Instituto Brasileiro de Geografia e Estatística
Escola Nacional de Ciências Estatísticas
Bacharelado em Estatística

Rafael Cabral Fernandez & João Victor Messa

A proficiência em Ciências Humanas dos alunos submetidos ao Exame Nacional do Ensino Médio em 2016: Uma modelagem linear

Brasil

2018, Julho

Resumo

O estudo aqui apresentado tem por objetivo analisar a proficiência em Ciências Humanas dos alunos que foram submetidos ao Exame Nacional do Ensino Médio, ou ENEM, no ano de 2016. A análise proposta é mediante a formulação de um Modelo de Regressão Linear Múltiplo. Com base em uma amostra de 350 estudantes, o modelo é calcado numa combinação de fatores interdisciplinares, informativos e educacionais.

Palavras-chaves: enem. regressão linear múltipla. ciências humanas

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

O Exame Nacional do Ensino Médio (ENEM) é um rito de passagem para todo jovem estudante brasileiro, aonde o mesmo é submetido a uma série de avaliações de forma que o resultado final pode ser determinante na formação da carreira profissional daquele indivíduo. De caráter amplo, geral e irrestrito, o ENEM é pivotal para a compreensão do sistema de ensino e educação nacional, tão bem quanto a sua interação com o mercado de trabalho, à luz das evidentes diferenças culturais e socioeconômicas históricas observadas na vastidão do universo estudantil.

A Teoria Interdisciplinar, difundida recentemente entre os pedagogos, assenta que nenhuma área da ciência caminha sozinha no que tange ao ensino, isto é, todas as disciplinas da base curricular nacional conversam entre si. A grosso modo, partindo deste pressuposto, o aprendizado total de um estudante é uma grandeza positiva, ou seja, podemos mensurar uma única disciplina, a partir das demais. Utilizaremos, para o fim deste estudo, as proficiências de Redação, Linguagens e Ciências da Natureza para mensurar Ciências Humanas.

Motivado pelo o que foi apresentado e baseado no referencial teórico de Neter *et al.* (1989), o documento a seguir apresenta uma abordagem por modelo de regressão linear múltiplo, descrito através de 6 covariáveis. São elas, três das já apresentadas ciências interdisciplinares aliadas a mais três variáveis categóricas, sendo estas: Sexo, presença de TV por assinatura e intenção de Ensino Superior Privado. Tais covariáveis serão motivadas na Revisão Bibliográfica.

Foram analisados alunos da rede privada de ensino que estavam cursando e concluiriam o Ensino Médio em 2016, e que realizaram o ENEM 2016.

1 Revisão Bibliográfica

Em Guerra *et al.* (2014) é apresentado uma proposta de aplicação de múltiplos modelos de regressão (ensembles) para prever a demanda potencial por vagas de ensino superior no ensino público brasileiro. Foram utilizadas variáveis socioeconômicas e educacionais disponibilizadas por MEC, INEP e IBGE para construir modelos de regressão que prevêem a quantidade atual de alunos matriculados em cada município brasileiro. Tornando possível a aplicação para o artigo aqui apresentado, mediante a não-hierarquização dos dados.

Em uma análise retrospectiva do ENADE, Exame Nacional de Desempenho, encontrada em Garcia (2014) aponta que os cursos dentro da Grande Área de Ciências Humanas tem alta renda familiar ao passo que a renda individual do aluno é baixa ou nula, em outras palavras, o aluno de um curso de humanas é, em geral, proveniente de uma família de classe média-alta ou alta, com razoável tempo livre, sendo alguém mais propenso, como menciona Fidalgo (1996), a consumir informação através de veículos como Televisão por Assinatura, sendo portanto justificado a presença de uma variável binária que calcule este efeito.

Sob uma ótica micro, em contrapartida, os cursos de humanas, em sua maioria, são demandados por mulheres brancas, negras e mulatas, como menciona Oliveira (2011) e de fato a amostra de 350 estudantes coaduna com a proposição. No entanto, mesmo em maioria, o desempenho da categoria mencionada é inferior ao desempenho masculino, ainda segundo Oliveira (2011), baseado na teoria do capital social. Sendo, portanto, possível analisar uma suspeita de que o sexo exerce alguma influência no desempenho dos alunos via vestibular.

Por fim, Neves (2002) faz uma reflexão sobre a expansão do PROUNI, analisando as condições de acesso aos estudantes pobres e questionando o programa como política pública de democratização. Aponta que grande parte do programa é voltado para a área Tecnológica, o que causa uma fuga dos pretendentes da área de Ciências Humanas para o Ensino Público, viabilizando portanto o estudo de uma variável que mensure tal evasão.

