Impact of Family Income on ENEM test scores

Gabriel de Souza Oliva
Isabella de Melo Sousa

Research Question

"How does family income affect students performance on the Brazilian National High School Examination (ENEM)?"

Literature Review

Previous literatures show that there are some factors that have great influence over the students performance. One of the most observed is the familiar background, specially the parents education. Some related articles are mentioned below.

- *Family background, regional inequality and performance in the national high school exam (ENEM)*
- *Determinants of the performance in ENEM of high school graduates in the city of Viçosa - MG*
- *The influence of parental education and family income on the performance of ENEM candidates*
- *Impact of Household Income on Standardized Test Scores*

Literature Review

Other literatures have checked other social factors such as skin color or type of high school. We've also looked for articles concerning the American test (SAT).

- Factors that influence the performance of participants in the ENEM test: a study for residents of Ceará in 2015
- Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance

Data

Data Source

Most of the Brazilian Government data is available in a platform of open data. Each brazilian organ has a open data website. As ENEM is part of National Institute of Educational Studies and Research Anísio Teixeira, microdata regarding the exam may be accessed in its website.

Variables Selected for the study

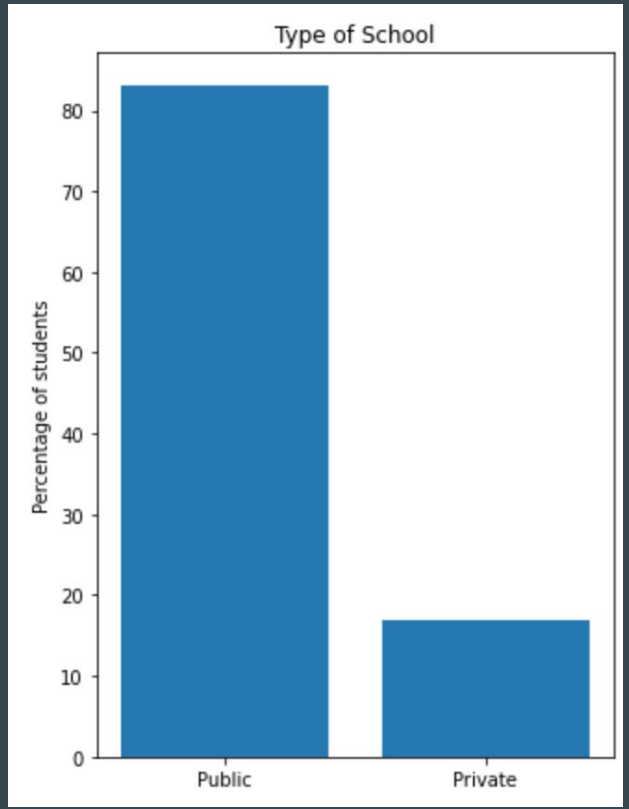
- Math test scores
- Income
- Type of school
- Students Region
- Color/race
- Father/mother formation
- Father/mother profession
- House Size
- Internet Access

Data Cleaning

- Initially 5 millions rows, after selecting the variables, filtering data and dropping Nan values, 1.2 millions rows.

