Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Thaís Zezza Barboza

Análise de Dados Longitudinais com Modelo Linear Misto: Aplicação à Taxa de Permanência Anual em Cursos de Graduação no Brasil

Curitiba
2024

Thaís Zezza Barboza

Análise de Dados Longitudinais com Modelo Linear Misto: Aplicação à Taxa de Permanência Anual em Cursos de Graduação no Brasil

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. José Luiz Padilha da Silva

Curitiba
2024

Análise de Dados Longitudinais com Modelo Linear Misto: Aplicação à Taxa de Permanência Anual em Cursos de Graduação no Brasil

Thaís Zezza Barboza[1]

[1]Universidade Federal do Paraná, Curitiba, PR, Brasil*

Produzidos a partir do Censo da Educação Superior realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), os indicadores de fluxo da educação superior permitem um acompanhamento detalhado de três principais dimensões: permanência, desistência e conclusão. Visando compreender possíveis fatores que afetam a taxa de permanência, este trabalho realiza uma análise longitudinal de cursos de bacharelado em instituições federais e privadas (com e sem fins lucrativos) nos anos de 2013 a 2021, voltando o foco para entender a diferença na variação do indicador de Taxa de permanência para cada uma das grandes áreas da classificação realizada pelo Cine Brasil e categorias administrativas referentes às universidades onde os cursos são ofertados.

Palavras-chave: Indicadores de fluxo, Análise longitudinal, Educação Superior

Originating from the College Education Census conducted by the National Institute for Educational Studies and Research Anísio Teixeira (INEP), the college education flow indicators allow for detailed monitoring of three main dimensions of interest: permanence, dropout, and completion. With the goal of understanding possible factors influencing the permanence rate, this study conducts a longitudinal analysis of bachelor's degree programs in federal and private (both for-profit and non-profit) institutions from 2013 to 2021, focusing on understanding the variation in the permanence rate indicator for each of the major areas classified by Cine Brasil and administration type of the universities where the courses are teached.

Palavras-chave: Flow indicators, Longitudinal Analysis, College Education

1. Introdução

Os indicadores de fluxo da educação superior são, essencialmente, métricas de desempenho desenvolvidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) para o acompanhamento da trajetória acadêmica de estudantes do ensino superior a nível nacional. As bases completas de indicadores em diversas faixas de tempo, sumarizados por curso e instituição de ensino, são disponibilizadas em *Indicadores de Fluxo da Educação Superior - INEP* (2024), e também documentos explicitando a metodologia de cálculo.

Os indicadores disponibilizados consistem em cinco medidas que visam sumarizar características de interesse. Essas medidas são: Taxa de Permanência (TAP), Taxa de Desistência Anual (TADA) , Taxa de Desistência Acumulada (TDA), Taxa de Conclusão Anual (TCAN), Taxa de Conclusão Acumulada (TCA). Os indicadores são calculados ano a ano e de maneira individual para cada curso, em cada uma das instituições de ensino que estavam presentes no Censo da Educação Superior.

Neste trabalho, investiga-se a hipótese de que uma dessas medidas, TAP, varia de maneira diferente considerando como variável explicativa a grande área do curso, incluindo os efeitos de agrupamento por estado em um modelo linear de efeitos mistos. Além disso, também é verificado se há diferença significativa na variação desses valores considerando a categoria da instituição (Federal/Privada) e a região geográfica onde o curso foi ministrado.

2. Dados utilizados e descrição das covariáveis

Os dados utilizados foram obtidos na página *Indicadores de Fluxo da Educação Superior - INEP* (2024), disponibilizados diretamente pelo INEP. No caso deste estudo, foi utilizada a base de dados de 2013 a 2021,

*thaiszezza@gmail.com

Tabela 1: Descrição das variáveis a serem utilizadas

Variável	Descrição
ANO_REF	Ano de referência do valor observado
CATEGORIA_ADMINISTRATIVA	Categoria administrativa da instituição
CODIGO_CURSO	Código identificador do Curso
CODIGO_GRANDE_AREA_CINE	Código da Grande área - CINE
NOME_GRANDE_AREA_CINE	Nome Grande Área - CINE
REGIAO_GEO_CURSO	Região Geográfica onde o curso é ministrado
TAP	Taxa de Permanência
UF_CURSO_COD	Código da UF

