Modelagem da evasão no ensino superior no Brasil

27/06/2023

Juan Belieni
juan.araujo@fgv.edu.br
FGV/EMAp

Sumário

1 Introdução

A evasão no ensino superior é um problema que afeta muitos cursos acadêmicos pelo Brasil, e possui diversas naturezas ao nível do estudante, tais como vocacionais, relativos ao desempenho ou até sociais [1]. No entanto, fatores macros como a modalidade de ensino do curso e o tipo de administração da instituição de ensino também podem ajudar a entender as dinâmicas do processo de evasão.

Modelar esse fenômeno dá a possibilidade de compreender como esse fatores se relacionam e influenciam a quantidade de desistências, permitindo identificar situações anômalas ou até tendências indesejáveis, que possibilitaria o desenvolvimento de planos de ação a nível nacional para a criação e aprimoramento de programas de combate a evasão em instituições públicas e privadas.

Dessa forma, a discussão e a modelagem iniciada nesse trabalho busca explorar os dados oferecidos pelo Inep para entender o comportamento da variável de interesse (quantidade de desistências de um determinado curso) por meio modelos de regressão para contagem, utilizando as variáveis disponíveis para estimar seu valor e depois analisando os coeficientes das covariáveis para entender seu comportamento.

2 Dados

Os dados desse trabalho foram coletados pelo Censo da Educação Superior, realizado anualmente pelo Inep [2], e disponibilizados no formato CSV no repositório do trabalho[1]. Cada entrada do conjunto de dados possui informações ao nível de curso por instituição em um determinado ano de referência. A cada ano, a partir do ano de ingresso, é registrado a quantidade de alunos que concluíram e desistiram

[1] https://github.com/juanbelieni/fgv-me-a2

do curso, além do número de falecidos nesse determinado ano. Para esse trabalho, o ano de ingresso escolhido para a modelagem foi 2012.

Cada curso no conjunto conta com os seguintes dados:
- identificação (código e nome);
- local onde o curso é ofertado (região, UF e município);
- grau acadêmico conferido ao diplomado (bacharelado, licenciatura ou tecnólogo);
- modalidade de ensino (presencial ou a distância);
- classificação segundo a Cine Brasil.

Em relação à instituição onde o curso é ofertado, os seguintes dados são disponibilizados:
- identificação (código e nome);
- categoria administrativa:
  1. pública federal;
  2. pública estadual;
  3. pública municipal;
  4. privada com fins lucrativos;
  5. privada sem fins lucrativos;
  6. especial.
- organização acadêmica:
  1. universidade;
  2. centro universitário;
  3. faculdade;
  4. instituto federal de educação, ciência e tecnologia;
  5. centro federal de educação tecnológica.

Na Seção 3.2 será visto quais informações serão utilizadas para ajustar um modelo que atinga os propósitos desse trabalho.

3 Modelagem

O foco dessa modelagem é entender a influência de algumas variáveis na quantidade de desistências dos cursos presentes no conjunto de dados. Mais especificamente, será modelado o número acumulado de desistências até o ano de integralização para um determinado curso i, que será denominado de D_i.

A escolha do ano de integralização como limite superior para o cálculo do número acumulado de desistências vem da necessidade de estipular uma base de comparação geral entre os diversos cursos, que possuem prazos de integralização diferentes.

3.1 Modelos para contagem

3.1.1 Regressão de Poisson e suas limitações

Para modelar um processo de contagem, natureza da variável de interesse desse trabalho, é conveniente utilizar um modelo linear generalizado (GLM) com a distribuição de Poisson [3], onde a função de ligação é do tipo $g(\mu) = \ln(\mu)$. Dessa forma, um modelo desse tipo para a variável de interesse desse trabalho pode ser descrito da seguinte forma:

$$E[D_i \mid \boldsymbol{X}_i] = \mu_i = \exp(\beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}'), \tag{1}$$

onde $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}')$ é o vetor de parâmetros a ser estimado.

Essa técnica é conhecida como regressão de Poisson e a estimação dos parâmetros ocorre por meio de máxima verossimilhança. Por não possuir fórmula fechada, a estimação depende métodos numéricos (no *R*, é utilizado *Fisher's scoring*).