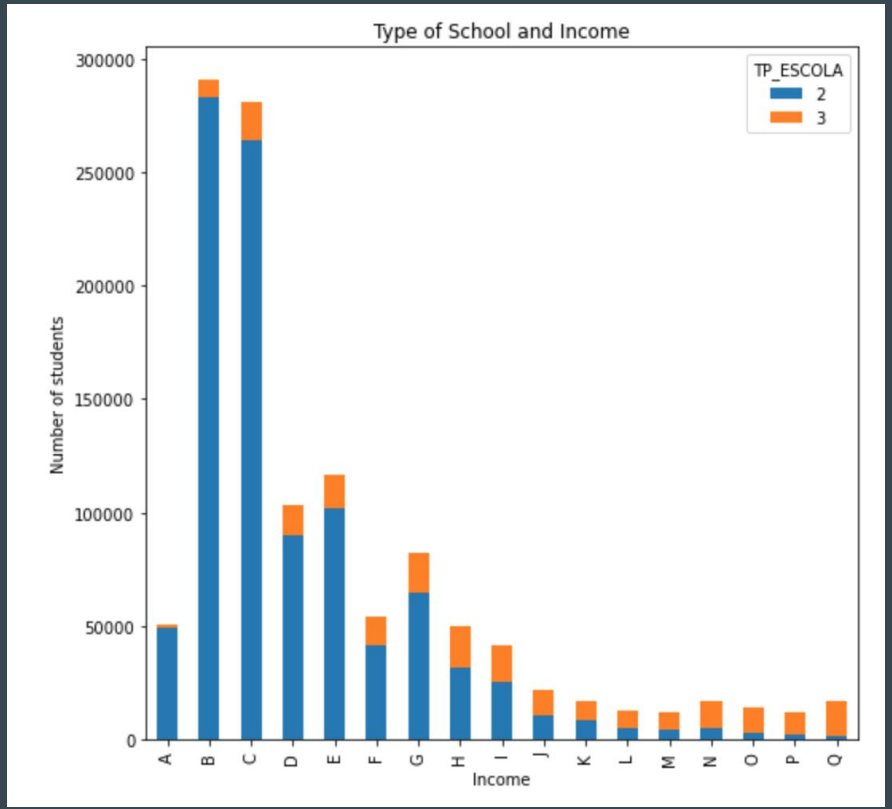
Descriptive Analysis

- We analysed the distribution of the variables into its categories using bar plots;



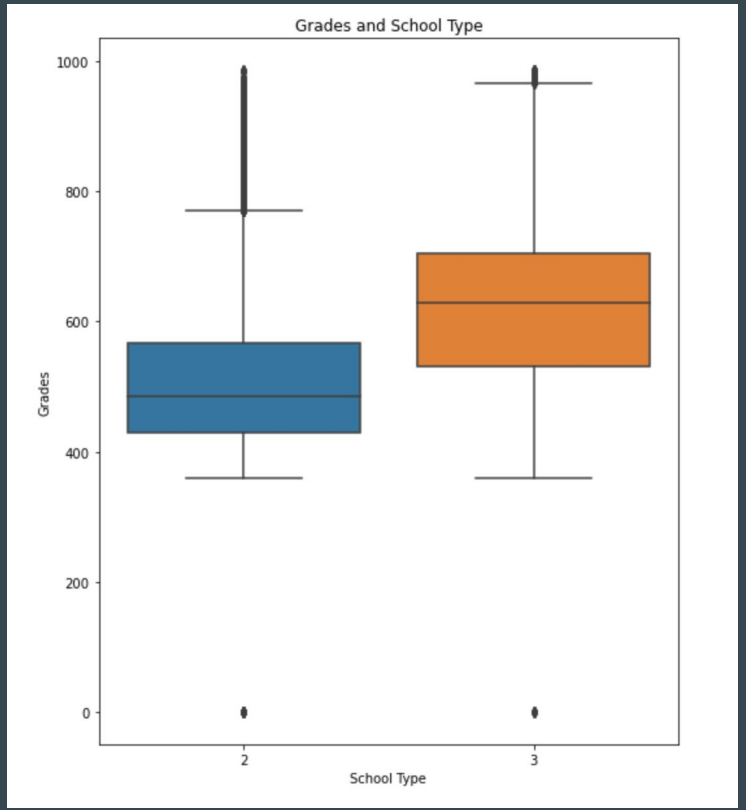
Descriptive Analysis

- We analysed the distribution of the variables categories within the income;