Tabela 2: Nº de observações por grande área do CINE

CODIGO_GRANDE_AREA_CINE	NOME_GRANDE_AREA_CINE	n
01	Educação	8
02	Artes e humanidades	555
03	Ciências sociais, comunicação e informação	786
04	Negócios, administração e direito	3243
05	Ciências naturais, matemática e estatística	596
06	Computação e Tecnologias da Informação e Comunicação (TIC)	780
07	Engenharia, produção e construção	97
08	Agricultura, silvicultura, pesca e veterinária	62
09	Saúde e bem-estar	1775
10	Serviços	91

onde o acompanhamento foi realizado considerando cursos que iniciaram no ano de 2013, e as variáveis de interesse foram computadas ao longo dos anos.

Para que a comparação fosse válida, a base foi filtrada para que apenas cursos de bacharelado presenciais com expectativa de integralização em 5 anos fossem analisados. Isso foi feito para que os valores da variável TAP pudessem ser interpretados da mesma maneira para todos os cursos analisados.

As observações também foram filtradas para que não houvessem registros com dados faltantes, garantindo que todos os cursos analisados em todas as instituições de ensino possuíssem registros em todos os anos observados.

3. Análise descritiva

3.1. Análise exploratória da relação entre covariáveis e resposta

O processo de investigação da hipótese se inicia por uma análise exploratória simples das variáveis consideradas na análise. Na Tabela 1, é feita uma breve descrição de todas as variáveis utilizadas. A análise será conduzida apenas para a variável de Taxa de permanência (TAP). Ao longo da análise, "n" será utilizado para determinar número de observações. Em seguida, na Tabela 2 é exibido o número de observações para cada uma das grandes áreas do CINE Brasil (Classificação Internacional Normalizada da Educação Adaptada para Cursos de Graduação e Sequenciais de Formação Específica) *Cine Brasil* (2024). O número é o mesmo para todos os anos analisados.

Inicia-se a análise descritiva verificando se há uma diferença visual clara entre as médias observadas para a variável TAP agrupando as médias por código de grande área do CINE na Figura 1.

A visualização parece tornar claro que algumas médias são diferentes e decrescem de maneira diferente

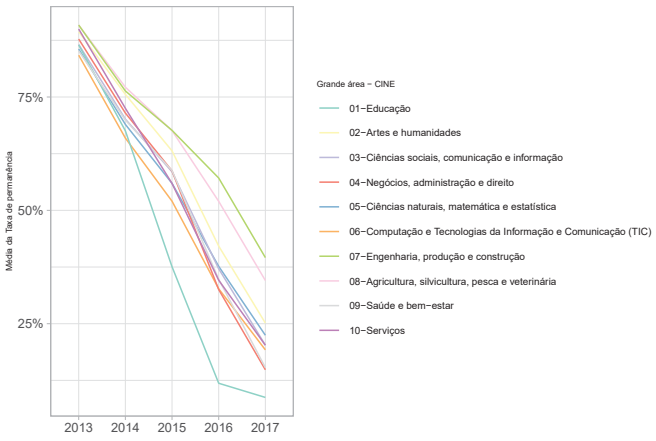

Figura 1: Média de Taxa de permanência por grande área

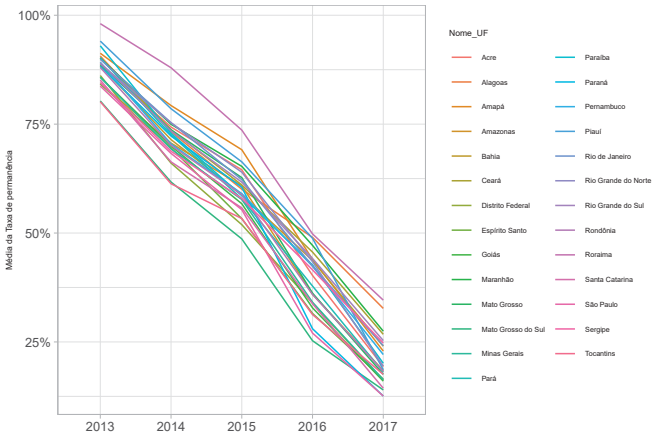

Figura 2: Média de Taxa de permanência por UF

ao longo dos anos. No geral, as taxas iniciam em valores altos e decaem com o passar dos anos.

Nas tabelas e gráficos em seguida, é realizada uma análise similar, porém agrupada por UF. Antes disso, verifica-se quantas observações existem por UF nas bases. Os valores são iguais para todos os anos analisados.

