UNIVERSIDADE FEDERAL DO PARANÁ

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

ANÁLISE DE MODELOS DE CONTAGEM PARA A COMPREENSÃO DE REPROVAÇÕES ESTUDANTIS

CURITIBA
24 de novembro de 2017

Resumo

Diversas são as características que influenciam no desempenho escolar dos estudantes. Compreendê-las permite que a implementação de medidas sejam feitas para favorecer o aprendizado dos alunos. Diante deste cenário, este estudo tem por finalidade avaliar a influência que as variáveis demográficas, sociais e relacionadas a escola têm sobre o número de reprovações de um estudante de ensino secundário considerando a disciplina de matemática. Ajustou-se quatro modelos considerando a Poisson e Binomial Negativa com e sem inflação de zeros. A análise favoreceu o modelo de Poisson sem inflação de zeros, constatando a relação de vários aspectos da vida do estudante com o seu histórico de dependências escolares.

Palavras-chave: Dados de contagem, desempenho estudantil, número de reprovações.

Sumário

1 Introdução

O desempenho escolar de crianças e adolescentes é um fenômeno que depende de diversos fatores. Um melhor ou pior desempenho pode estar associado ao contexto social, econômico, cultural e até familiar. Um dos grandes desafios da modernidade é compreender como adaptar o sistema de ensino para atender as necessidades dos mais diversos tipos de estudantes.

Para entender melhor como o contexto vivenciado pelo aluno afeta o seu desempenho escolar, o presente estudo tem por objetivo compreender a influência que as variáveis demográficas, sociais e relacionadas a escola têm sobre o número de reprovações dos estudantes em uma certa disciplina, utilizando uma base de dados disponível na plataforma digital *Kaggle*, coletada em dois colégios de ensino médio da cidade do Porto, em Portugal, no ano de 2008. Foram considerados para a coleta de dados estudantes cursando a disciplina de matemática, e portanto o número de reprovações modelado é referente a esta disciplina.

2 Material e Métodos

2.1 Material

A base de dados contém 395 observações e 33 variáveis relativas as notas dos alunos na avaliação, características demográficas, sociais e relacionadas à escola, sendo elas:

- **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)
- **G1:** Nota obtida pelo aluno no primeiro perído (0 a 20)
- **G2:** Nota obtida pelo aluno no segundo período (0 a 20)
- **G3:** Nota final obtida pelo aluno (0 a 20)

A coleta dos dados deu-se através da utilização de questionários e relatórios. Para a análise, utilizou-se o *software* R e os respectivos pacotes *gamlss*, *hnp*, *statmod*, *MASS* e *lme4*, além de algoritmos como *stepwise* para seleção de variáveis do modelo.

2.2 Métodos

Ao considerar o número de reprovações como sendo a variável resposta, ajustou-se modelos de regressão para dados de contagem, utilizando-se de métodos computacionais para a análise e diagnóstico destes ajustes.

3 Resultados e discussão

3.1 Análise descritiva e exploratória

A análise descritiva e exploratória permite observar o comportamento e peculiaridades que envolvem as variáveis do estudo. A figura 1 apresenta o número de reprovações, onde é possível constatar um número elevado de zeros indicando um alto índice de alunos que nunca reprovaram.

