Tutorial 2 - Lucas Teruo Imaizumi

Lucas Teruo Imaizumi

02/03/2021

## Paper - Os determinantes do Desempenho Escolar e Efeitos de Tamanho da sala no desempenho dos alunos: Evidências usando regressões descontínuas.

**Principais Pontos/ Resumo Curto:** Menezes Filho 2007 e Matavelli/Menezis Filho 2020

O Papar trabalha a questão da educação, o primeiro (2007) traz uma revisão do panorama da educação usando estatística descritivas do Brasil em comparação com outros países, após um breve relato histórico de como a educação foi construída no país, o autor tentar descobrir o impacto de diversas variáveis sobre a nota dos alunos, durante a pesquisa ele tem achados importantes, tendo como principal o impacto das variáveis: escolaridade *das mães*, o número de horas na escola e algumas outras de menor relevância sobre a nota dos alunos. Além de que os modelos predizem apenas 20% da nota, sendo o restante determinado por variáveis não observadas (como esforço e facilidade em aprendizado). Fora isso o autor também traz uma série de variáveis que não são relevantes, apesar de intuitivamente parecerem, como o salário dos professores, número de computadores e tamanho da sala, sendo esse último alvo de sua tese seguinte, onde este explora melhor essa relação utilizando método da regressão descontinuada, este acha que não relevância sobre o tamanho da turma na nota dos alunos.

## Tutorial 2:

# 1

#Mudar o diretório quando trocar de PC!  
library(readxl)  
notas <- read\_excel("C:/Users/Daquisu/Documents/nota-alagoas.xls")

Primeiramente é importante falar que o ponto ótimo dos gráficos apresentados não implica que o tamanho da turma afeta negativamente o desempenho, Menezis Filho (2007,2020) traz evidências que quando adicionados outros controles, o tamanho da turma não tem significância sobre a nota (pelo menos para os 3 Estados estudados), sendo assim é preciso verificar se essa hipótese é verdadeira para Alagoas também, as turmas maiores podem estar atreladas a municípios/regiões com menos escolas e consequentemente mais pobres afetando assim a nota dos alunos. Para observar isso podemos criar um modelo de previsão das notas com os dados apresentados.

Observamos que com os dados apresentados podemos construir um modelo para as escolas e um modelo para os alunos. Mas vamos focar nas notas dos alunos.

library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(sjPlot)

## Warning: package 'sjPlot' was built under R version 4.0.4

Modelo1 <- lm(nota\_lp ~ tamturma + tamturma2 + mulher + negro + computador + moramae + educmae1 + educmae2 + educmae3 + educmae4 + educmae5 + educmae6 + morapai + educpai1 + educpai2 + educpai3 + educpai4 + educpai5 + educpai6 + iraula + livro + reprovou + trabalha + abandonou + reforço + extracurricular + merenda + interrupção + semprof + nivel\_socio\_economico, data=notas, na.action=na.exclude)  
summary(Modelo1)