O gráfico de linhas da Figura 2 demonstra que a variável UF é uma provável variável de agrupamento dos efeitos aleatórios, dado que as médias apresentam va-

Tabela 3: Nº de observações por UF

UF_CURSO_COD	Nome_UF	n
11	Rondônia	77
12	Acre	23
13	Amazonas	111
14	Roraima	27
15	Pará	130
16	Amapá	27
17	Tocantins	39
21	Maranhão	116
22	Piauí	119
23	Ceará	188
24	Rio Grande do Norte	128
25	Paraíba	123
26	Pernambuco	244
27	Alagoas	80
28	Sergipe	67
29	Bahia	350
31	Minas Gerais	859
32	Espírito Santo	180
33	Rio de Janeiro	695
35	São Paulo	2117
41	Paraná	637
42	Santa Catarina	367
43	Rio Grande do Sul	560
50	Mato Grosso do Sul	116
51	Mato Grosso	154
52	Goiás	256
53	Distrito Federal	203

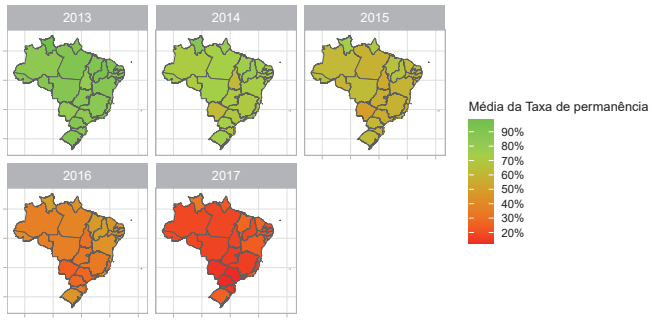

Figura 3: Variação da taxa de permanência ao longo dos anos nos estados Brasileiros

lores muito diferentes ao longo dos anos. Apesar disso, o gráfico é de difícil interpretação pelo número de UFs.

Na Tabela 4 mostrada posteriormente, é apresentada de maneira tabular a diferença entre os valores observados por UF para a média da variável em questão, possibilitando ver a diferença de valores de maneira mais clara.

Tabela 4: Média da taxa de permanência ano a ano por UF

UF_CURSO_COD	Nome_UF	2013	2014	2015	2016	2017
11	Rondônia	88%	70.1%	58.8%	35.9%	18.2%
12	Acre	88.7%	74.3%	64.5%	40.3%	19.3%
13	Amazonas	84.3%	70.6%	60.9%	35.9%	18.4%
14	Roraima	98.1%	87.9%	73.6%	49.8%	34.6%
15	Pará	90.3%	72.5%	57.9%	37.8%	18%
16	Amapá	91.3%	79.3%	69.1%	44%	22.9%
17	Tocantins	80.1%	61.3%	53.3%	31.3%	17.6%
21	Maranhão	90.1%	75%	65.3%	47.2%	27.5%
22	Piauí	94.1%	78.5%	66.3%	48.8%	18.6%
23	Ceará	90.3%	73.7%	61%	45.6%	26.7%
24	Rio Grande do Norte	88.9%	75.3%	64%	43.4%	19.4%
25	Paraíba	93%	72.9%	58.6%	42.7%	20.1%
26	Pernambuco	89.2%	70.5%	59.1%	42.4%	22.1%
27	Alagoas	90.6%	73.3%	60.1%	49%	32.7%
28	Sergipe	84.6%	68.8%	57.9%	41.5%	24.8%
29	Bahia	88.4%	71.4%	58.8%	43.7%	24%
31	Minas Gerais	85.8%	69.3%	57.7%	34.1%	16.3%
32	Espírito Santo	84.2%	69.6%	53.3%	31.5%	16.5%
33	Rio de Janeiro	89.9%	75.3%	62.3%	42.4%	24.1%
35	São Paulo	83.7%	68.3%	55.4%	27%	12.6%
41	Paraná	88.3%	72.3%	60.5%	28%	12.6%
42	Santa Catarina	85.1%	66.3%	55.8%	33.8%	14.4%
43	Rio Grande do Sul	89.1%	73.5%	61.7%	44.1%	25.3%
50	Mato Grosso do Sul	80.3%	61.7%	48.7%	25.3%	14%
51	Mato Grosso	88.6%	73.7%	62.7%	36.2%	17.6%
52	Goiás	85.8%	70%	56.7%	32.8%	16%
53	Distrito Federal	86.2%	66.1%	51.9%	33.9%	16.2%