2 Análise Exploratória de Dados

2.1 Covariâncias

A primeira análise será feita a partir da correlação entre as covariaveis contínuas com a variável resposta. Embora, a rigor, não seja uma análise exploratória.

Tabela 1 – Correlação entre as ciências

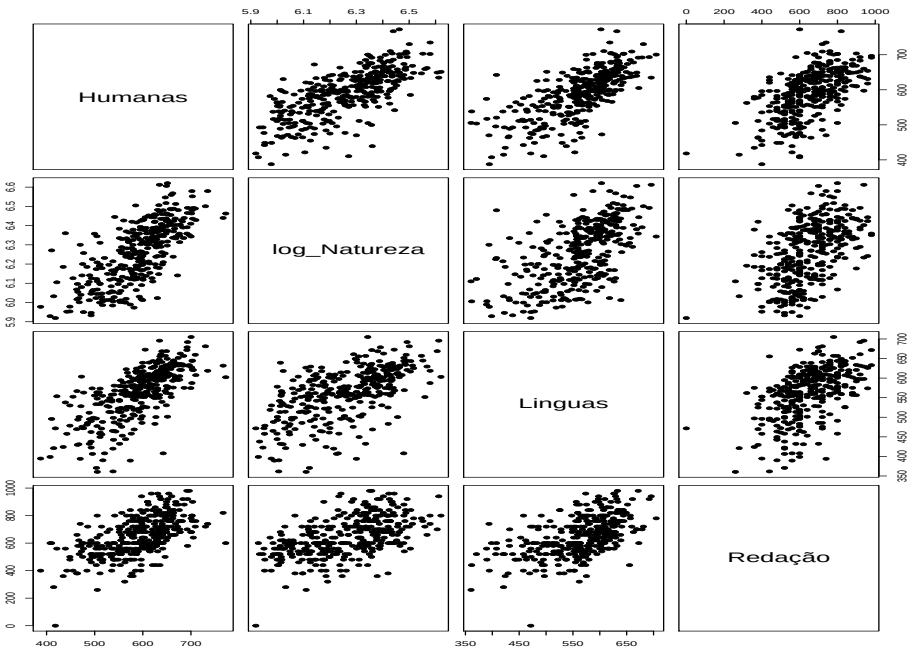
Correlação	Natureza	Línguas	Redação
Humanas	0.71	0.72	0.6

Fonte: Microdados ENEM 2016 - INEP

Por mais que seja tentador estender o conceito univariado da correlação para uma regressão múltipla, não é possível. Contudo, para efeito de conclusão, a proficiência em Humanas é bem explicada, ressalta-se, individualmente, pelas demais ciências, sendo estas: Ciências da Natureza, Linguagens e Redação. Como visto na tabela 1.

2.2 Dispersão

Figura 1 – Gráfico de Dispersão dois-a-dois

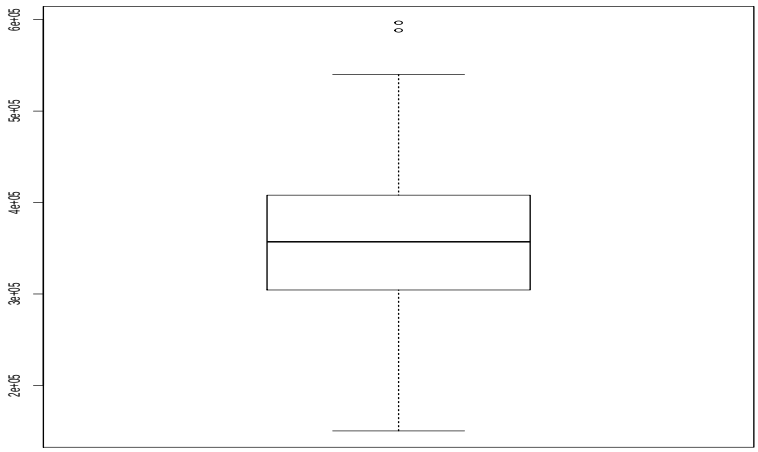


Fonte: Microdados ENEM 2016 - INEP

A figura 1 diz respeito ao gráfico de dispersão, dois a dois. Nota-se uma relação linear positiva entre as covariáveis e a variável resposta, quando observadas individualmente, embora, mais uma vez, não podemos generalizar o conceito para um modelo de regressão múltiplo. Uma transformação logarítmica foi feita sob a variável Ciências da Natureza para torná-la mais linear. É possível também detectar a presença de alguns outliers.