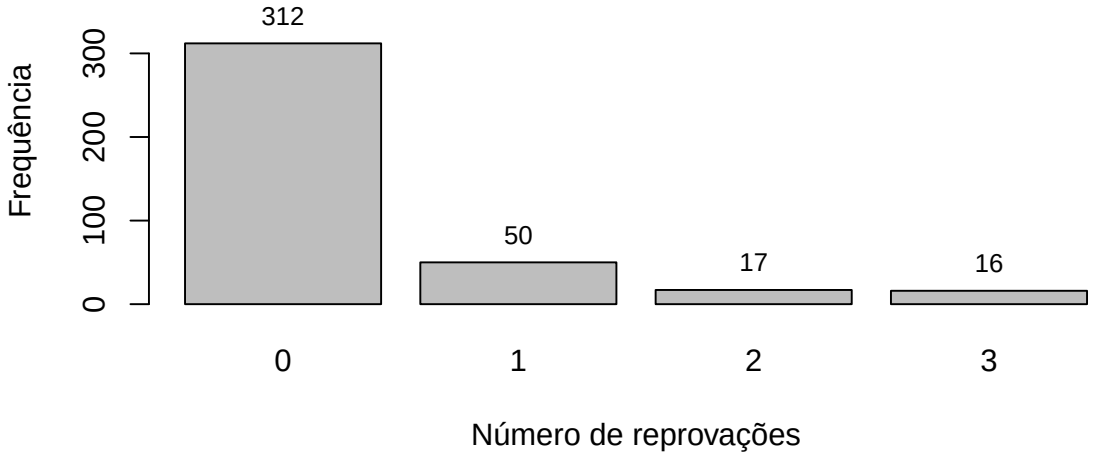

Figura 1: Frequência do número de reprovações

Diante das informações apresentadas na figura 1, fez-se necessário o ajuste de modelos considerando o número elevado de zeros.

3.2 Modelagem

Ajustou-se em um primeiro momento quatro modelos lineares generalizados considerando todas as variáveis, sendo a *Poisson* e a *Binomial Negativa* com e sem inflação de zeros. A tabela 1 apresenta uma comparação entre os modelos considerando o AIC, os graus de liberdade residual (*df residual*) e a deviance residual.

Tabela 1: Comparação entre modelos

	Poisson	Binomial Negativa	Zero Inflated Poisson	Zero Inflated Negative Binomial
AIC	500.3165	502.3190	502.0533	504.0534
Deviance	224.0140	223.9661	416.0533	416.0534
df residual	353	353	352	351

A tabela 1 revela circuntâncias favoráveis aos modelos Poisson e Binomial Negativa, já que eles obtiveram os menores valores de AIC e deviance residual. Contudo, a simplicidade do modelo de Poisson em conjunto ao menor AIC nos faz aderir ao mesmo. O mesmo padrão destes modelos foi observado pelos modelos com variáveis selecionadas pelo método de *stepwise*.

Após a seleção de um modelo poisson foram testados diversos ajustes, dentre eles verificou-se que o mais adequado apresentava termos quadráticos e algumas interações significativas. O AIC para este modelo foi de 455.76 com deviance residual de 225.46 e 376 graus de liberdade. A tabela 2 apresenta os valores estimados dos parâmetros, os respectivos erros padrão e p-valor associado (em percentual).

Tabela 2: Resumo das Estimativas para o Modelo Ajustado

	Estimativa	Erro padrão	P-valor
(intercepto)	2.792025	0.709406	0.00829%
EndereçoUrbano	-0.356563	0.213707	9.5223%
Fedu	-0.356563	0.138060	0.00810%
MjobÁreaDaSaúde	0.190719	0.428190	65.6026%
MjobOutro	-0.398876	0.256979	12.0620%
MjobServiçosCivis	0.539502	0.256110	3.5158%
MjobProfessor	-1.141169	0.640659	7.4873%
GuardiãoMãe	-0.440386	0.250914	7.9238%
GuardiãoOutro	0.919126	0.282140	0.1123%
PaySim	-0.711257	0.219491	0.1123%
FamRel	-0.141753	0.102723	16.7603%
TempoLivre	0.189524	0.098676	5.4774%
Walc	0.138343	0.069107	4.5299%
Ausências	-0.047829	0.055015	31.7673%
G1	-0.221147	0.055015	0.00582%
G3	-0.109612	0.029774	0.0232%
Ausênciasˆ2	-0.001728	0.000873	4.7732%
Fedu:Ausências	0.026583	0.012463	3.2926%
Ausências:G3	0.009314	0.004374	3.3225%

3.3 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. A figura 2 permite avaliar a qualidade do ajuste do modelo com base nos resíduos quantílicos aleatorizados e no envelope simulado.