No entanto, diferente de outras distribuições como a Normal e a Binomial Negativa, a distribuição de Poisson não possui um parâmetro de dispersão. Por esse motivo, em uma regressão de Poisson, assume-se que os dados são equidispersos, i.e., que a média condicional seja igual à variância condicional [4]. Caso isso não seja verdade e esse fato não for levado em conta, podemos acabar tendo valores incorretos para as estimativas de erro padrão, para os intervalos de confiança, etc.

É possível verificar um caso de sobredispersão ou subdispersão em um modelo já treinado por meio de um teste proposto por Cameron e Trivedi [5] de seguinte teor:

$$\begin{aligned} H_0 &: \mathrm{Var}(y_i) = \mu_i, \\ H_1 &: \mathrm{Var}(y_i) = \mu_i + \alpha \cdot g(\mu_i), \end{aligned} \tag{2}$$

onde $g(\cdot)$ é uma função positiva qualquer. No *R*, é possível realizar esse teste por meio do método `dispersiontest` disponibilizado na biblioteca *AER* [6].

Caso haja evidência para os casos citados, i.e., que os dados não sejam equidispersos, uma possível alternativa para contornar essa limitação do modelo é incluir um parâmetro de dispersão ϕ. Com essa correção, o novo modelo, chamado de regressão Quasi-Poisson [7], possui a seguinte descrição:

$$\begin{aligned} E[D_i \mid \boldsymbol{X}_i] &= \mu_i, \\ \mathrm{Var}(D_i \mid \boldsymbol{X}_i) &= \phi \cdot \mu_i. \end{aligned} \tag{3}$$

Outra alternativa é construir uma regressão com a distribuição Binomial Negativa, que também possui um parâmetro de dispersão. Por possuir uma parametrização relativamente parecida com a regressão de Poisson, como será visto posteriormente, e por não depender de métodos de quasi-verossimilhança, foi escolhida sua utilização.

3.1.2 Regressão Binomial Negativa

A regressão Binomial Negativa começa com uma modelagem similar para média de D_i, mas introduz um novo parâmetro κ que descreve sua variância [7]. Para a modelagem que está sendo feito nesse trabalho, o modelo em questão seria descrito da seguinte forma:

$$\begin{aligned} E[D_i \mid \boldsymbol{X}_i] &= \mu_i = \exp(\beta_0 + \boldsymbol{X}_i^T \boldsymbol{\beta}'), \\ \mathrm{Var}(D_i \mid \boldsymbol{X}_i) &= \mu_i + \kappa \mu_i^2. \end{aligned} \tag{4}$$

Essa definição dá à regressão Binomial Negativa uma maior flexibilidade em modelar o comportamento envolvendo a variável de interesse e as covariáveis, se comparado com uma regressão de Poisson tradicional [3], sendo ainda possível utilizar máximo verossimilhança para estimar os parâmetros necessários.

No *R*, a biblioteca *MASS* oferece uma implementação que permite ajustar modelos desse tipo [8], que pode ser feito utilizando o método `glm.nb`, uma modificação do método `glm` que estima um parâmetro $\theta = \frac{1}{\kappa}$ por máxima verossimilhança, utilizado posteriormente para ajustar os coeficientes e calcular outros valores associados ao modelo.

3.2 Escolha das covariáveis

Já explorado na Seção 2, a base de dados possui informações relativas ao curso e à instituição onde este é ofertado. No entanto, nem toda informação é útil, e a escolha de qual incluir para que o ajuste do modelo seja construído corretamente passou por uma investigação. A princípio, é esperado uma

correlação substantiva entre a variável de interesse e a quantidade de ingressantes, como é possível ver na Figura 1.