##   
## Call:  
## lm(formula = nota\_lp ~ tamturma + tamturma2 + mulher + negro +   
## computador + moramae + educmae1 + educmae2 + educmae3 + educmae4 +   
## educmae5 + educmae6 + morapai + educpai1 + educpai2 + educpai3 +   
## educpai4 + educpai5 + educpai6 + iraula + livro + reprovou +   
## trabalha + abandonou + reforço + extracurricular + merenda +   
## interrupção + semprof + nivel\_socio\_economico, data = notas,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -108.079 -28.721 -3.536 24.955 171.542   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 198.297915 11.332853 17.498 < 2e-16 \*\*\*  
## tamturma 0.426164 0.457713 0.931 0.351835   
## tamturma2 -0.005148 0.007054 -0.730 0.465507   
## mulher 5.281982 0.769420 6.865 7.00e-12 \*\*\*  
## negro -7.120530 1.226528 -5.805 6.59e-09 \*\*\*  
## computador 1.625921 1.405726 1.157 0.247443   
## moramae 3.241048 1.214153 2.669 0.007610 \*\*   
## educmae1 -8.683650 1.768302 -4.911 9.20e-07 \*\*\*  
## educmae2 -6.403976 1.500420 -4.268 1.99e-05 \*\*\*  
## educmae3 -4.723035 1.489328 -3.171 0.001522 \*\*   
## educmae4 -3.835757 1.604841 -2.390 0.016859 \*   
## educmae5 7.341187 1.572107 4.670 3.05e-06 \*\*\*  
## educmae6 NA NA NA NA   
## morapai -0.548660 0.829715 -0.661 0.508457   
## educpai1 -5.722261 1.612732 -3.548 0.000389 \*\*\*  
## educpai2 -2.616924 1.472504 -1.777 0.075563 .   
## educpai3 0.505326 1.505708 0.336 0.737174   
## educpai4 -0.067738 1.574122 -0.043 0.965676   
## educpai5 8.293491 1.590829 5.213 1.89e-07 \*\*\*  
## educpai6 NA NA NA NA   
## iraula 14.775579 1.515082 9.752 < 2e-16 \*\*\*  
## livro 12.021457 1.662185 7.232 5.05e-13 \*\*\*  
## reprovou -19.325402 0.854930 -22.605 < 2e-16 \*\*\*  
## trabalha -18.583575 0.981361 -18.937 < 2e-16 \*\*\*  
## abandonou -3.401860 1.240154 -2.743 0.006096 \*\*   
## reforço -1.012373 1.096435 -0.923 0.355854   
## extracurricular 1.608248 0.750814 2.142 0.032214 \*   
## merenda 1.878573 1.305869 1.439 0.150302   
## interrupção -1.979157 0.879566 -2.250 0.024458 \*   
## semprof -3.359566 0.873635 -3.846 0.000121 \*\*\*  
## nivel\_socio\_economicoBaixo -43.368818 8.706243 -4.981 6.41e-07 \*\*\*  
## nivel\_socio\_economicoMédio -33.633924 8.668694 -3.880 0.000105 \*\*\*  
## nivel\_socio\_economicoMédio Alto -39.825659 9.377145 -4.247 2.18e-05 \*\*\*  
## nivel\_socio\_economicoMédio Baixo -39.448908 8.669501 -4.550 5.41e-06 \*\*\*  
## nivel\_socio\_economicoMuito Baixo -49.521692 10.257665 -4.828 1.40e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.02 on 11503 degrees of freedom  
## (27846 observations deleted due to missingness)  
## Multiple R-squared: 0.1918, Adjusted R-squared: 0.1896   
## F-statistic: 85.33 on 32 and 11503 DF, p-value: < 2.2e-16

Modelo2 <- lm(nota\_lp ~ tamturma + tamturma2 + mulher + negro + computador + iraula + livro + reprovou + trabalha + abandonou + reforço + extracurricular + merenda + interrupção + semprof + nivel\_socio\_economico, data=notas, na.action=na.exclude)  
summary(Modelo2)