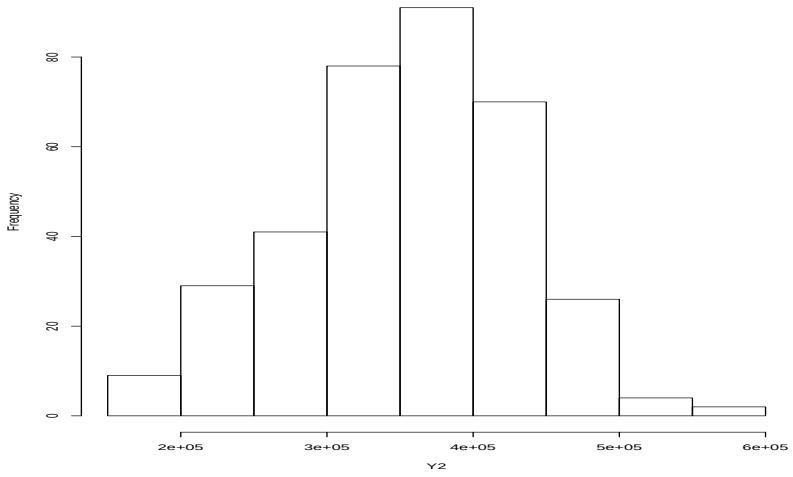
2.3 Normalidade da variável resposta

Figura 2 – Boxplot das observações de Ciências Humanas ao quadrado



Fonte: Microdados ENEM 2016 - INEP

Figura 3 – Histograma das observações de Ciências Humanas ao quadrado



Fonte: Microdados ENEM 2016 - INEP

As figuras 2 e 3 foram obtidas através de uma transformação Box Cox na variável Y, tendo como parâmetro igual a 2 (ver figura 4). Visualmente, não há indícios da violação de normalidade para a variável resposta, dado que não consta assimetria. Mais uma vez, é possível identificar a presença de alguns outliers.

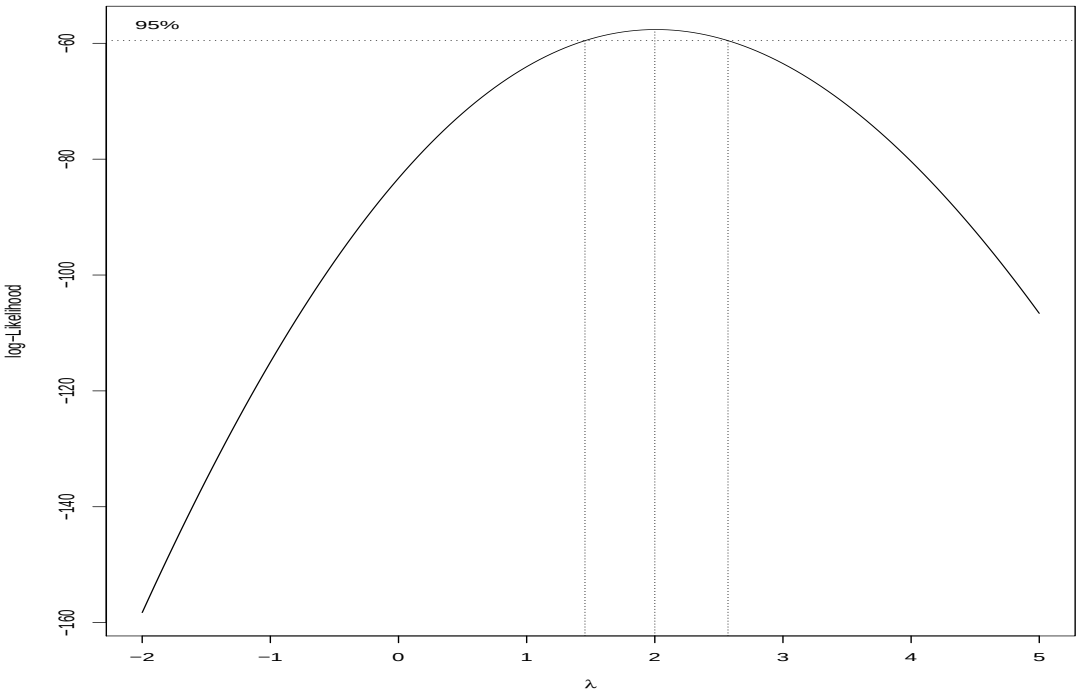
Tabela 2 – Teste de Shapiro Wilk para normalidade da variável resposta

W: 0.99349	p-value = 0.1362

Fonte: Microdados ENEM 2016 - INEP

Por meio de um Teste de Hipótese onde o β_0 é igual a normalidade das observações e através da tabela 2, verificamos por meio do Teste de Shapiro-Wilk que não rejeitamos a hipótese nula aos níveis usuais, ou seja, não temos evidências para rejeitar a normalidade dos dados.