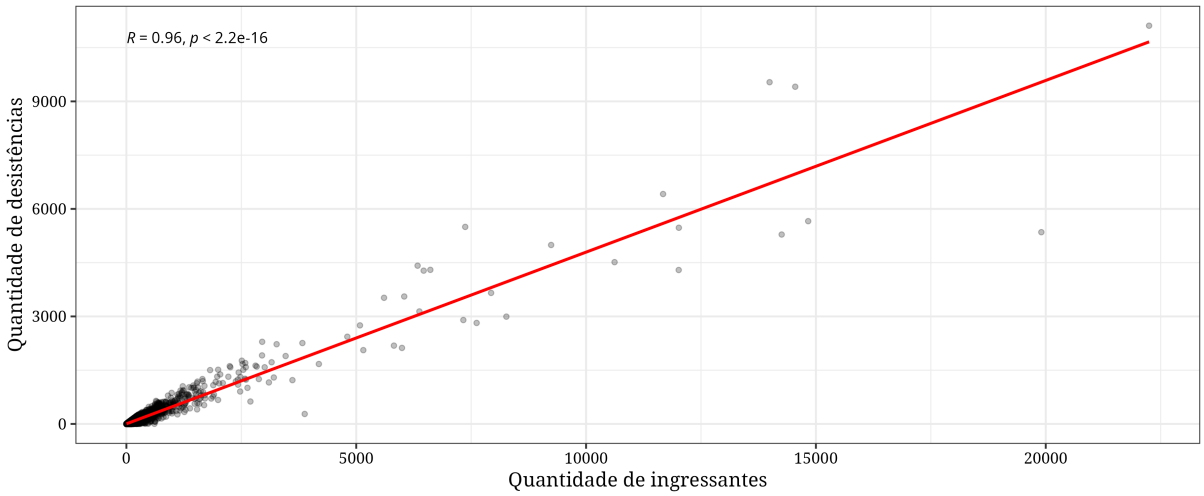

Figura 1: Gráfico de dispersão entre a quantidade de ingressantes e a quantidade de desistências, acompanhado por uma linha de regressão linear.

Porém, é importante notar que, diferente do que acontece em uma regressão linear tradicional, uma mudança unitária do valor de uma covariável qualquer β_j não resulta em uma mudança aditiva proporcional a β_j, e sim acarreta uma mudança multiplicativa de fator e^{β_j} [4]. Dessa forma, devido à forte relação linear entre as variáveis, é interessante modelar a covariável relativa à quantidade de ingressantes como $\log(\cdot)$.

Outra variável que deveria apresentar uma relevância considerável é o prazo de integralização. No entanto, não existe uma forma funcional óbvia entre essa quantidade e a quantidade de desistências, como é possível ver na figura Figura 2. Além disso, apenas os cursos com prazo de integralização entre 3 e 6 anos apresentam uma quantidade considerável de dados.

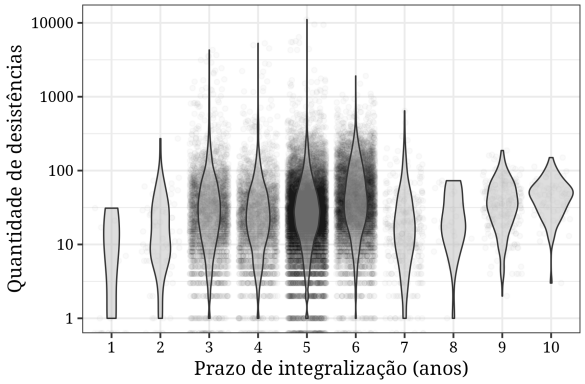

Figura 2: Gráfico da distribuição da quantidade de desistências em relação ao prazo de integralização.

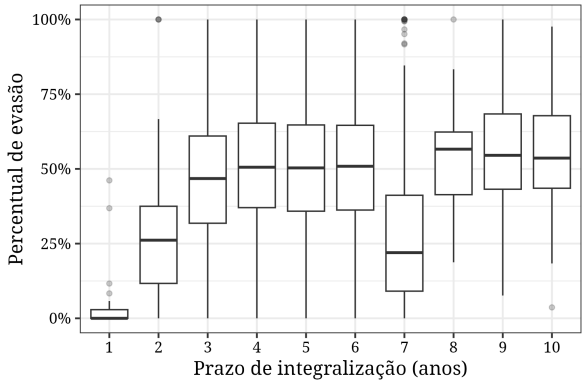

Figura 3: Gráfico da distribuição do percentual de evasão em relação ao prazo de integralização.

Se formos visualizar essa variável em relação ao percentual de evasão (Figura 3), fica ainda menos óbvio qual seria a forma ideal de representar esse valor de maneira ordinária. Portanto, é perceptível que modelar essa variável como um valor não-categórico não seria o ideal.

