Modelo de regressão para avaliar a mobilidade social de alunos ingressantes em cursos de engenharia de alta seletividade da Unicamp.

Palavras-Chave: Mobilidade Social, Engenharias, Regressão Linear

Autores(as):

Marcos José Grosso Filho, IMECC – Unicamp

Prof. Rafael Pimentel Maia (orientador), IMECC – Unicamp

Profa. Helena Maria Sant´Ana Sampaio Andery (co-orientadora), FE – Unicamp

Pesquisadora Cibele Yahn de Andrade (co-orientadora), NEPP – Unicamp

INTRODUÇÃO:

A Universidade estadual de Campinas (Unicamp) é uma das maiores e melhores instituições de ensino superior da américa latina, atuando em um dos maiores polos tecnológicos do Brasil e sendo responsável pela formação de profissionais de diversas áreas que assumem protagonismo no apoio e desenvolvimento do país. Em um estudo gerado pela Coordenadoria Geral da Universidade (CGU) em 2021 mostra que a Unicamp foi responsável pela geração de 13,8 bilhões de reais em riqueza para a região de Campinas em 2019. Junto com a geração de riqueza, Campinas assim como todo o Brasil, conta com sérios problemas relacionados à desigualdade social. Segundo (Balbachevsky *et al.*, 2019), o acesso ao ensino superior promove significativas implicações para mobilidade social.

Este estudo, que tem como foco alunos que ingressaram na Unicamp por cursos de engenharia em alta seletividade (Engenharia Civil, Engenharia Mecânica, Engenharia Química, Engenharia Elétrica e Engenharia de Computação) durante os anos de 2003 até 2006, contempla diversas análises que buscam responder à pergunta central deste estudo “Cursos de engenharia de alta seletividade da Unicamp promovem mobilidade social? ”. Foi utilizado técnicas de análise descritiva para entendimento e visualização de dados, além de técnicas inferenciais para modelagem estatística. Como apontado no próprio título, será aplicado um modelo de regressão linear, que tem como objetivo estimar os parâmetros desconhecidos do modelo e será utilizado para investigar a relação entre variáveis e modelar variáveis que possam apontar mobilidade social (como por exemplo renda).

MATERIAIS:

Para tal estudo, será considerado 3 conjuntos de dados que auxiliarão para investigar a questão central deste projeto:

1. Dados relativos ao ingresso à faculdade disponibilizados pela Comissão Permanente para os Vestibulares da Unicamp (COMVEST).
2. Dados relativos à situação acadêmica dos estudantes no egresso fornecidos pela Diretoria Acadêmica (DAC) da Unicamp.
3. Dados relativos às condições de trabalho obtidos pela Relação Anual de Informações Sociais (RAIS) do Ministério da Economia.

Como citado na introdução, será considerado estudantes relativos a 5 cursos de engenharia de alta seletividade da Unicamp (Engenharia Civil, Engenharia Mecânica, Engenharia Química, Engenharia Elétrica e Engenharia de Computação) que ingressaram durante os anos de 2003 a 2006. Na Tabela 1 é possível visualizar todos os 2241 estudantes que foram considerados durante as etapas deste estudo e quantos deles se posicionam em cada curso e ano de ingresso.

Tabela 1: Número de ingressantes em cada um dos cursos de Engenharia considerados por ano de ingresso.

	2003	2004	2005	2006
ENGENHARIA DE COMPUTACAO	92	92	90	94
ENGENHARIA CIVIL	81	82	81	81
ENGENHARIA MECANICA	142	141	136	135
ENGENHARIA DE CONTROLE E AUTOMACAO	50	50	48	51
ENGENHARIA QUIMICA	102	104	95	102
ENGENHARIA ELETRICA	100	102	93	97
TOTAL POR ANO	567	571	543	560

Já através da Tabela 2, observamos que aproximadamente 17% dos matriculados não chegaram a concluir o curso, desta forma, a fim de evitar certos vieses em nossa investigação de mobilidade social na Unicamp, será pertinente uma análise para avaliar qual o perfil socioeconômico destes estudantes que não chegaram a finalizar a graduação.