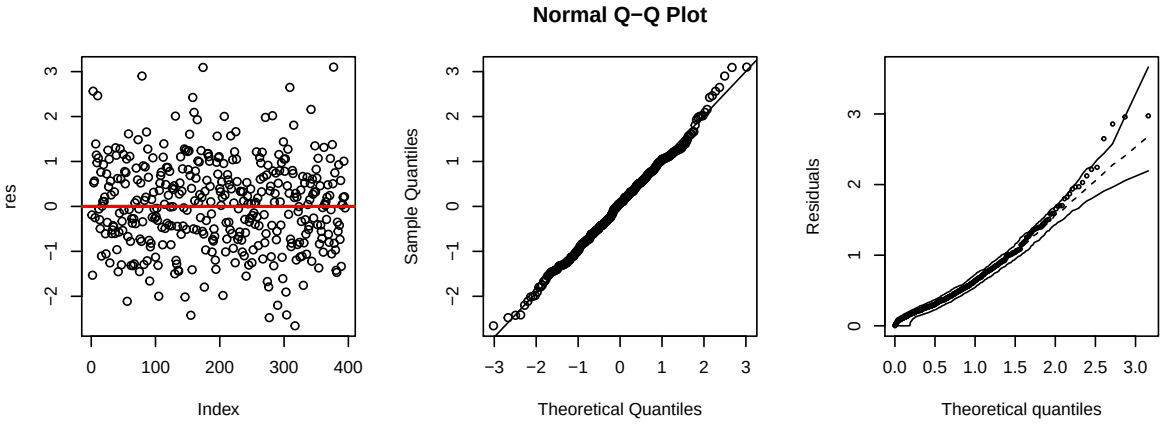

Figura 2: Gráfico dos resíduos quantílico aleatorizado e envelope simulado

Através da figura 2, nota-se no primeiro gráfico que os resíduos quantílicos aleatorizados estão bem dispersos entre -3 e 3 e o segundo gráfico evidencia um padrão de normalidade dos mesmos. Quanto ao gráfico do envelope simulado, constata-se que os resíduos estão bem dispersos no interior do envelope, exceto a presença de dois pontos que excederam as bandas de 95% de confiança, nada que impedisse o uso deste modelo. Uma análise de medidas de influência evidenciou não haver indicativos de outliers ou observações influentes.

De acordo com o modelo obtido, fatores relacionados ao contexto familiar do estudante foram altamente relevantes para o desempenho dos alunos. Ajustadas pelo efeito das demais covariáveis, permaneceram significativos: o nível educacional do pai, o tipo de emprego da mãe, o responsável pelo estudante (mãe, pai ou outro) e a qualidade das relações familiares. Outras variáveis socioeconômicas também influenciaram no resultado obtido: a moradia do estudante (urbana ou rural), a realização de aulas particulares pagas e o tempo livre do aluno. O nível do consumo de álcool pelo estudante também apresentou efeito significativo, podendo aumentar em até 74% a contagem média de reprovações, para alunos cujo consumo de álcool em finais de semana é "muito alto", em relação aos cujo consumo é "muito baixo". Além disso, o número de faltas e as notas obtidas pelos alunos também foram relevantes na estimação dessa contagem.

Foi possível observar que a área urbana, em relação a área rural foi um dos fatores que diminuiu a contagem média de reprovações. O consumo de álcool e o tempo livre, por outro lado, foram considerados fatores que podem aumentar a média de reprovações dos alunos.

4 Considerações finais

Na análise deste fenômeno, foi importante a consideração de diversos tipos de variáveis, que considerassem vários aspectos da vida do estudante. Observa-se que o desempenho escolar está relacionado tanto com as condições sociais e econômicas do aluno, quanto com as características da sua vida familiar e pessoal. As relações evidenciadas no estudo podem ser úteis para a reflexão e melhoria do sistema educacional, para que estudantes de diferentes origens possam usufruir do ensino da mesma maneira. Além disso, podemos

observar a necessidade de ações de conscientização quanto ao uso de álcool na adolescência e a importância das relações familiares para o desempenho educacional.

Quanto à análise estatística, observamos que neste caso o melhor modelo foi também o mais usual para este tipo de dados, apesar da grande quantidade de valores nulos da variável resposta, estes foram comportados de maneira satisfatória ao considerar a distribuição Poisson.