Figura 4 – Gráfico da Função de Verossimilhança Perfilada do λ

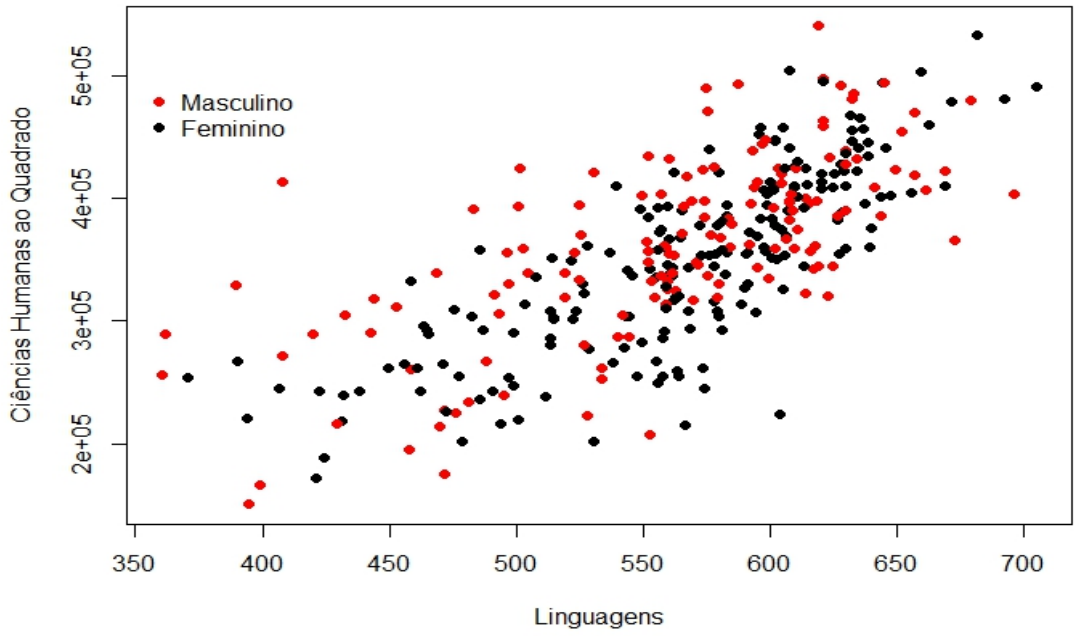


Fonte: Microdados ENEM 2016 - INEP

2.4 Variável de Interação

A suspeita que a covariável sexo exerce influência na relação entre Ciências Humanas e Linguagens é confirmadas visualmente através da Figura 5, sendo, portanto, justificável a inclusão de uma interação entre o Sexo e Linguagens. Observe que os maiores recortes de notas são dominados pelos setores femininos.

Figura 5 – Relação entre Línguas e Ciências Humanas, desagregada por sexo



Fonte: Microdados ENEM 2016 - INEP

3 Metodologia

3.1 Medidas

Para o estudo, foram utilizados os microdados do ENEM 2016 fornecidos pelo INEP (2017). Neles, são disponibilizadas as provas, os gabaritos, as informações sobre os itens, e as notas e o questionário respondido pelos inscritos no Enem. Ao todo, a base continha 20 variáveis de Dados do Participantes (Sexo, Idade, etc), 50 variáveis advindas do questionário socioeconômico e 5 variáveis contínuas provenientes das notas em cada grande área do conhecimento mensurados pelo ENEM (Matemática, Linguagens, Ciências da Natureza, Ciências Humanas e Redação). Foram excluídos da base aqueles que deixaram a redação em branco ou não fizeram alguma das quatro provas objetivas.

3.2 Amostragem

Foi retirada uma amostra de tamanho 350 de INEP (2017) através de Amostragem Aleatória Simples sem Reposição, recortada para o estado do Rio de Janeiro.

3.3 Análise Estatística

Matrizes de coeficientes de correlação de Pearson e de dispersão foram criadas para verificar a relação entre as variáveis contínuas. Após isso, foi utilizado o método de seleção backward por AIC com a nota em Ciências Humanas como variável dependendente e todas as outras como variáveis independentes para encontrar os fatores que influenciavam a nota. O software R foi utilizado para a análise com o nível de significância estabelecido em $\alpha = 0.05$.

3.4 Modelo de Regressão

Regressão linear múltipla é um mecanismo estatístico para o entendimento da relação entre uma variável dependente e mais de uma variável independente, ela pode ser escrita da forma: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon$

Onde ϵ é um vetor de erros com distribuição normal multivariada com vetor de médias nulo e matriz de variâncias e covariâncias, e cada β_i representa um parâmetro para a respectiva variável X_i.

4 Resultados

4.1 Pontos Influentes

A primeira análise será feita sob os pontos influentes, através da *Influence Measures*. O critério será: se um ponto for detectado em pelos menos dois dos testes: *DFFIT*, *Covariance Ratio* ou *Leverage* ele será removido do modelo. A Tabela 3 apresenta todos os pontos que foram detectados.