##   
## Call:  
## lm(formula = nota\_lp ~ tamturma + tamturma2 + mulher + negro +   
## computador + iraula + livro + reprovou + trabalha + abandonou +   
## reforço + extracurricular + merenda + interrupção + semprof +   
## nivel\_socio\_economico, data = notas, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -108.91 -29.12 -3.50 25.86 167.32   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 191.281750 7.895292 24.227 < 2e-16 \*\*\*  
## tamturma 0.958126 0.300437 3.189 0.00143 \*\*   
## tamturma2 -0.013241 0.004639 -2.854 0.00431 \*\*   
## mulher 4.089525 0.513704 7.961 1.78e-15 \*\*\*  
## negro -5.864422 0.843515 -6.952 3.68e-12 \*\*\*  
## computador -2.854452 1.001741 -2.849 0.00438 \*\*   
## iraula 16.796975 1.058699 15.866 < 2e-16 \*\*\*  
## livro 14.042306 1.067008 13.160 < 2e-16 \*\*\*  
## reprovou -22.202779 0.563641 -39.392 < 2e-16 \*\*\*  
## trabalha -19.255273 0.693851 -27.751 < 2e-16 \*\*\*  
## abandonou -5.237573 0.856038 -6.118 9.59e-10 \*\*\*  
## reforço -0.264868 0.740240 -0.358 0.72049   
## extracurricular 1.631633 0.502440 3.247 0.00117 \*\*   
## merenda 0.138645 0.871118 0.159 0.87355   
## interrupção -1.453651 0.583307 -2.492 0.01271 \*   
## semprof -3.011887 0.580547 -5.188 2.14e-07 \*\*\*  
## nivel\_socio\_economicoBaixo -46.606133 6.323316 -7.371 1.75e-13 \*\*\*  
## nivel\_socio\_economicoMédio -32.479844 6.308321 -5.149 2.64e-07 \*\*\*  
## nivel\_socio\_economicoMédio Alto -37.310146 6.735408 -5.539 3.06e-08 \*\*\*  
## nivel\_socio\_economicoMédio Baixo -40.471227 6.305630 -6.418 1.40e-10 \*\*\*  
## nivel\_socio\_economicoMuito Baixo -50.232922 7.263541 -6.916 4.76e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.41 on 26096 degrees of freedom  
## (13265 observations deleted due to missingness)  
## Multiple R-squared: 0.1597, Adjusted R-squared: 0.1591   
## F-statistic: 248 on 20 and 26096 DF, p-value: < 2.2e-16

Foram criados dois modelos tentando prever a nota em língua portuguesa, a diferença entre os 2 está na omissão da variável dummies escolaridade dos pais, talvez ao utilizá-la ocorremos no risco de homoestaticidade (ou seja alta correlação entre as variáveis), já que na média pessoas com mais escolaridade possuem maior produtividade e consequentemente renda conforme diz a teoria econômica do capital humano. Além disso no segundo modelo o tamanho da classe torna-se relevante, sendo assim, o ponto máximo da turma é a , ou seja , resolvendo a equação achamos que . O valor aparenta estar um pouco mais à direita do valor apresentado e siginifica que levando em contas as variáveis do modelo e tudo o mais constante, o valor que maximiza a nota média em português é , seguindo essa mesma metodologia vamos tentar achar o valor que maximiza a nota em matemática.

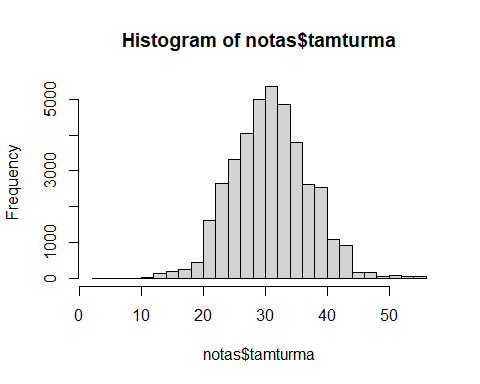
Modelo3 <- lm(nota\_mt ~ tamturma + tamturma2 + mulher + negro + computador + iraula + livro + reprovou + trabalha + abandonou + reforço + extracurricular + merenda + interrupção + semprof + nivel\_socio\_economico, data=notas, na.action=na.exclude)  
summary(Modelo3)