Tabela 5: Descrição de cada código de categoria administrativa

CATEGORIA_ADMINISTRATIVA	CATEG_NOME	n
1	Pública Federal	1424
2	Pública Estadual	601
3	Pública Municipal	100
4	Privada com fins lucrativos	3071
5	Privada sem fins lucrativos	2756
7	Especial	41

Em seguida, na Figura 3 são exibidos mapas de calor ano a ano para que seja demonstrada de maneira mais intuitiva a variação da variável em cada estado.

Por fim, foi realizada a mesma análise feita anteriormente para a variável CATEGORIA_ADMINISTRATIVA, tentando também evidenciar visualmente a diferença entre médias para cada uma de suas categorias na Figura 4.

4. Modelagem dos dados

4.1. Objetivo da modelagem e justificativa do modelo utilizado

Considerando as aparentes diferenças em médias anuais de TAP (Taxa de permanência) observadas quando os dados são agrupados pelas covariáveis analisadas, parece razoável supor que existe uma relação entre as covariáveis e a resposta.

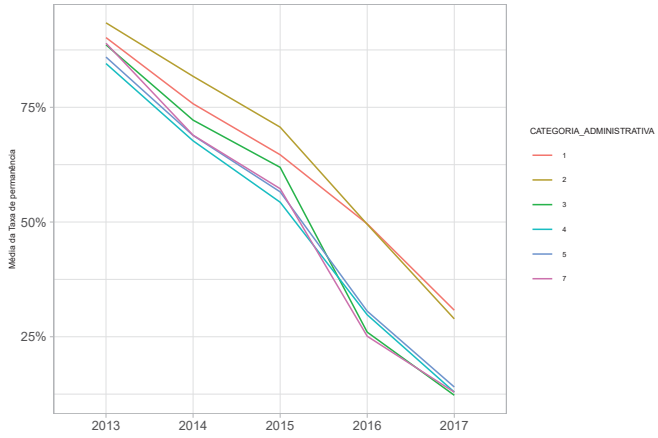

Figura 4: Média de Taxa de permanência por Categoria

Algumas características da estrutura dos dados devem ser consideradas na fase de construção de um modelo que explique a variabilidade da resposta. O fato da variável resposta ser observada múltiplas vezes ao longo de anos para cada indivíduo (curso) requer uma estratégia de modelagem que leve em consideração medidas repetidas ao longo do tempo. Outro ponto observado é a possível correlação entre indivíduos de mesmo estado, dada por fatores não observados.

Por conta dessas características, este trabalho utilizará modelos mistos lineares na fase de modelagem, permitindo a estimação de efeitos fixos (efeitos estimados para todas as observações) e efeitos aleatórios (efeitos estimados para cada um dos indivíduos, representados pelos cursos).

4.2. Exclusão de categorias e categorias de referência

Em relação aos dados originais, foi desconsiderada a categoria "Educação" da variável de grande área do CINE dado que existem apenas 8 cursos nessa categoria. Também foi removida a categoria "Especial" da variável de categoria administrativa, que possui apenas 41 cursos no total.

A diferença de médias por UF também fornece um indicativo de que um modelo considerando a covariância dos indivíduos em função de UF pode ser mais razoável. O modelo será ajustado aninhando os efeitos aleatórios de indivíduo por estado. Os modelos são ajustados utilizando o pacote *nlme* Pinheiro, Bates, and R Core Team (2024).

Como as variáveis preditoras analisadas são categóricas, é importante definir quais são as categorias

Tabela 6: Média do TAP por categoria administrativa

CATEGORIA_ADMINISTRATIVA	2013	2014	2015	2016	2017	Média de TAP
1	90.19	75.77	64.66	49.60	30.78	62.20
2	93.40	81.75	70.67	49.50	28.90	64.85
3	88.61	72.20	61.91	26.00	12.26	52.20
4	84.54	67.70	54.34	29.82	12.97	49.87
5	85.94	68.87	56.60	30.60	14.05	51.21