Tabela 2: Número e proporção da situação dos matriculados nos cursos de Engenharia da Unicamp, por motivo de saída.

Situação	N	%
Concluiu o curso	1876	83.71
Não concluiu o curso	365	16.29
TOTAL	2241	100

Para toda a análise foram consideradas as seguintes variáveis:

- Sexo (Masculino e Feminino).
- Raça (PPI - Pretos, Pardos e Indígenas e BA - Brancos e Amarelos).
- Tipo de escola no ensino médio (Inteiramente pública ou não).
- Tipo de escola no ensino fundamental (Inteiramente pública ou não).
- Se o estudante recebeu isenção na taxa do vestibular da Unicamp (Sim ou Não).
- Ocupação profissional do pai do estudante antes do ingresso na faculdade.
- Ocupação profissional da mãe do estudante antes do ingresso na faculdade.
- Nível de escolaridade do pai do estudante antes do ingresso na faculdade.
- Nível de escolaridade da mãe do estudante antes do ingresso na faculdade.
- Nota de Física na segunda fase do vestibular da Unicamp.
- Nota de Química na segunda fase do vestibular da Unicamp.
- Motivo de saída (concluiu ou não concluiu o curso)
- Vínculo ativo (possui vínculo empregatício ativo ou não).
- Renda Mensal Média em Salários Mínimos.

METODOLOGIA:

Podemos separar o estudo em basicamente 2 etapas e em cada uma delas foram utilizados os seguintes métodos:

1. Análise Descritiva: Foi realizado uma extensa e detalhada análise exploratória dos dados, visando entender amplamente o perfil socioeconômico dos estudantes, assim como, possíveis variáveis correlacionadas com a evasão do curso. Para isto foram utilizadas estatísticas sumárias, métodos gráficos, tabelas de contingência, entre outros.
2. Análise Inferencial: Foi construído testes de hipótese, intervalos de confiança e modelos estatísticos a fim de entender e nos possibilitar tirar conclusões sobre a população relativa aos nossos dados amostrais. Em especial, os modelos aplicados neste estudo são:
    - Modelo de Regressão Linear Múltipla (Weisberg, 2013):

        Utilizado quando há interesse em modelar a relação linear entre a variável de interesse (variável dependente ou variável resposta) com outras variáveis (independentes ou explicativas), para isto, os dados devem respeitar os seguintes pressupostos: homocedasticidade, multicolinearidade entre as variáveis explicativas (a correlação entre elas não pode estar perto de uma correlação perfeita), independência de erros, variância constante e independente das variáveis preditoras.

        Para uma amostra de n indivíduos, o modelo de regressão linear múltipla pode ser expresso por:

        $$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_m X_{mi} + \varepsilon_i,$$

        onde m é o número de variáveis independentes consideradas, X_j é a j-ésima variável independente, β_j é o j-ésimo coeficiente referente à uma das m variáveis explicativas e Y é a variável resposta com $i = 1, \dots, n$ e $j = 1, \dots, m$.

        O objetivo deste modelo é estimar os coeficientes β_j, para este estudo foi considerado o método de mínimos quadrados.
    - Modelo de Regressão Logística Dicotômica (Giolo, 2017):

        Utilizado para calcular a chance de um determinado evento categórico associado à uma variável resposta em função de uma ou mais variáveis explicativas, o modelo pode ser expresso por:

        $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m = \ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right),$$

        onde β_j são os parâmetros de regressão, $\boldsymbol{X} = (X_1, \dots, X_m)$ e $p(\boldsymbol{X})$ é a probabilidade do evento de interesse ocorrer.

        Tal transformação é chamada de logito e como a razão de $p(\boldsymbol{X})$ e $1 - p(\boldsymbol{X})$ define uma chance (Giolo, 2017), segue que a chance é definida por:

        $$chance = \frac{p(\boldsymbol{X})}{1 - p(\boldsymbol{X})} = e^Y$$

É pertinente salientar que para ambos os modelos foi realizado o método stepwise de seleção de modelos. Especificadamente o método Backward Selection, ou seja, foi considerado inicialmente um modelo completo e foi sendo retirado as variáveis explicativas que não apresentavam iteração significativa com a variável resposta e/ou levando em consideração o coeficiente de determinação ajustado. Será utilizada como ferramenta principal para realizar análises propostas e ajuste dos modelos a linguagem de programação R (R Core Team, 2022).