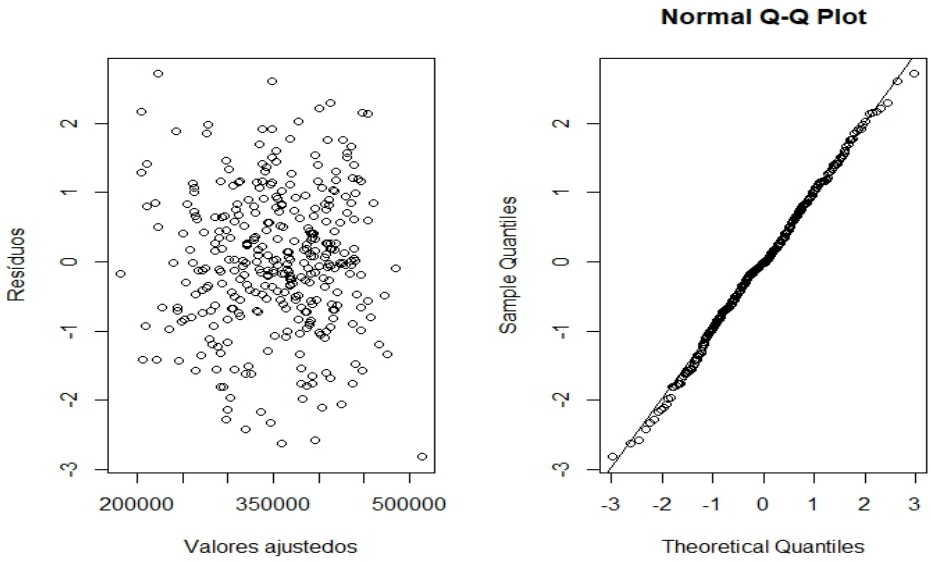
Tabela 3 – Tabela de Pontos Influentes

Resíduo	*DFFIT*	*Cov.R*	*Leverage*
56738	-0.53_*	0.92_*	
149455		1.07_*	0.08_*
196931	-0.50_*	0.77_*	
199043	-0.54_*	0.81_*	
346280		1.09_*	0.08_*
463040		1.09_*	0.08_*

Fonte: Microdados ENEM 2016 - INEP

4.2 Normalidade dos erros

Figura 6 – Gráfico de Resíduos e QQPlot



Fonte: Microdados ENEM 2016 - INEP

Segue da Figura 6 que, aparentemente, não há violação da suposição dos Erros normalmente distribuídos, uma vez que os resíduos se comportam aleatoriamente no *scatterplot* e distribuem-se sob os quantis teóricos da normal, visto no QQplot.

Tabela 4 – Teste de Shapiro Wilk para normalidade dos resíduos

W: 0.9965 p-value = 0.6565

Fonte: Microdados ENEM 2016 - INEP

Assim como anteriormente, é efetuado um Teste de Hipótese onde β_0 é a Hipótese de Normalidade contra a Não Normalidade das Observações. A tabela 4 confirma a intuição da figura 6, através do p-valor, não rejeitaremos a hipótese nula aos níveis usuais, ou seja, não há evidências para afirmar que os resíduos não são normalmente distribuídos, portanto, a suposição da normalidade dos erros não é violada.

4.3 Multicolinearidade

Tabela 5 – *Variance Inflation Factors*

Natureza	**Linguas**	**Redação**	**Tv Assinatura**	**Ensino Privado**	**Sexo:Línguas**
1.85	1.76	1.67	1.04	1.01	1.12

Fonte: Microdados ENEM 2016 - INEP

A literatura sugere atenção conforme o *VIF* se aproxima de 5, no modelo aqui sugerido, todas os valores são abaixo de 2, o que pressupõe a não multicolinearidade, como indicado na tabela 5. Em outras palavras, as covariáveis propostas na regressão não são combinações lineares de si próprias, a grosso modo, elas não explicam o mesmo fenômeno, o que as tornaria redundantes.

4.4 Heterocedasticidade

Tabela 6 – *Studentized Breusch-Pagan test*

BP = 3.1942 p-value = 0.7841

Fonte: Microdados ENEM 2016 - INEP

A tabela 6 indica o teste de Hipótese de Breusch-Pagan para heterocedasticidade, onde a Hipótese Nula é a de que a Variância dos Erros é constante contra a Variância dos erros não é constante. De acordo com o p-valor obtido, não rejeitaremos a hipótese nula aos níveis usuais, ou seja, o modelo proposto não viola a suposição de variância constante.