Em relação às variáveis categóricas, algumas destas acabam tendo naturalmente muitas opções de valores possíveis, como a variável contendo a UF onde o curso é ofertado. Por esse motivo, essas variáveis foram preteridas em favor de outras mais gerais, como a da região, no caso citado. Para a última, nota-se uma ligeira variação do percentual de evasão, como é possível observar na Figura 4.

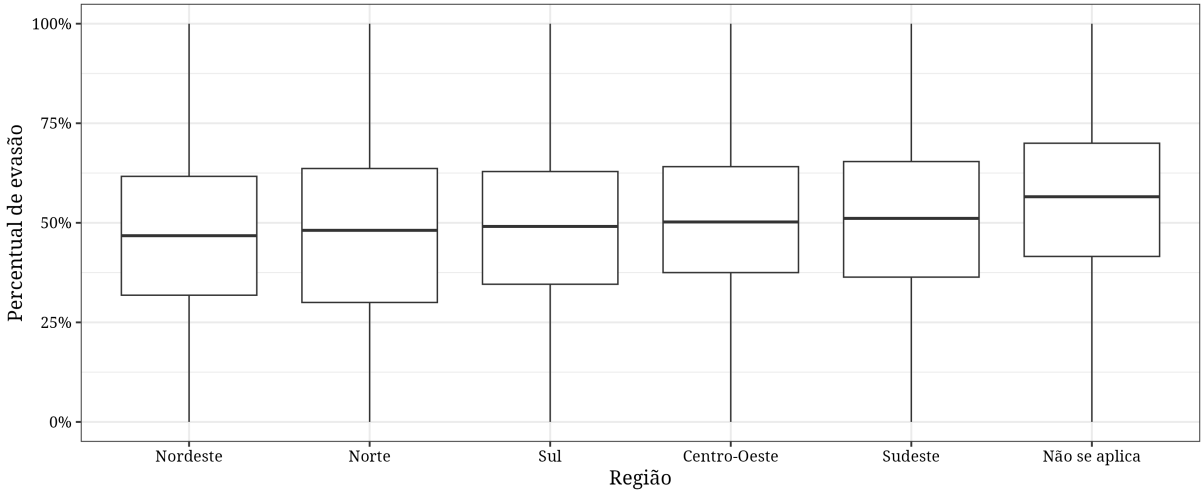

Figura 4: Gráfico da proporção da taxa de evasão em relação à região.

A diferença que se destaca, nesse caso, é uma maior média do percentual de evasão quando a classificação de região não se aplica. No conjunto de dados, essa classificação significa que o curso é ofertado a distância. Dessa maneira, conseguímos utilizar a variável que informa a modalidade de ensino para codificar essa discrepância.

Também é possível observar variações significativas no percentual de evasão quando analisamos esse valor em relação ao tipo de administração na Figura 5.

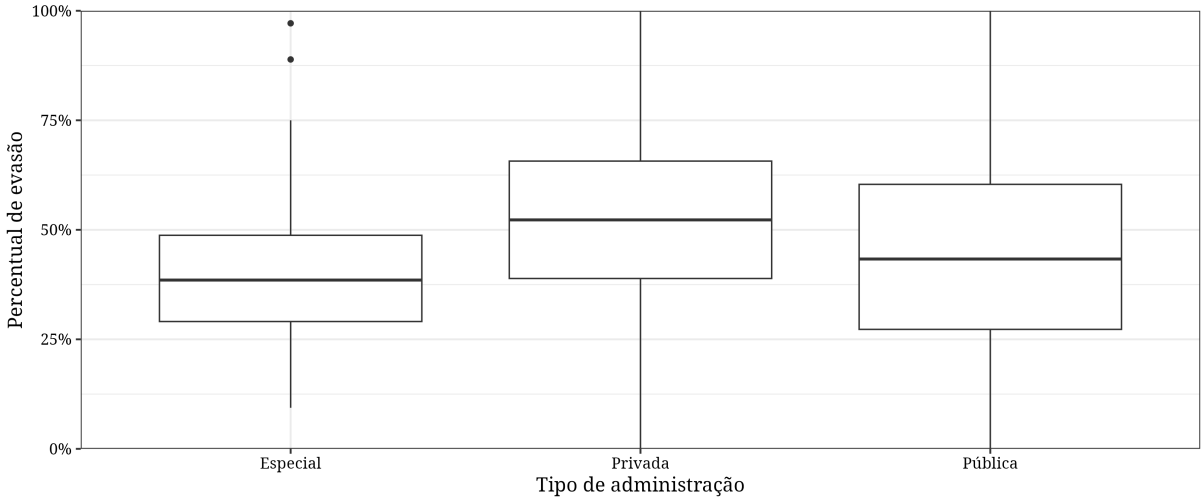

Figura 5: Gráfico da proporção da taxa de evasão em relação ao tipo de administração.