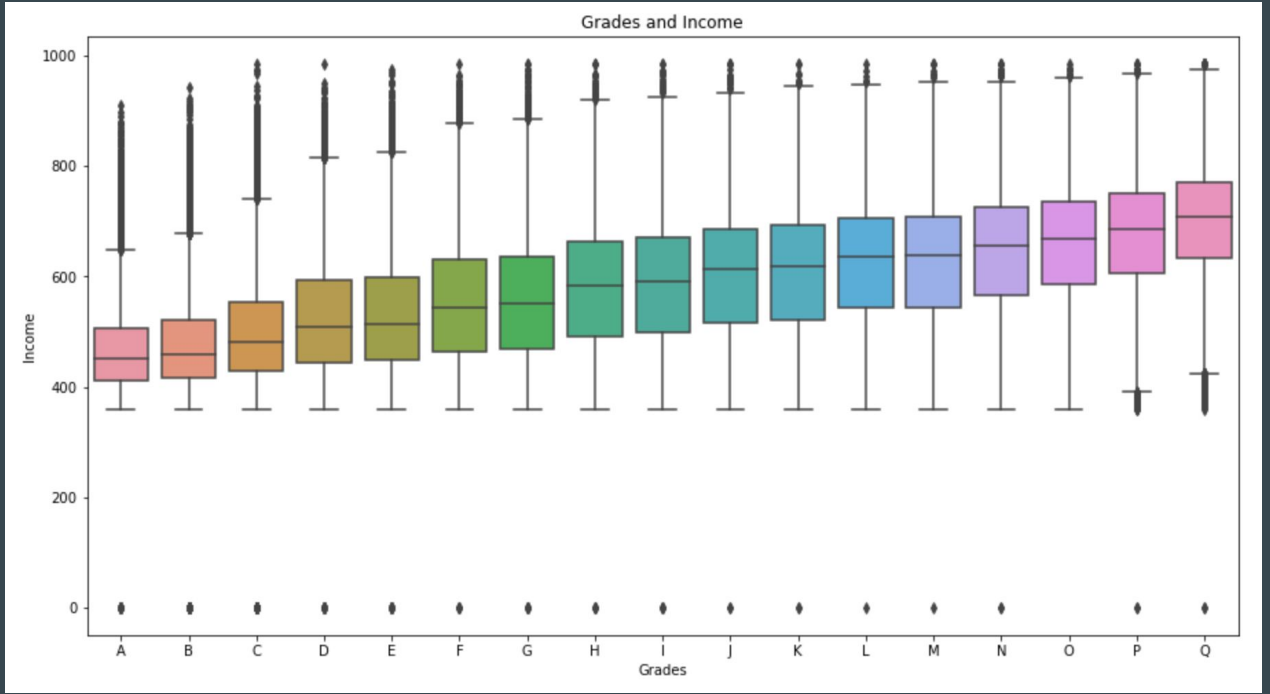


Descriptive Analysis

- By using box plots we could observe that the test scores mean increase with the income ; attendance in a private school ; the richness of the region; the most privileged race/color ; the parents level of education / the profession's renown; the size of the house and the presence of internet



Income


Grades and Income
Income
1000
800
600
400
200
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
Grades

Descriptive Analysis

- We used Cramér's V statistic to measure the association between two categorical variables, concluding that there no strong correlations, but some moderate ones specially between type of school and income, parents profession/formation and house size. Also, internet access and income; and mother and father professions

	TP_COR_RACA	TP_ESCOLA	Father Formation	Mother Formation	Father Profession	Mother Profession	Income	Internet	REGION	HOUSE
TP_COR_RACA	1.000000	0.226946	0.125181	0.112035	0.131315	0.122596	0.172413	0.209209	0.192789	0.126462
TP_ESCOLA	0.226946	1.000000	0.434276	0.399716	0.483039	0.447310	0.542185	0.198160	0.104049	0.438451
Father Formation	0.125181	0.434276	1.000000	0.334806	0.371897	0.251329	0.250031	0.299455	0.116470	0.169654
Mother Formation	0.112035	0.399716	0.334806	1.000000	0.241594	0.338806	0.222055	0.297337	0.101549	0.153960
Father Profession	0.131315	0.483039	0.371897	0.241594	1.000000	0.421263	0.340969	0.350573	0.147475	0.233476
Mother Profession	0.122596	0.447310	0.251329	0.338806	0.421263	1.000000	0.318574	0.351145	0.154046	0.213590
Income	0.172413	0.542185	0.250031	0.222055	0.340969	0.318574	1.000000	0.435576	0.216334	0.259067
Internet	0.209209	0.198160	0.299455	0.297337	0.350573	0.351145	0.435576	1.000000	0.305758	0.328646
REGION	0.192789	0.104049	0.116470	0.101549	0.147475	0.154046	0.216334	0.305758	1.000000	0.110793
HOUSE	0.126462	0.438451	0.169654	0.153960	0.233476	0.213590	0.259067	0.328646	0.110793	1.000000