4.5 Modelo Final

Tabela 7 – Modelo de Regressão para Ciências Humanas

Coeficientes	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
Intercepto	-1.109e+06	1.056e+05	-10.501	0.0000
log(Natureza)	1.832e+05	1.906e+04	9.613	0.0000
Linguas	4.202e+02	4.452e+01	9.437	0.0000
Redação	1.004e+02	1.979e+01	5.073	0.0000
Televisão	1.041e+04	4.917e+03	2.117	0.0350
Particular	-2.736e+03	1.349e+03	-2.027	0.0434
Sexo:Linguagem	1.852e+01	8.237e+00	2.248	0.0252
Residual standard error	40470 on 337 degrees of freedom			
Multiple R-squared	0.7046	Adjusted R-Squared	0.6993	
F-Statistics	134 on 6 and 336 DF	p-value	<2.2e-16	

Fonte: Microdados ENEM 2016 - INEP

Dado a equação genérica:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 ln(X_1) + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6$$

A Tabela 7 sumariza, de forma categórica, a validade do modelo proposto neste artigo, passado as etapas estabelecidas até aqui, todos os coeficientes propostos são significantes ao nível usual de $\alpha = 0.05$ e portanto, o modelo de regressão múltiplo também ser-lo-á. O R^2 de aproximadamente 0.7 é considerada pela literatura como um bom número, embora isoladamente devemos adotar precauções ao utilizá-lo. Para fins de orientação ao leitor, a variável binária Televisão tem o Não como categoria de orientação e o Sexo tem o Feminino como categoria de Referência.

5 Aplicação

No âmbito escolar, pode haver interesse das redes privadas do Ensino Médio em analisar as disciplinas que mais afetam o desempenho do aluno nas matérias de Ciência Humanas. A partir disso, modificar o modo de aprendizagem em busca de melhores resultados poderia ser fundamental para alavancar as condições da instituição. Além disso, também pode ser exequível a construção de intervalos de predição para cada aluno, em particular. Já no âmbito estadual, é relevante entender o comportamento das variáveis não relacionadas aos dados da prova objetiva. Tendo conhecimento dessas variáveis, torna-se possível redirecionar os recursos a serem aplicados em políticas públicas educacionais.

A validação de um modelo linear múltiplo não hierárquico com aplicações em base de dados reais abre os horizontes para os estudos e pesquisas no âmbito educacional, à luz das dificuldades que permeiam o tratamento das bases oriundas da área citada, dado a pouca aplicação de modelagem estatística encontrada na literatura no que tange ao ensino pedagógico.

O modelo aqui proposto tem por finalidade de aplicação refletir sobre o atual cenário de ensino tradicional fluminense, principalmente a matriz curricular do ensino médio que por décadas segmentou as áreas do conhecimento, tornando-as incomunicáveis entre si e ademais, ignorando fatores socioeconômicos como determinante no desempenho da proficiência. Através do Modelo de Regressão Linear aqui proposto, vê-se que é possível transformar a nota de Ciências Humanas como combinação de outras ciências e fatores socioeconômicos, abrindo margem para modelagem das demais áreas do conhecimento.

6 Conclusão

O modelo proposto é capaz de explicar 70% da variabilidade nas notas observadas de Ciências Humanas. As outras notas demonstraram papel fundamental na predição, sendo três delas significativas ao nível considerado. Um dos resultados esperados foi o impacto que a vontade em querer ingressar na educação superior privada pode causar na nota em estudo, pois quanto maior for a motivação em ingressar, menor será o valor esperado da nota. Outra variável interessante refere-se a interação entre o sexo e a nota em Línguas. baseado nela, ao considerarmos as demais variáveis constantes, as médias das notas poderiam ser traçadas por retas paralelas para cada sexo, em função da nota em Línguas. Portanto, o estudo mostra que as Ciências Humanas, por meio do Exame Nacional do Ensino Médio, podem ser explicadas de maneira multifacetada, através de outras ciências, fatores socioeconômicos e fatores culturais captados pelo questionário do mesmo, a partir do qual, novas investigações podem ser feitas.

Modelo final:

$$Y = -1109177 + 183196ln(X_1) + 420X_2 + 100X_3 + 10407X_4 - 2735X_5 + 8.5X_6$$

Referências

FIDALGO, António. O consumo de informação. interesse e curiosidade. *Internet. Disponível em http://www. bocc. ubi. pt/pag/fidalgo-antonio-interesse-curiosidade-informacao. pdf (consultado em 13 de fevereiro de 2012)*, 1996. Citado na página 2.

GARCIA, Maurício. Análise retrospectiva dos resultados do exame nacional de desempenho de estudantes (enade) de 2007 a 2012. *Revista Cientefico*, v. 14, n. 27, p. 11–22, 2014. Citado na página 2.