Com essa investigação, as variáveis escolhidas serão transformadas para formar as covariáveis do modelo final, que será descrito a seguir e contará com 14 covariáveis.

3.3 Modelo final

Por fim, o modelo escolhido foi uma regressão Binomial Negativa que leva em conta as influências das variáveis da quantidade de ingressantes, do prazo de integralização, da modalidade de ensino e do tipo de administração da universidade. No *R*, tal modelo é construído da seguinte maneira:

```
glm.nb(
  qt_desistencias ~ 1
    + modalidade_ensino
    + tipo_administracao
    + as.factor(prazo_integralizacao)
    + log(qt_ingressantes),
  data = data,
)
```

Antes de mostrar seus resultados, um modelo de regressão Poisson análogo ao apresentado para a regressão Binomial Negativa também será mostrado, no qual será aplicado o teste de dispersão apresentado na Seção 3.1.1 para verificar a necessidade o uso de um modelo que considere sobredisperão.

4 Resultados

O modelo de regressão de Poisson foi ajustado em *R* utilizando a biblioteca padrão com o seguinte comando:

```
glm(
  qt_desistencias ~ 1
    + modalidade_ensino
    + tipo_administracao
    + as.factor(prazo_integralizacao)
    + log(qt_ingressantes),
  data = data,
  family = poisson,
)
```

Por padrão, o *R* considera que o parâmetro de dispersão é igual a 1 [9]. Com esse valor e após 5 iterações de *Fisher's scoring*, foi produzido um modelo com AIC igual a 339.177, BIC igual a 339.292 e com as estimativas para os coeficientes que estão presente na Tabela 1.

Covariável	**Estimativa**	**Erro padrão**	***z-value***
Intercepto	-3,0975812	0,1103542	-28,069
Modalidade de ensino (presencial)	-0,0946402	0,0033707	-28,077
Tipo de administração (Privada)	0,3438696	0,0152867	22,495
Tipo de administração (Pública)	0,1222539	0,0153891	7,944
Prazo de integralização (2 anos)	1,7122496	0,1120452	15,282
Prazo de integralização (3 anos)	2,1894498	0,1091304	20,063
Prazo de integralização (4 anos)	2,2939549	0,1091393	21,019
Prazo de integralização (5 anos)	2,3490929	0,1091144	21,529
Prazo de integralização (6 anos)	2,3586600	0,1091174	21,616
Prazo de integralização (7 anos)	1,8257122	0,1095745	16,662

Prazo de integralização (8 anos)	2,3518163	0,1153984	20,380
Prazo de integralização (9 anos)	2,4495912	0,1095914	22,352
Prazo de integralização (10 anos)	2,4405378	0,1103832	22,110
Quantidade de ingressantes (log)	0,9768587	0,0008204	1190,769

Tabela 1: Estimativa dos coeficientes do modelo de regressão de Poisson.

Realizando o teste de dispersão apresentado na Seção 3.1.1 com o modelo acima, temos que seu p-valor é menor que 2,2e16, com o valor estimado de 6,948751 para ϕ. Ou seja, temos bastante evidência para rejeitar a hipótese nula. Com isso, o modelo final especificado na Seção 3.3 pode ser finalmente ajustado.

O método `glm.nb` começou estimando o valor de 5,9732 para o parâmetro θ. Depois disso, o modelo ajustado, com AIC igual a 215.558 e BIC igual a 215.681, apresentou como estimativa dos coeficientes os valores presentes na Tabela 2.