Empirical Strategy

Base Model

$$Test\ Scores = \beta_0 + \beta_1 . Family\ Income$$

Variables

- Y: test scores
- Variable of interest: family income
- Control variables:
  - Type of school
  - Students Region
  - Color/race
  - Father/mother formation
  - Father/mother profession
  - Internet Access
  - House Size
    - number of bedrooms + numbers of bathrooms

Regression Results

Income and Socioeconomic Classes

- Class E(1) : R$ 0 - 1996 (0 to 2 minimum wages)
- Class D(2) : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class C(3) : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class B(4) : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class A(5) : R$ 19960,01 (+ than 20 minimum wages)

Income

Legend

- *Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted*
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

OLS Regression Results

Dep. Variable:	NU_NOTA_MT	**R-squared:**	0.212
Model:	OLS	**Adj. R-squared:**	0.212
Method:	Least Squares	**F-statistic:**	8.043e+04
Date:	Sun, 30 May 2021	**Prob (F-statistic):**	0.00
Time:	14:14:03	**Log-Likelihood:**	-7.1396e+06
No. Observations:	1193326	**AIC:**	1.428e+07
Df Residuals:	1193321	**BIC:**	1.428e+07
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	490.3702		4351.298	0.000	490.149	490.591
C(Income)[T.2]	51.8326		233.857	0.000	51.398	52.267
C(Income)[T.3]	105.6913		393.385	0.000	105.165	106.218
C(Income)[T.4]	165.5238		349.220	0.000	164.595	166.453
C(Income)[T.5]	204.7005		275.376	0.000	203.244	206.157

51,83

53,86

59,83

39,18

Omnibus:	36164.081	**Durbin-Watson:**	1.947
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39550.671
Skew:	0.445	**Prob(JB):**	0.00
Kurtosis:	3.061	**Cond. No.**	8.77

Income/TypeSchool

Legend

- *Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted*
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

	Model 1	Model 2
Intercept	490.37***	551.45***
	(0.11)	(0.28)
C(Income)[T.2]	51.83***	43.88***
	(0.22)	(0.22)
C(Income)[T.3]	105.69***	80.29***
	(0.27)	(0.28)
C(Income)[T.4]	165.52***	119.37***
	(0.47)	(0.50)
C(Income)[T.5]	204.70***	149.26***
	(0.74)	(0.76)
Public		-64.61***
		(0.27)
R-squared	0.21	0.25
	0.21	0.25
No. observations	1193326	1193326

Income/House

Legend

- *Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted*
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

	Model 1	Model 2
Intercept	490.37***	458.01***
	(0.11)	(0.29)
C(Income)[T.2]	51.83***	44.68***
	(0.22)	(0.23)
C(Income)[T.3]	105.69***	90.32***
	(0.27)	(0.29)
C(Income)[T.4]	165.52***	139.68***
	(0.47)	(0.52)
C(Income)[T.5]	204.70***	169.52***
	(0.74)	(0.79)
House		9.82***
		(0.08)
R-squared	0.21	0.22

Income/ Father and Mother Formation

Legend

Income

- ***Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted***
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

Level of Study

- Group 0: Never studied.
- Group 1: Did not complete the 4th grade/5th grade of elementary school.
- Group 2: Completed the 4th grade/5th grade, but didn't complete the 8th grade/9th grade of elementary school.
- Group 3: Completed 8th grade/9th grade of elementary school, but did not complete high school.
- Group 4: Completed High School, but did not complete college.
- Group 5: Completed college, but did not complete graduate school.
- Group 6: Completed graduate school.
- Group 7: Did not know.

