos. O padrão apresentado no primeiro gráfico (forma não linear) é decorrente do fato de a variáveç resposta ser dicotômica, ou seja, valores 0 ou 1. O segundo gráfico evidencia que não há indícios de tendência. Nota-se também uma acomodação aleatória dos pontos, não evidenciando uma relação média variância. No terceiro gráfico, tem-se pela análise da distância de Cook a não presença de valores discrepantes.

Outra forma de avaliar a qualidade do ajuste é com base nos resíduos quantílicos aleatorizados e por meio de um envelope simulado. O primeiro gráfico da figura 2 apresenta que os resíduos estão dispostos dispersos em torno de -3 e 3 e o segundo gráfico mostra que os resíduos estão bem dispersos no interior do envelope simulado, evidenciando que o modelo está bem ajustado, considerando a banda de 95% de confiança.

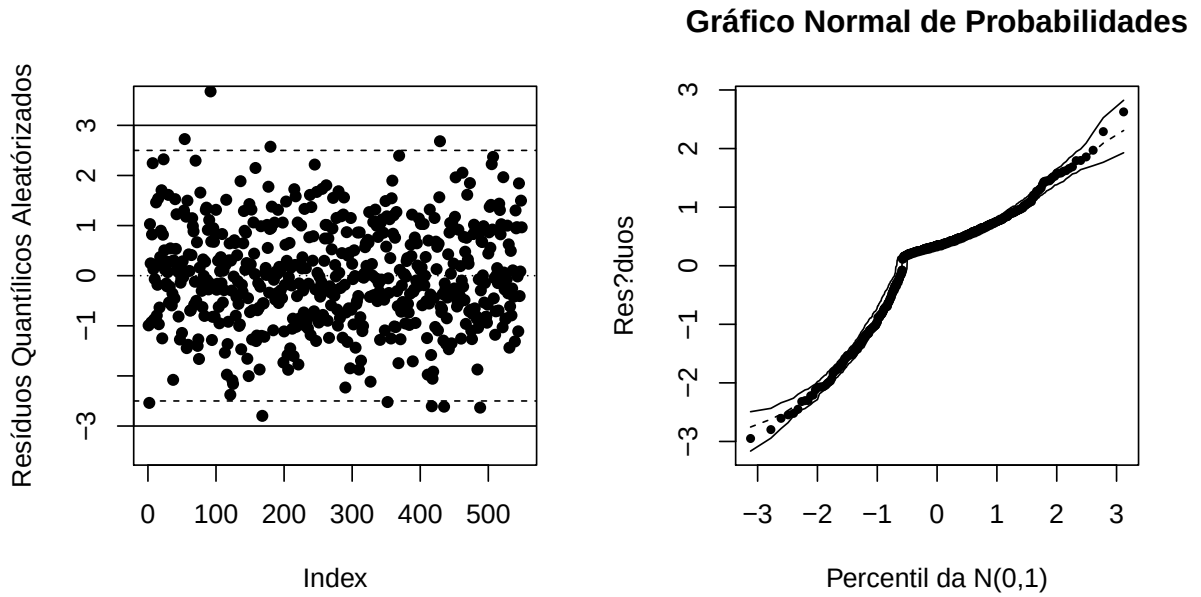

Figura 2: Gráfico dos resíduos quantílico aleatorizado e envelope simulado.

Para complementar a análise de diagnóstico, verificou-se a presença de valores discrepantes através dos gráficos de medidas de influência.

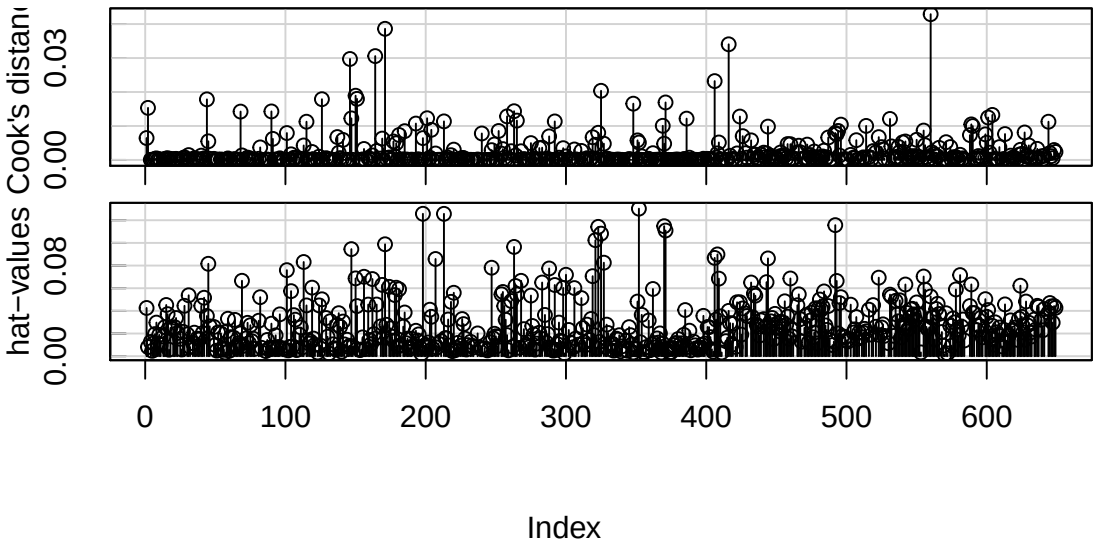

Figura 3: Medidas de Influência.

De acordo com a figura 3, conclui-se que não há observações que apresentam discrepância relativamente alto e, consequentemente, uma boa adequação do modelo.