Tabela 7: Média do TAP por grande área do CINE

CODIGO_GRANDE_AREA_CINE	2013	2014	2015	2016	2017	Média de TAP
02	89.39	75.58	63.13	42.15	25.17	59.08
03	86.64	70.19	58.55	37.23	20.31	54.58
04	87.75	71.44	58.69	32.56	14.84	53.06
05	85.57	68.95	55.92	37.66	22.45	54.11
06	84.19	65.93	52.01	32.75	19.25	50.83
07	90.87	76.33	67.64	57.13	39.58	66.31
08	90.85	77.17	67.58	51.99	34.53	64.42
09	85.15	69.92	58.78	34.85	15.35	52.81
10	90.44	73.31	56.56	34.98	20.48	55.15

Tabela 8: Média do TAP por região geográfica

REGIAO_GEO_CURSO	NOME_REGIAO_GEO	2013	2014	2015	2016	2017	Média de TAP
1	Norte	87.89	72.08	60.43	37.66	19.53	55.52
2	Nordeste	89.81	72.98	60.86	44.50	23.61	58.35
3	Sudeste	85.28	69.82	57.03	31.74	15.76	51.93
4	Sul	87.84	71.36	59.90	35.19	17.57	54.37
5	Centro-Oeste	85.62	68.35	55.37	32.64	16.06	51.61

utilizadas como base de comparação no modelo. Para isso, foram geradas as Tabelas 6, 7 e 8.

Com base nas médias anuais do TAP por variável categórica, foi definido que as categorias que se encontram na posição mediana em termos da média geral ao longo dos anos serão utilizadas como categorias de referência na estimação do modelo. Sendo assim, o ajuste do modelo levou em consideração o a categoria administrativa "3" (Universidades Públicas Municipais) como base de comparação para as outras categorias assim como o código de grande área "03" (Ciências sociais, comunicação e informação) para a variável "CODIGO_GRANDE_AREA_CINE" e a região 4 (Sul) para a variável "REGIAO_GEO_CURSO".

4.3. Modelagem da Taxa de permanência

Para estimar a relação entre a variável resposta TAP (Taxa de permanência) e as covariáveis, foram empregados modelos lineares com efeitos mistos.

Como mencionado anteriormente, os modelos mistos consideram aninhamento de cada curso em suas respectivas UFs, considerando a estrutura dos dados.

Como covariáveis de efeitos fixos, foram empregadas as variáveis Categoria Administrativa, Código de grande área CINE e Região geográfica do curso. Também foram avaliadas as diferenças entre cada um dos anos observados.

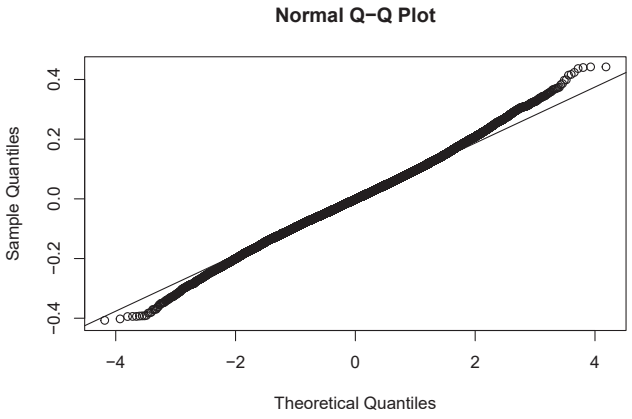

Figura 5: Q-Q plot do modelo

Como a base de dados original possui muitos cursos em que as taxas apresentam valor zero em alguns anos, algumas medidas foram adotadas para reduzir os cursos observados e melhorar a capacidade de ajuste dos modelos.

Primeiramente, a modelagem restringiu os anos levados em consideração aos anos de 2013 a 2017, que correspondem ao período esperado de formação dos alunos dos cursos de duração de 5 anos. Em seguida, todos os cursos que apresentaram taxa de permanência igual a zero em algum dos anos foram removidos do modelo. Isso foi feito para que o modelo de regressão aqui utilizado não encontrasse problemas quando realizassem a estimação dos parâmetros, considerando que não foram feitos para dados inflacionados de zeros.

4.3.1. Modelo ajustado

O modelo ajustado é um modelo de regressão linear com efeitos mistos, levando em conta as variáveis já comentadas como efeitos fixos, e a variável de código de curso aninhada por UF nos efeitos aleatórios, para que a correlação entre cursos de um mesmo estado seja levada em consideração. A Tabela 9 mostra os coeficientes de regressão, bem como seus valores e seus níveis de significância.