	Model 1	Model 2
Intercept	490.37***	463.04***
	(0.11)	(0.46)
C(Income)[T.2]	51.83***	40.89***
	(0.22)	(0.23)
C(Income)[T.3]	105.69***	82.52***
	(0.27)	(0.29)
C(Income)[T.4]	165.52***	129.77***
	(0.47)	(0.51)
C(Income)[T.5]	204.70***	164.64***
	(0.74)	(0.78)
C(FatherFormation)[T.1]		14.11**
		(0.50)
C(FatherFormation)[T.2]		25.05***
		(0.51)
C(FatherFormation)[T.3]		30.26***
		(0.52)
C(FatherFormation)[T.4]		43.02***
		(0.49)
C(FatherFormation)[T.5]		71.85***
		(0.57)
C(FatherFormation)[T.6]		75.09***
		(0.63)
C(FatherFormation)[T.7]		16.90***
		(0.54)
R-squared	0.21	0.24

	Model 1	Model 2
Intercept	490.37***	456.62***
	(0.11)	(0.61)
C(Income)[T.2]	51.83***	40.04***
	(0.22)	(0.23)
C(Income)[T.3]	105.69***	83.99***
	(0.27)	(0.29)
C(Income)[T.4]	165.52***	136.82***
	(0.47)	(0.50)
C(Income)[T.5]	204.70***	173.19***
	(0.74)	(0.76)
C(MotherFormation)[T.1]		14.81***
		(0.66)
C(MotherFormation)[T.2]		24.52***
		(0.65)
C(MotherFormation)[T.3]		30.59***
		(0.65)
C(MotherFormation)[T.4]		45.35***
		(0.63)
C(MotherFormation)[T.5]		70.54***
		(0.67)
C(MotherFormation)[T.6]		69.59***
		(0.69)
C(MotherFormation)[T.7]		13.14**
		(0.77)
R-squared	0.21	0.24

	Model 1	Model 2
Intercept	490.37***	451.33***
	(0.11)	(0.61)
C(Income)[T.2]	51.83***	35.72***
	(0.22)	(0.23)
C(Income)[T.3]	105.69***	72.90***
	(0.27)	(0.30)
C(Income)[T.4]	165.52***	117.56***
	(0.47)	(0.52)
C(Income)[T.5]	204.70***	151.17***
	(0.74)	(0.78)
C(FatherFormation)[T.1]		7.78***
		(0.55)
C(FatherFormation)[T.2]		14.86***
		(0.55)
C(FatherFormation)[T.3]		17.67***
		(0.56)
C(FatherFormation)[T.4]		26.26***
		(0.54)
C(FatherFormation)[T.5]		50.03***
		(0.61)
C(FatherFormation)[T.6]		52.00***
		(0.68)
C(FatherFormation)[T.7]		8.41***
		(0.59)
C(MotherFormation)[T.1]		10.14***
		(0.69)
C(MotherFormation)[T.2]		15.98***
		(0.70)
C(MotherFormation)[T.3]		19.81***
		(0.70)
C(MotherFormation)[T.4]		30.44***
		(0.68)
C(MotherFormation)[T.5]		50.78***
		(0.73)
C(MotherFormation)[T.6]		49.59***
		(0.75)
C(MotherFormation)[T.7]		7.77***
		(0.84)
R-squared	0.21	0.25
No. observations	1193326	1193326

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Regression with All Variables

	Model 1	Internet	Public	Region	ColorRace	House	FatherFormation	MotherFormation	FatherProfession	MotherProfession
Intercept	490.37***	470.86***	532.47***	519.33***	531.75***	513.05***	492.11***	480.31***	479.29***	478.94***
	(0.11)	(0.19)	(0.32)	(0.41)	(0.43)	(0.51)	(0.67)	(0.80)	(0.80)	(0.80)
C(Income)[T.2]	51.83***	44.41***	37.67***	34.39***	32.71***	30.00***	25.33***	22.24***	20.40***	20.07***
	(0.22)	(0.23)	(0.22)	(0.23)	(0.23)	(0.23)	(0.24)	(0.24)	(0.24)	(0.24)
C(Income)[T.3]	105.69***	97.00***	73.55***	70.00***	66.92***	60.75***	50.36***	44.87***	40.95***	40.45***
	(0.27)	(0.28)	(0.29)	(0.29)	(0.29)	(0.31)	(0.32)	(0.32)	(0.33)	(0.33)
C(Income)[T.4]	165.52***	156.50***	113.02***	109.28***	104.87***	94.07***	78.13***	71.85***	67.33***	66.38***
	(0.47)	(0.48)	(0.50)	(0.50)	(0.50)	(0.53)	(0.55)	(0.55)	(0.57)	(0.57)
C(Income)[T.5]	204.70***	195.65***	143.19***	139.24***	133.75***	118.39***	101.29***	94.78***	90.65***	88.90***
	(0.74)	(0.74)	(0.76)	(0.76)	(0.76)	(0.79)	(0.81)	(0.82)	(0.84)	(0.85)
R-squared	0.21	0.22	0.26	0.26	0.27	0.27	0.28	0.29	0.29	0.29