Covariável	**Estimativa**	**Erro padrão**	***z-value***
Intercepto	-3,183171	0,172868	-18,414
Modalidade de ensino (presencial)	-0,151741	0,016649	-9,114
Tipo de administração (Privada)	0,300960	0,037509	8,024
Tipo de administração (Pública)	0,110856	0,037685	2,942
Prazo de integralização (2 anos)	1,939972	0,177700	10,917
Prazo de integralização (3 anos)	2,364277	0,167391	14,124
Prazo de integralização (4 anos)	2,517725	0,167434	15,037
Prazo de integralização (5 anos)	2,511231	0,167270	15,013
Prazo de integralização (6 anos)	2,507559	0,167309	14,988
Prazo de integralização (7 anos)	1,949313	0,169240	11,518
Prazo de integralização (8 anos)	2,540016	0,192112	13,222
Prazo de integralização (9 anos)	2,635591	0,169901	15,513
Prazo de integralização (10 anos)	2,658740	0,174851	15,206
Quantidade de ingressantes (log)	0,976705	0,003276	298,181

Tabela 2: Estimativa dos coeficientes do modelo de regressão Binomial Negativa.

É possível perceber que os dois métodos apresentaram valores muito próximos para os coeficientes, o que é o esperado dado que os dois métodos apresentam a mesma expressão para a média. Outro comportamento esperado é o erro padrão menor para os coeficientes, pois, como visto na Seção 3.1.1, modelos de regressão de Poisson com dados sobredispersos acabam atribuindo valores menos corretos para essa informação se comparados com modelos que consideram esse fenômeno.

Também é interessante analisar o modelo por meio da utilização de um gráfico de quantis dos resíduos. Para esse diagnóstico, o cálculo dos resíduos foi feito utilizado o *randomized quantile residual* (RQR), proposto por Dunn e Smyth em 1996 [10], que serve para trabalhar com modelos de contagem [11],

com a visualização disponível na Figura 6, no qual os resíduos foram calculados utilizando o método `qresiduals` da biblioteca *countreg* [12].

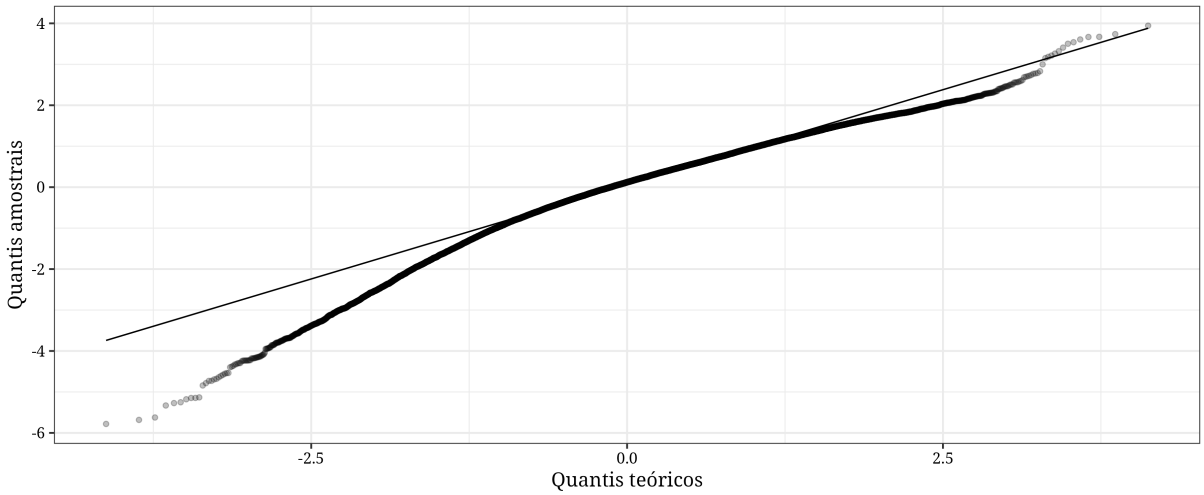

Figura 6: Gráfico de quantis dos resíduos do tipo RQR.

É notável que o modelo não se ajustou bem aos dados, pois existe um desvio considerável dos pontos em relação da linha de identidade. Isso pode ter acontecido por diversos fatores, sendo um deles a hipótese dos dados não seguirem realmente uma distribuição Binomial Negativa.