##   
## Call:  
## lm(formula = nota\_mt ~ tamturma + tamturma2 + mulher + negro +   
## computador + iraula + livro + reprovou + trabalha + abandonou +   
## reforço + extracurricular + merenda + interrupção + semprof +   
## nivel\_socio\_economico, data = notas, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -95.836 -27.235 -4.159 23.483 180.936   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 206.717130 7.446469 27.760 < 2e-16 \*\*\*  
## tamturma 0.805404 0.283358 2.842 0.00448 \*\*   
## tamturma2 -0.011317 0.004375 -2.587 0.00969 \*\*   
## mulher -7.817233 0.484501 -16.135 < 2e-16 \*\*\*  
## negro -5.476124 0.795564 -6.883 5.98e-12 \*\*\*  
## computador -2.114450 0.944795 -2.238 0.02523 \*   
## iraula 12.091420 0.998516 12.109 < 2e-16 \*\*\*  
## livro 11.413426 1.006352 11.341 < 2e-16 \*\*\*  
## reprovou -18.332369 0.531600 -34.485 < 2e-16 \*\*\*  
## trabalha -14.474889 0.654408 -22.119 < 2e-16 \*\*\*  
## abandonou -4.357819 0.807375 -5.398 6.82e-08 \*\*\*  
## reforço -1.053940 0.698160 -1.510 0.13116   
## extracurricular 2.930347 0.473878 6.184 6.35e-10 \*\*\*  
## merenda 1.623022 0.821597 1.975 0.04823 \*   
## interrupção -2.167643 0.550148 -3.940 8.17e-05 \*\*\*  
## semprof -4.557701 0.547545 -8.324 < 2e-16 \*\*\*  
## nivel\_socio\_economicoBaixo -34.457298 5.963855 -5.778 7.66e-09 \*\*\*  
## nivel\_socio\_economicoMédio -25.012463 5.949713 -4.204 2.63e-05 \*\*\*  
## nivel\_socio\_economicoMédio Alto -29.696833 6.352521 -4.675 2.96e-06 \*\*\*  
## nivel\_socio\_economicoMédio Baixo -29.139078 5.947175 -4.900 9.66e-07 \*\*\*  
## nivel\_socio\_economicoMuito Baixo -39.320695 6.850631 -5.740 9.59e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 38.11 on 26096 degrees of freedom  
## (13265 observations deleted due to missingness)  
## Multiple R-squared: 0.1107, Adjusted R-squared: 0.11   
## F-statistic: 162.4 on 20 and 26096 DF, p-value: < 2.2e-16

Usando as mesmas variáveis do modelo anterior podemos realizar o mesmo procedimento de maximização através da derivada parcial da equação de regressão, ou seja, , ou seja, sendo assim chegamos no valor de .

Com relações ao custo podemos usar duas métricas para observar se os números apontados como máximo são maiores ou menores do que as médias das turmas atuais, caso sejam maiores o ponto ótimo das turmas implicará na média uma redução de custos (visto que se poderá colocar mais alunos numa mesma sala e talvez reduzir turmas) e caso menores um aumento nos custos para se atingir o ótimo.

hist(notas$tamturma, breaks = 20)



mean(notas$tamturma, na.rm = TRUE)

## [1] 31.32535

sd(notas$tamturma, na.rm = TRUE)

## [1] 6.14504

A distribuição aparenta ser uma normal, então podemos utilizar a função de probabilidade para estimar quantas escolas aumentariam suas turmas para chegar ao máximo e quantas teriam que diminuir

pnorm(35.58, mean = 31.32535, sd = 6.14504)

## [1] 0.7556479

pnorm(47.9, mean = 31.32535, sd = 6.14504)

## [1] 0.9965042

Segundo esses dados 75% das escolas enfrentariam uma redução da turma, todavia como na prática, como o eixo x do gráfico só existe nos inteiros, é provável que esse número esteja um pouco superestimado, todavia a garantia é que pelos mais da metade das escolas poderia aumentar as turmas e talvez enfrentar algum ganho, mas se usarmos o ótimo da turma de português como referencial, nesse caso maioria 99% das escolas poderiam ter ganhos advindos do aumento de alunos por turma e consequente redução de classes, todavia conforme a concavidade da função esse ganho poderia prejudicar o desempenho dos alunos em matemática. E é importante lembrar da premissa que turmas com mais alunos implicam em redução de custos.