Robustness Check

- Interactions between variables except income
- Interactions between the income and other variables

Interaction between variables except Income

- TypeSchool/Internet
- House/Region
- House/Mother Formation
- ColorRace/TypeSchool
- ColorRace/Region

=> The interactions variables barely had effect on the income coefficients.

Interaction Income/ColorRace

Class

- ***Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted***
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

Color and Race

- ***Color/Race 1: White - omitted***
- ***Color/Race*** 2: Black
- ***Color/Race*** 3: Brown
- ***Color/Race*** 4: Yellow
- ***Color/Race*** 5: Indigenous

Not White

Model 3:

- Class3: 103.52 - 14.54*(NotWhite)
  - if NotWhite = 0 => 103.52
  - if NotWhite = 1 => 88.98
- Class4: 163.37 - 14.50*(NotWhite)

	Model 1	Model 2	Model 3
Intercept	489.60***	508.90***	506.58***
	(0.50)	(0.78)	(0.93)
C(Income)[T.2]	53.10***	47.11***	49.28***
	(0.99)	(1.00)	(1.51)
C(Income)[T.3]	106.27***	97.02***	103.52***
	(1.20)	(1.22)	(1.65)
C(Income)[T.4]	170.38***	158.12***	163.37***
	(2.09)	(2.11)	(2.52)
C(Income)[T.5]	205.63***	190.72***	193.04***
	(3.41)	(3.41)	(3.77)
C(NotWhite)[T.1]		-27.10***	-23.83***
		(0.84)	(1.10)
C(Income)[T.2]:C(NotWhite)[T.1]			-2.96
			(2.02)
C(Income)[T.3]:C(NotWhite)[T.1]			-14.54***
			(2.49)
C(Income)[T.4]:C(NotWhite)[T.1]			-14.50***
			(4.72)
C(Income)[T.5]:C(NotWhite)[T.1]			-3.25
			(9.13)
R-squared	0.22	0.23	0.23
R-squared Adj.	0.22	0.23	0.23
No. observations	59666	59666	59666

Interaction Income/TypeSchool

Legend

Class

- ***Class 1 : R$ 0 - 1996 (0 to 2 minimum wages) - omitted***
- Class 2 : R$ 1996,01 - 3992 (2 to 4 minimum wages)
- Class 3 : R$ 3992,01 - 9980 (4 to 10 minimum wages)
- Class 4 : R$ 9980,01 - 19960 (10 to 20 minimum wages)
- Class 5 : R$ 19960,01 (+ than 20 minimum wages)

Model 3:

- Class5: 147.09 + 26.60*(Public)
  - if Public= 0 => 147.09
  - if Public= 1 => 174,69

	Model 1	Model 2	Model 3
Intercept	489.60***	549.14***	551.14***
	(0.50)	(1.24)	(2.10)
C(Income)[T.2]	53.10***	45.57***	42.68***
	(0.99)	(0.98)	(2.90)
C(Income)[T.3]	106.27***	81.18***	79.59***
	(1.20)	(1.27)	(2.63)
C(Income)[T.4]	170.38***	125.75***	122.53***
	(2.09)	(2.22)	(3.09)
C(Income)[T.5]	205.63***	151.00***	147.09***
	(3.41)	(3.49)	(4.02)
C(Public)[T.1]		-63.00***	-65.11***
		(1.20)	(2.16)
C(Income)[T.2]:C(Public)[T.1]			3.19
			(3.08)
C(Income)[T.3]:C(Public)[T.1]			1.36
			(3.04)
C(Income)[T.4]:C(Public)[T.1]			7.28
			(5.15)
C(Income)[T.5]:C(Public)[T.1]			26.60**
			(12.48)
R-squared	0.22	0.25	0.25
R-squared Adj.	0.22	0.25	0.25
No. observations	59666	59666	59666