RESULTADOS E DISCUSSÃO:

Inicialmente foi feita uma análise sobre a evasão do curso, isto é, foi feito uma análise descritiva detalhada e testes de independência a partir de tabelas de contingência para verificar a associação entre a variável “Motivo de saída” e as demais variáveis explicativas. Este estudo inicial apontou que dois indicadores

socioeconômicos ("Sexo" com um $p - valor$ de 0,018 e "Raça" $p - valor$ de 0,002) possuem estatisticamente uma associação significativa (considerando um nível de significância $\alpha = 0{,}05$) com a evasão do curso.

Em seguida, com o mesmo objetivo, foi modelado um modelo de Evasão (Regressão Logística Dicotômica). Na Tabela 3 podemos ver as estimativas para os coeficientes do modelo, assim como o erro padrão, a razão de chances (RC) e o respectivo intervalo de confiança de 95% para a RC.

Tabela 3: Estimativas e razão de chance (RC) obtidos a partir do Modelo de Regressão Logística para a variável "Motivo de saída". RC – Razão de Chances, BA – Brancos e Amarelos, PPI – Pretos, Pardos e Indígenas, f2 – Segunda Fase do Vestibular da Unicamp, 2,5% representa o limite inferior do intervalo de confiança de 95% e 97,5% representa o limite superior do intervalo de confiança de 95%.

	Nível	Coeficiente	Erro Padrão	RC	2,5%	97,5%
Intercepto		-1.235	0.364	0.291	0.141	0.590
Sexo	Masculina (ref: Feminina)	0.547	0.182	1.728	1.223	2.498
Raça	PPI (ref: BA)	0.583	0.166	1.791	1.286	2.464
Nota física f2		-0.011	0.009	0.989	0.972	1.007
Nota química f2		-0.016	0.007	0.985	0.971	0.998

O modelo nos aponta que um estudante do sexo masculino possui uma chance de aproximadamente 73% maior de evadir, quando comparado com a um estudante do sexo feminino. De maneira análoga, o grupo de Pretos, Pardos e Indígenas possui uma chance maior de evasão de curso (cerca de 79%) quando comparado ao grupo de Brancos e Amarelos. Para as variáveis independentes numéricas (nota de física e de química) iremos considerar de maneira hipotética duas pessoas, uma que tirou uma nota 20 e outra que tirou nota 50. Desta forma, a chance de evasão será calculada da seguinte forma:

$$chance = \frac{\exp(20\beta)}{\exp(50\beta)},$$

onde β deve ser substituído pelas respectivas estimativas dos coeficientes de cada uma das variáveis. Assim, percebemos que uma pessoa que tirou 20 em física na segunda fase do vestibular, possui uma chance de aproximadamente 38% maior de evadir quando comparada a alguém que tirou 50. De maneira análoga, quando analisamos a prova de química esta chance aumenta para aproximadamente 59%.

Todos os outros indicadores socioeconômicos que foram apresentados ao início deste relatório e que não apareceram nestes resultados sobre a evasão de curso, não evidenciaram através dos dados utilizados associações significativas com a variável resposta.

A fim de avaliar a mobilidade social, foi realizado um corte de 6 anos para a coleta dos dados do conjunto da Rais a partir do ano de formatura do estudante, isto é, para aqueles que chegaram a completar a graduação, estaremos analisando o indivíduo no mercado de trabalho 6 anos após sua formatura. Na Tabela 4 podemos ver a disposição destes indivíduos dentre os anos que concluíram o curso.

Tabela 4: Número de pessoas que aparecem no conjunto de dados da Rais, considerando uma janela temporal de 6 anos após o ano de conclusão de curso.