4.3.2. Interpretação dos parâmetros

Iniciando pelos coeficientes estimados para a variável ANO_REF, o modelo evidencia o decrescimento da taxa de permanência de maneira geral, considerando as observações com categorias de referência.

Tabela 9: Coeficientes de estimados no modelo 1

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.88	0.02	27892	38.07	0.00
ANO_REF2014	-0.15	0.00	27892	-81.00	0.00
ANO_REF2015	-0.27	0.00	27892	-144.88	0.00
ANO_REF2016	-0.50	0.00	27892	-270.35	0.00
ANO_REF2017	-0.69	0.00	27892	-371.09	0.00
CATEGORIA_ADMINISTRATIVA1	0.07	0.01	6935	4.70	0.00
CATEGORIA_ADMINISTRATIVA2	0.11	0.01	6935	7.11	0.00
CATEGORIA_ADMINISTRATIVA4	-0.04	0.01	6935	-2.47	0.01
CATEGORIA_ADMINISTRATIVA5	-0.02	0.01	6935	-1.57	0.12
CODIGO_GRANDE_AREA_CINE02	0.02	0.01	6935	2.72	0.01
CODIGO_GRANDE_AREA_CINE04	0.02	0.00	6935	3.39	0.00
CODIGO_GRANDE_AREA_CINE05	-0.03	0.01	6935	-4.35	0.00
CODIGO_GRANDE_AREA_CINE06	-0.01	0.01	6935	-2.24	0.03
CODIGO_GRANDE_AREA_CINE07	0.07	0.01	6935	5.38	0.00
CODIGO_GRANDE_AREA_CINE08	0.05	0.02	6935	3.04	0.00
CODIGO_GRANDE_AREA_CINE09	0.01	0.01	6935	1.92	0.06
CODIGO_GRANDE_AREA_CINE10	-0.04	0.01	6935	-2.60	0.01
REGIAO_GEO_CURSO1	0.03	0.02	22	1.46	0.16
REGIAO_GEO_CURSO2	0.03	0.02	22	1.36	0.19
REGIAO_GEO_CURSO3	0.00	0.02	22	0.19	0.85
REGIAO_GEO_CURSO5	-0.03	0.02	22	-1.24	0.23

Para a variável de categoria administrativa, excetuando a categoria 5 (Privada sem fins lucrativos), todas as outras categorias apresentaram diferenças significativas do valor de referência (“Pública Municipal”). As universidades públicas federais e estaduais têm uma taxa de permanência média maior (em 11% e 7%, respectivamente). Em contrapartida, as universidades privadas com fins lucrativos possuem TAP média estimada 4% inferior às universidades públicas municipais.

Tomando como referência a grande área CINE “3” (Ciências sociais, comunicação e informação), todas as outras grandes áreas diferem significativamente quanto aos seus coeficientes estimados. As áreas “02”, “04”, “07” e “08” possuem médias estimadas maiores do que a área 3 em 2%, 2%, 7% e 5%, respectivamente. As áreas “05” ,“06” e “10” apresentam diferenças médias menores estimadas em 3%, 1% e 4%.

As estimativas para região geográfica não foram significativas, o que pode ser explicado pelo maior detalhamento fornecido na especificação dos efeitos aleatórios pela variável referente a UF.

4.4. Conclusão

Após análise descritiva e ajuste dos dados sobre a taxa de permanência através de um modelo linear de efeitos mistos, podemos observar significância estatística para as covariáveis categoria administrativa e grande área do CINE para explicar a variação entre taxa de permanência anual em cursos de bacharelado com duração de 5 anos.

O modelo indica que, de forma geral, universidades públicas possuem uma taxa de permanência mais elevada do que universidades privadas com fins lucrativos. Em relação à grande área do CINE, áreas associadas a ciências humanas, artes, admnistração, engenharia e saúde possuem uma taxa de permanência mais elevada do que cursos relacionados às ciências naturais, exatas e computação.

Este trabalho teve como objetivo a modelagem do TAP, mas trabalhos futuros podem explorar os outros indicadores de fluxo da educação superior no Brasil.

Referências

Cine Brasil. (2024). Retrieved from https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil

Indicadores de Fluxo da Educação Superior - INEP. (2024). Retrieved from https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior

Pinheiro, J., Bates, D., & R Core Team. (2024). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=nlme (R package version 3.1-165)