GUERRA, Pedro Calais *et al.* Estimativa de demanda potencial de matrículas em ensino superior usando dados públicos e múltiplos modelos de regressão. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO-SBC. *Symposium on Knowledge Discovery, Mining and Learning, 2th*. [S.l.], 2014. Citado na página 2.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANISIO TEIXEIRA. *Microdados do Enem 2016*. Brasília: Inep, 2017. Disponível em: < http://portal.inep.gov.br/web/guest/microdados>. Acesso em: 01 nov. 2017. Citado na página 7.

NETER, John; WASSERMAN, William; KUTNER, Michael H. Applied linear regression models. Irwin Homewood, IL, 1989. Citado na página 1.

NEVES, Lúcia Maria Wanderley. *O empresariamento da educação: novos contornos do ensino superior no Brasil dos anos 1990*. [S.l.]: Coletivo de Estudos sobre Política Educacional, 2002. Citado na página 2.

OLIVEIRA, Mara Janaina Gomes de. Um perfil de concluintes do curso superior com base no enade (2005). Universidade Estadual Paulista (UNESP), 2011. Citado na página 2.

Anexos

ANEXO A – Códigos Utilizados

```
## PROCESSO DE GARIMPAGEM DO BANCO DE DADOS

enem = enemAMOSTRA

attach(enem)

y = NU_NOTA_CH
x1 = NU_NOTA_CN
x2 = NU_NOTA_MT
x3 = NU_NOTA_LC
x4 = NU_NOTA_REDACAO

model = lm(y~x1+x2+x3)
summary(model)

setwd("C:/Users/201511282-00/Documents/Modelos
Lineares/Trabalho de Modelos Lineares/Microdados_enem_2016/DADOS")
load("enemAMOSTRA.Rda")
#load("enemLIMPO.Rda")
#load("enemRJ.Rda")
#install.packages("sqldf")
#library(sqldf)
#enem=read.csv.sql("microdados_enem_2016.csv",
sql = "select * from file where CO_UF_RESIDENCIA = 33 ",sep=";")
#save(enem,file="enemRJ.Rda")

#2
enem=enem[-which(enem$NU_IDADE==0),]

#7
enem=enem[(enem$TP_PRESENCA_CN==1 & enem$TP_PRES
ENCA_LC==1),]

#8
enem=enem[(enem$TP_ST_CONCLUSAO==2 & enem$TP_E
```

```
SCOLA==3),]

#9
enem=enem[!enem$TP_STATUS_REDACAO==4,]

#10
enem=enem[(enem$NU_NOTA_CN>0 & enem$NU_NOTA_CH>0
& enem$NU_NOTA_LC>0 & enem$NU_NOTA_MT>0),]

################################################
##############################################

#11

enem=enem[,c('NU_IDADE','TP_SEXO','TP_ESTADO_CIVIL
','TP_COR_RACA','TP_NACIONALIDADE',
             'TP_ENSINO','TP_LOCALIZACAO_ESC','TP_
             SIT_FUNC_ESC','IN_CERTIFICADO','NU_NOTA_CN'
             ,'NU_NOTA_CH','NU_NOTA_LC','NU_N
             OTA_MT','TP_LINGUA','NU_NOTA_REDACAO',
             'Q001','Q002','Q003','Q004','Q005',
             'Q006','Q007','Q008','Q009','Q010','Q011',
             'Q012','Q013','Q014','Q015','Q016',
             'Q017','Q018','Q019','Q020','Q021','Q022',
             'Q023','Q024','Q025','Q026','Q034','Q035'
             ,'Q036','Q037','Q038','Q039','Q040',
             'Q042','Q043','Q044','Q045','Q046','Q047'
             ,'Q048','Q049','Q050')]

#12
for(i in 2:9)
  enem[,i]=factor(enem[,i])
enem$TP_LINGUA=factor(enem$TP_LINGUA)
for(i in 16:19)
  enem[,i]=factor(enem[,i])
for(i in 21:41)
  enem[,i]=factor(enem[,i])
for(i in 49:57)
  enem[,i]=factor(enem[,i])
```

```
enem$IN_CERTIFICADO=as.numeric(as.character
(enem$IN_CERTIFICADO))
enem$TP_LINGUA=as.numeric(as.character(enem$TP_LINGUA))

#####################################################

#13
library(TeachingSampling)
N=dim(enem)[1]
n=350
amostra=S.SI(N,n)
enemAMOSTRA=enem[amostra,]

#####################################################
save(enem, file="enemAMOSTRA.Rda")
#####################################################