3.3 Definição do ponto de corte

O uso do ponto de corte faz-se necessário para separar, a partir da probabilidade predita pelo modelo, os alunos que serão aprovados ou reprovados. Para definir um ponto de corte adequado, fez-se o uso da curva ROC, conforme apresetado na figura 4.

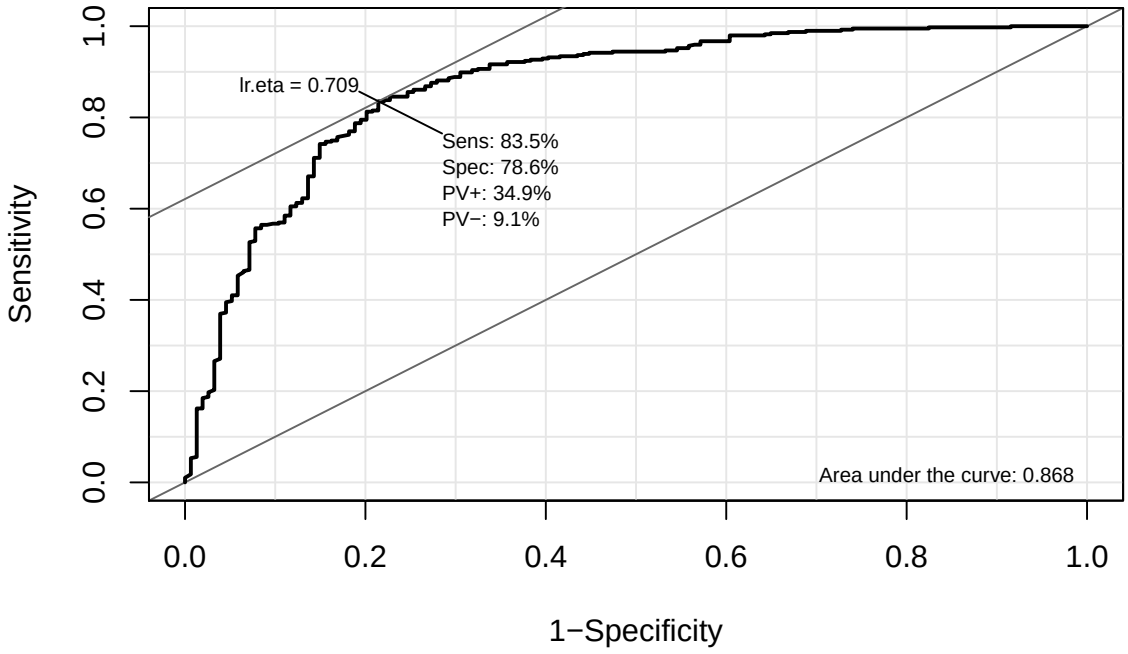

Figura 4: Curva ROC

Conforme as informações apresentadas na figura 4, pode-se definir que o ponto de corte é 0.709. Além disso, a sensibilidade é 0.835 e a especificidade é 0.786.

3.4 Validação do modelo

A validação faz-se necessária para avaliar o poder preditivo do modelo. O modelo ajustado foi estimado com a retirada de 15% das observações. Esses 15% serão utilizados nesta etapa para a validação do modelo. A tabela 2 abaixo ilustra a concordância do modelo com os dados observados.

Tabela 2: Tabelas de contingência

		Observado		
		Aprovado	Reprovado	Total
Predito	Aprovado	330	34	364
	Reprovado	65	120	185
	Total	395	154	549

		Observado		
		Aprovado	Reprovado	Total
Predito	Aprovado	54	8	62
	Reprovado	19	19	38
	Total	73	27	100

A tabela 2 auxilia no cálculo do intervalo de confiança para sensibilidade e especificidade. O intervalo obtido para espcificidade e sensibilidade, estimados pelo modelo, foi respectivamente (0.7135; 0.8449) e (0.7988; 0.8720). Já o intervalo de confiança obtido para especificidade e sensibilidade, considerando a predição para os dados de validação, foi respectivamente (0.5282; 0.8792) e (0.6384; 0.8411). Como os intervalos de confiança se cruzam, tem-se forte evidÊncia de que o modelo possui um poder preditivo satisfatório.

4 Considerações finais

O modelo nos permite identificar possíveis fatores que influenciam o desempenho dos alunos e alunas no curso de português. Pelo seu caráter transversal, o estudo não permite que sejam tiradas conclusões muito concretas, mas é possível utilizá-lo como evidência para estudos mais aprofundados. Os resultados indicam associações significativas entre 10 das variáveis coletadas e a probabilidade de aprovação do aluno.

Os fatores que influenciam no aumento da probabilidade de um indivíduo da população em estudo passar de ano são: mãe trabalha com serviços, ensino ou outros; possui ocupações extracurriculares; tem pretenções de cursar o ensino superior.

Entre as escolas consideradas, a que apresentou efeito negativo sobre a probabilidade de aprovação foi a Escola Mouzinho da Silveira. Os fatores sexo masculino, apoio da escola, e mãe que trabalha na área de saúde apresentaram efeito negativo em relação as outras categorias das mesmas variáveis. As categorias de guardião também apresentaram efeito negativo quando comparadas com os estudantes cujo guardião identificado foi o pai. O aumento do número de reprovações, do tempo livre e do número de faltas relatados também contribuiram na diminuição da probabilidade de aprovação dos estudantes.

A análise do estudo também sugere associação positiva entre a probabilidade de aprovação e a intenção do aluno em cursar ensino superior.