	Ano de Conclusão	Numero de pessoas (Conclusão)	Ano Rais	Numero de pessoas (Rais)
1	2006	6	2012	3
2	2007	114	2013	67
3	2008	224	2014	142
4	2009	248	2015	174
5	2010	228	2016	148
6	2011	221	2017	151
7	2012	85	2018	60
Total		1126		745

A partir destes dados, foi ajustado um Modelo de Regressão Linear Múltipla para modelar o logaritmo natural da variável "Renda Mensal Média em Salários Mínimos". Na Tabela 5, visualizamos as estimativas para os coeficientes do modelo, o intervalo de confiança de 95% para as estimativas, o erro padrão e o $p - valor$ do Teste T de significância, que basicamente testa se a variável tem um impacto nulo ou não no modelo.

Tabela 5: Resultados do Modelo de Regressão Linear para o logaritmo natural da variável "Renda Mensal Média em Salários Mínimos". ref. – Categoria de Referência, 2,5% representa o limite inferior do intervalo de confiança de 95%, 97,5% representa o limite superior do intervalo de confiança de 95%, EM - Ensino Médio e Pública – Inteiramente Pública.

	Nível	Coeficientes	2,5%	97,5	Erro Padrão	P-Valor
Intercepto		1.825	1.720	1.929	0.053	0.000
Sexo	Feminina (ref: Masculina)	-0.227	-0.368	-0.085	0.072	0.002
Tipo de escola (EM)	Pública (ref: Particular)	-0.219	-0.397	-0.041	0.091	0.016

Podemos ver que para a variável sexo, a categoria de referência é "Masculino" e o coeficiente é negativo (-0,227). Já para a variável tipo de escola no ensino médio, a categoria é "Inteiramente Pública" e o coeficiente também é negativo (-0,219).

Como modelamos o logaritmo da renda mensal média em salários mínimos, para interpretar os coeficientes devemos aplicar a função exponencial nestes valores a fim de obter uma aproximação da relação entre as variáveis explicativas com a variável resposta ($EFEITO_i = e^{\beta_i}$, onde $i = \{1, 2, 3\}$ representa os coeficientes em ordem). Desta forma, podemos afirmar que $EFEITO_2 = e^{-0,227} \cong e^{-0,219} = EFEITO_3 \cong 0,8$. Ou seja, tanto as mulheres, quanto as pessoas que fizeram o ensino médio inteiramente em escola pública, tendem a receber em salários mínimos mensais algo em torno de 20% a menos do que homens e pessoas que estudaram inteiramente ou parcialmente em escola particular, respectivamente

CONCLUSÕES:

Destarte, a partir do Modelo de Regressão Linear Múltipla, vemos que, após 6 anos da conclusão de curso, a variável raça não possui uma relação linear significativa com o logaritmo da variável renda mensal em salários mínimos, porém, não podemos esquecer que de maneira significativa, pessoas PPI estão mais suscetíveis a não concluir o curso durante a graduação quando comparados às pessoas BA (como visto durante a análise descritiva e inferencial). Entretanto, quando analisamos a variável "Sexo", vemos uma certa vantagem do sexo feminino em completar a graduação quando comparado ao sexo masculino, porém, em contrapartida, uma certa desvantagem em salários mínimos mensais após de formadas. Por fim, a variável "Tipo de escola no ensino médio" aponta que pessoas que tem condição de estudar em escolas particulares em algum momento durante o ensino médio, tendem a ter uma melhor renda quando comparados às pessoas que realizaram o ensino médio inteiramente em escola pública.

BIBLIOGRAFIA

Balbachevsky, E., Sampaio, H., and Andrade, C. Y. (2019). *Expanding Access to Higher Education and Its (Limited) Consequences for Social Inclusion: The Brazilian Experience*. Social Inclusion, 7(1).

R Core Team (2022). R: *A language and environment for statistical computing. R* Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Weisberg, S. (2013) Applied Linear Regression, 4th Edition, John Wiley & Sons.

GIOLO, Suely R. Introdução à análise de dados categóricos com aplicações. 1° ed. 2017.