#Análise Exploratória
#table()
enem$TP_NACIONALIDADE=NULL
enem$TP_LOCALIZACAO_ESC=NULL
enem$TP_ENSINO=NULL
enem$TP_ESTADO_CIVIL=NULL
enem$TP_SIT_FUNC_ESC=NULL

enem$Q035=NULL
enem$Q043=NULL
enem$Q044=NULL
enem$Q046=NULL
enem$Q048=NULL

for(i in 11:14)
  enem[,i]=as.integer(enem[,i])
for(i in 16:35)
  enem[,i]=as.integer(enem[,i])

#Fazendo variáveis categóricas ficarem em
forma de número para correlação
```

```
enemcor=enem
enemcor$TP_SEXO=as.numeric(enem$TP_SEXO)
enemcor$TP_COR_RACA=as.numeric(enem$TP_COR_RACA)
enemcor$Q026=as.numeric(enem$Q026)
enemcor$Q042=as.numeric(enem$Q042)
enemcor$Q045=as.numeric(enem$Q045)
enemcor$Q047=as.numeric(enem$Q047)
enemcor$Q049=as.numeric(enem$Q049)
enemcor$Q050=as.numeric(enem$Q050)
correlacoes=cor(enemcor)

a <- which(correlacoes>0.7 & correlacoes<1,arr.ind=T)
#b <- which(correlacoes<(-0.3),arr.ind=T)
#Correlações mais altas em modulo entre NU_NOTA_CH e
NU_NOTA_CN,
#e entre NU_NOTA_CH e NU_NOTA_LC

##################################################
library(car)
library(MASS)
library(lmtest)
library(lawstat)
library(nortest)
library(leaps)
library(stats)
##################################################


####### código do modelo final

#Análise Exploratória

attach(enem)

Humanas = NU_NOTA_CH
Natureza = NU_NOTA_CN
Linguas = NU_NOTA_LC
Redação = NU_NOTA_REDACAO
```

```
#Scatter
pairs(cbind(Humanas,log_Natureza
,Linguas,Redação),pch=19)

#Boxcox
bcx=boxcox(regbackward,lambda=seq
(-2,5,0.05))
transf.bcx = bcx$x[which.max(bcx$y)]
CH_QUADRADO = NU_NOTA_CH^2
Y2 =NU_NOTA_CH^2

#Normalidade
boxplot(CH_QUADRADO,title="")
hist(Y2, main = "")
shapiro.test(Y2)

#Interação
plot(enem_sem$NU_NOTA_LC,CH_QUADRADO_sem,
pch=19,col=enem_sem$TP_SEXO,xlab="Linguagens"
,ylab="Ciências Humanas ao Quadrado")
legend(350,5e+05,legend=c("Feminino","Masculino"
),col=c(2,1),text.col="black",lty=c(-1, -1),
pch=19,bty="n")


######Validação do Modelo

#Pontos Influentes
infmed=influence.measures(regbackward)
infmed
summary(infmed)
enemteste=enem
enemteste$num=1:350
pl=enemteste[c("56738","149455","196931"
,"199043","346280","463040"),]$num
enem_sem=enem[-c(50,164,165),]
enem_sem=enem_sem[-c(62,78,159),]

#Normalidade dos erros
```

```
par(mfrow=c(1,2))
plot(fitted.values(final),rstudent(final),xlab =
'Valores ajustedos' , ylab='Resíduos')
qqnorm(rstudent(final))
qqline(rstudent(final))
shapiro.test(final$residuals)


vif(final)

bptest(final)

#Modelo final
final=lm(formula = CH_QUADRADO_sem ~
NU_NOTA_LC:TP_SEXO + log_NOTA_CN_sem  + N
U_NOTA_LC
         + NU_NOTA_REDACAO  + Q021 + Q036 ,
         data = enem_sem)

summary(final)
```

ANEXO B – Questionário

DICIONÁRIO DE VARIÁVEIS - ENEM 2016					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DO PARTICIPANTE					
TP_SEXO	Sexo	M	Masculino	1	Alfanumérica
		F	Feminino		
DADOS DA PROVA OBJETIVA					
NU_NOTA_CN	Nota da prova de Ciências da Natureza			9	Numérica
NU_NOTA_CH	Nota da prova de Ciências Humanas			9	Numérica
NU_NOTA_LC	Nota da prova de Linguagens e Códigos			9	Numérica
NU_NOTA_REDACAO	Nota da prova de redação			9	Numérica
DADOS DO QUESTIONÁRIO SOCIOECONÔMICO					
Q021	Na sua residência tem TV por assinatura?	A	Não.	1	Alfanumérica
		B	Sim.		
Q036	Indique os motivos que levaram você a participar do ENEM: Ingressar na Educação Superior privada.	0	0 indica o fator menos relevante e 5 o fator mais relevante	1	Numérica
		1			
		2			
		3			
		4			
		5			