Também podemos avaliar a hipótese de normalidade dos resíduos por meio de um teste estatístico. Utilizando o teste de Anderson–Darling [13] para essa finalidade, calculamos seu p-valor em *R* utilizando o método `ad.test` da biblioteca *nortest* [14], que considera que esse valor é menor que 2,2e-16. Ou seja, temos bastante evidência para rejeitar a normalidade dos resíduos.

É possível visualizar esse comportamento dos resíduos ao analisar sua densidade (Figura 7). Mesmo tendo uma distribuição aparentemente gaussiana, a média tende para a direita e a cauda esquerda é mais pesada.

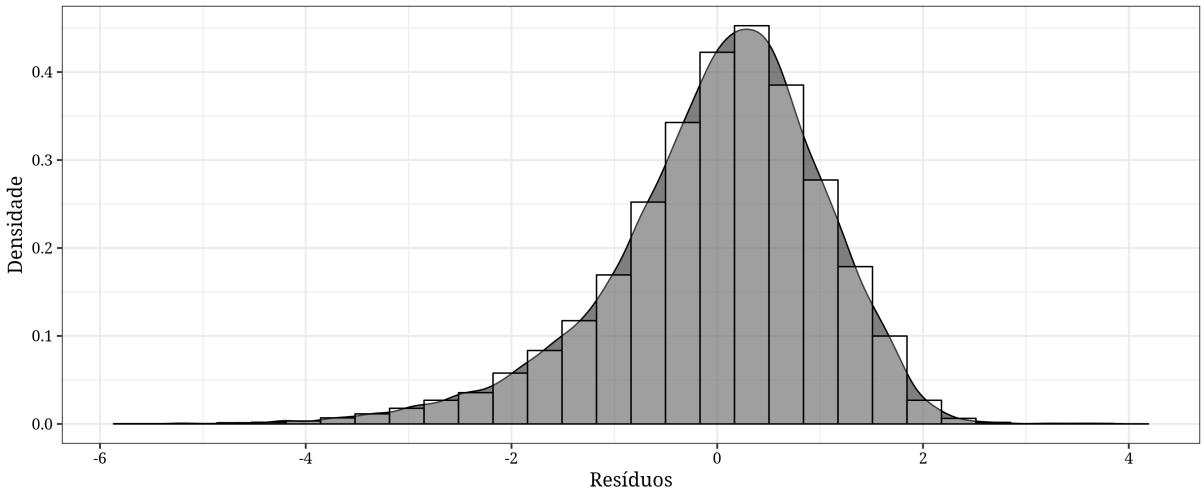

Figura 7: Gráfico de densidade dos resíduos do tipo RQR.

5 Conclusão

5.1 Interpretação dos resultados

Mesmo com o ajuste não ideal do modelo, ainda é pertinente interpretar seus coeficientes e resultados. Primeiramente, a hipótese inicial da forte relação entre a quantidade de ingressantes e desistências é percebido claramente na covariável "quantidade de ingressantes (log)", pois a estimativa para seu coeficiente foi a que mais teve evidência de ser diferente de zero.

Em relação as variáveis categóricas, podemos começar percebendo que cursos com a modalidade de ensino presencial possuem uma taxa menor de desistências. Esse resultado pode ser compreendido ao perceber que cursos a distância possuem características únicas que tornam essa modalidade mais suscetíveis à evasão, como a falta de uma infraestrutura física robusta para o aprendizado, dificuldades inerentes ao meio digital como a falta de *feedbacks* e apoio ao aluno, a demografia dos estudantes e outros fatores que prejudicam o processo de aprendizado [15].

Analisando os coeficientes para instituições com administrações públicas e privadas, é observado que cursos de instituições privadas são aqueles com maior taxa de evasão. Muitos podem ser os motivos para esse fenômeno, porém diferenças em como cada tipo de administração lida com avaliações institucionais e realiza processos de estudo do motivo de saída dos alunos podem ser uma pista [16].

Os prazos de integralização, por sua vez, retornaram resultados coerentes para os coeficientes em relação à análise exploratória dos dados. No entanto, assim como aconteceu na análise exploratória, cursos com prazo de integralização igual a 7 anos possuem uma quantidade de desistências menor do que em relação a cursos que possuem esse valor igual a 6 ou 8 anos. Isso pode ser devido à abundância de cursos de medicina nessa categoria, nos quais já foi observado uma menor taxa de evasão em relação a outros cursos [17].