\*Obs: Como trata-se de microdados, as turmas são repetidas na base, visto que 1 turma possuiu vários alunos, mas eu imagino que isso não gere uma distorção nos dados

# 2.

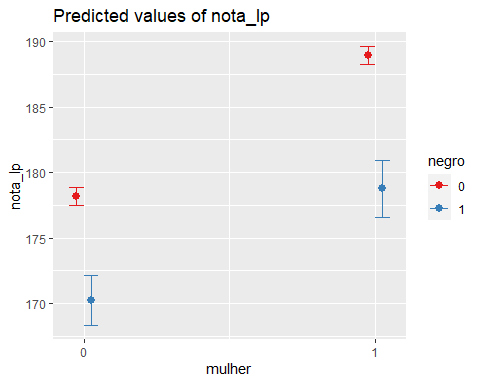
Em seguida vamos estimar os termos de interação entre alguns grupos de interesse da secretaria.

Primeiro iremos observar a interação entre alunos e alunas negros e não negros na nota de português.

library(sjPlot)  
notas3 <- notas  
notas3$mulher <- as.factor(notas3$mulher)  
notas3$negro <- as.factor(notas3$negro)  
modelo4 <- lm(data = notas3, formula = nota\_lp ~ mulher + negro + mulher:negro)  
summary(modelo4)

##   
## Call:  
## lm(formula = nota\_lp ~ mulher + negro + mulher:negro, data = notas3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -100.22 -32.88 -5.39 28.52 167.06   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 178.1818 0.3333 534.555 < 2e-16 \*\*\*  
## mulher1 10.7732 0.4809 22.402 < 2e-16 \*\*\*  
## negro1 -7.9525 1.0225 -7.777 7.61e-15 \*\*\*  
## mulher1:negro1 -2.2379 1.5379 -1.455 0.146   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.73 on 36755 degrees of freedom  
## (2623 observations deleted due to missingness)  
## Multiple R-squared: 0.01835, Adjusted R-squared: 0.01827   
## F-statistic: 229.1 on 3 and 36755 DF, p-value: < 2.2e-16

plot\_model(modelo4, type = "pred", terms = c("mulher","negro"))



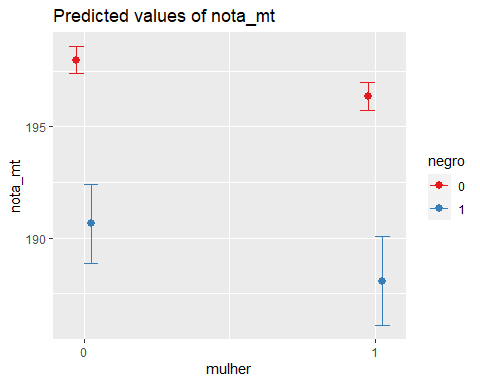
É possível tirar duas conclusões do gráfico apresentado, primeiro que existe uma diferença grande entre as notas das meninas e meninos, todavia isso já era esperado, como apontado no paper de FILHO (2007) onde essa diferença se apresentava até o momento em todos os dados já analisados. Outro ponto importante é que existe uma diferença grande entre alunos negros e não negros em ambos os gêneros.

Seguindo para a nota de matemática temos:

library(sjPlot)  
modelo5 <- lm(data = notas3, formula = nota\_mt ~ mulher + negro + mulher:negro)  
summary(modelo5)

##   
## Call:  
## lm(formula = nota\_mt ~ mulher + negro + mulher:negro, data = notas3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -76.997 -29.934 -5.721 24.288 171.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 197.9911 0.3094 639.848 < 2e-16 \*\*\*  
## mulher1 -1.6389 0.4464 -3.671 0.000242 \*\*\*  
## negro1 -7.3539 0.9492 -7.747 9.64e-15 \*\*\*  
## mulher1:negro1 -0.9336 1.4277 -0.654 0.513161   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.59 on 36755 degrees of freedom  
## (2623 observations deleted due to missingness)  
## Multiple R-squared: 0.003656, Adjusted R-squared: 0.003575   
## F-statistic: 44.96 on 3 and 36755 DF, p-value: < 2.2e-16

plot\_model(modelo5, type = "pred", terms = c("mulher","negro"))



Nesse caso se inverte a situação e os meninos se saem um pouco melhor em matemática, o que novamente é apontada por FILHO (2007), apesar disso, as diferenças de cor continuam significantes, e não negros se saem melhor tanto para meninos quanto para meninas.