5.2 Limitações e trabalho futuro

As limitações da modelagem desenvolvida nesse trabalho se concentram principalmente na construção de um modelo que corresponda corretamente aos dados. Possivelmente, teria sido mais interessante ter construído um modelo que levasse mais em conta as diferenças entre as modalidades de ensino presencial e a distância. Mais do que isso, a construção de um modelo que conseguisse lidar individualmente com o número de desistências ao longo dos anos de acompanhamento e modelar essa evolução produziria, possivelmente, uma melhor análise do processo de evasão.

Isto posto, para continuar o desenvolvimento da modelagem feito nesse trabalho, é imprescindível a utilização de dados de outros anos disponibilizados pelo Inep, já que seria possível também estudar a mudança no comportamento da evasão no ensino superior ao longo dos últimos anos e relacionar com eventos e mudanças importantes da última década, como o aumento significativo do acesso à Internet.

Bibliografia

[1] R. A. Ambiel, P. A. Cortez, e A. P. Salvador, "Predição da potencial evasão acadêmica entre estudantes trabalhadores e não trabalhadores," *Psicologia: Teoria E Pesquisa*, vol. 37, 2021.

[2] "Indicadores de fluxo da educação superior," Inep. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior

[3] W. Gardner, E. P. Mulvey, e E. C. Shaw, "Regression analyses of counts e rates: poisson, overdispersed poisson, e negative binomial models.," *Psychological Bull.*, vol. 118, no. 3, p. 392, 1995.

[4] S. Coxe, S. G. West, e L. S. Aiken, "The analysis of count data: a gentle introduction to poisson regression e its alternatives," *J. Personality Assessment*, vol. 91, no. 2, pp. 121–136, 2009.

[5] A. C. Cameron, e P. K. Trivedi, "Regression-based tests for overdispersion in the poisson model," *J. Econometrics*, vol. 46, no. 3, pp. 347–364, 1990.

[6] C. Kleiber, e A. Zeileis, *Applied Econometrics With R*, New York: Springer-Verlag, 2008. [Online]. Disponível em: https://cran.r-project.org/package=AER

[7] J. M. Ver Hoef, e P. L. Boveng, "Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?," *Ecology*, vol. 88, no. 11, pp. 2766–2772, 2007.

[8] W. N. Venables, e B. D. Ripley, *Modern Applied Statistics With S*, Fourth, New York: Springer, 2002. [Online]. Disponível em: https://www.stats.ox.ac.uk/pub/MASS4/

[9] R Core Team, *R: A Language e Environment for Statistical Computing*, (2023). [Online]. Disponível em: https://www.r-project.org/

[10] P. K. Dunn, e G. K. Smyth, "Randomized quantile residuals," *J. Comput. Graphical Statist.*, vol. 5, no. 3, pp. 236–244, 1996.

[11] C. Feng, L. Li, e A. Sadeghpour, "A comparison of residual diagnosis tools for diagnosing regression models for count data," *BMC Med. Res. Methodology*, vol. 20, no. 1, pp. 1–21, 2020.

[12] A. Zeileis, e C. Kleiber, *Countreg: Count Data Regression*, (2023). [Online]. Disponível em: https://r-forge.r-project.org/projects/countreg/

[13] T. W. Anderson, "Anderson-darling tests of goodness-of-fit.," *Int. Encyclopedia Statistical Sci.*, vol. 1, pp. 52–54, 2011.

[14] J. Gross, e U. Ligges, *Nortest: Tests for Normality*, (2015). [Online]. Disponível em: https://cran.r-project.org/package=nortest

[15] O. C. d. S. de Almeida, G. Abbad, P. P. M. Meneses, e T. Zerbini, "Evasão em cursos a distância: fatores influenciadores," *Revista Brasileira De Orientação Profissional*, vol. 14, no. 1, pp. 19–33, 2013.

[16] C. A. d. S. Baggi, e D. A. Lopes, "Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica," *Avaliação: Revista Da Avaliação Da Educação Superior (Campinas)*, vol. 16, no. 2, pp. 355–374, 2011.

[17] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, e M. B. d. C. M. Lobo, "A evasão no ensino superior brasileiro," *Cadernos De Pesquisa*, vol. 37, pp. 641–659, 2007.