Interaction Income/House

	Model 1	Model 2	Model 3
Intercept	489.60***	456.12***	459.18***
	(0.50)	(1.28)	(1.75)
C(Income)[T.2]	53.10***	45.66***	39.80***
	(0.99)	(1.02)	(3.54)
C(Income)[T.3]	106.27***	90.34***	76.16***
	(1.20)	(1.32)	(4.40)
C(Income)[T.4]	170.38***	144.06***	145.87***
	(2.09)	(2.27)	(8.98)
C(Income)[T.5]	205.63***	169.23***	209.78***
	(3.41)	(3.62)	(18.59)
House		10.15***	9.22***
		(0.36)	(0.51)
C(Income)[T.2]:House			1.62*
			(0.89)
C(Income)[T.3]:House			3.21***
			(0.95)
C(Income)[T.4]:House			0.10
			(1.54)
C(Income)[T.5]:House			-5.41**
			(2.69)
R-squared	0.22	0.23	0.23
R-squared Adj.	0.22	0.23	0.23
No. observations	59666	59666	59666

Interaction Income/Father Formation

	Model 1	Model 2	Model 3
Intercept	489.60***	482.03***	481.86***
	(0.50)	(0.55)	(0.59)
C(Income)[T.2]	53.10***	44.24***	46.41***
	(0.99)	(1.00)	(1.37)
C(Income)[T.3]	106.27***	84.28***	81.60***
	(1.20)	(1.30)	(2.34)
C(Income)[T.4]	170.38***	135.40***	129.26***
	(2.09)	(2.25)	(6.93)
C(Income)[T.5]	205.63***	166.24***	143.19***
	(3.41)	(3.52)	(13.51)
C(FatherFormation)[T.2]		22.11***	23.29***
		(0.91)	(1.17)
C(FatherFormation)[T.3]		53.70***	50.80***
		(1.36)	(2.51)
C(Income)[T.2]:C(FatherFormation)[T.2]			-2.99
			(2.16)
C(Income)[T.3]:C(FatherFormation)[T.2]			0.41
			(3.08)
C(Income)[T.4]:C(FatherFormation)[T.2]			-7.55
			(8.25)
C(Income)[T.5]:C(FatherFormation)[T.2]			19.87
			(16.84)
C(Income)[T.2]:C(FatherFormation)[T.3]			-6.17*
			(3.59)
C(Income)[T.3]:C(FatherFormation)[T.3]			8.61**
			(3.79)
C(Income)[T.4]:C(FatherFormation)[T.3]			13.73*
			(7.72)
C(Income)[T.5]:C(FatherFormation)[T.3]			28.10**
			(14.20)
R-squared	0.22	0.24	0.24
R-squared Adj.	0.22	0.24	0.24
No. observations	59666	59666	59666

Conclusions

Some Conclusions

- ↑ Income ↑ Test scores
- The binary variable internet basically does not influence the income and the test scores
- The variable TypeSchool is the one with the biggest influence in the income and in the test scores.
    - Students in public schools are penalized with -64.61 points
- People that have color/race different from white are penalized in the test scores
    - Indigenous have the biggest penalization with -41 points
- Students from the region Southeast and South have higher scores
- ↑ Parents Formation ↑ Income ↑ Test scores
- ↑ Parents Profession ↑ Income ↑ Test scores
    - Mother and Father have very similar results
- The r-squared from the all the variations of models is low, indicating that we do not have a good model for predictions
- The interactions variables (without income) barely had effect on the income coefficients, which means that our study variable is not biased by the interactions between variables, only by themselves.
- The interaction variables (with income) showed more significative effects in the influence of the income in the test scores, but did not affect predictability

Final Model

Y = Grades

Variables: Income, Type of School, House, Mother Formation

Final Model

Dep. Variable:	NU_NOTA_MT	**R-squared:**	0.272
Model:	OLS	**Adj. R-squared:**	0.272
Method:	Least Squares	**F-statistic:**	1714.
Date:	Thu, 24 Jun 2021	**Prob (F-statistic):**	0.00
Time:	02:53:34	**Log-Likelihood:**	-3.5472e+05
No. Observations:	59666	**AIC:**	7.095e+05
Df Residuals:	59652	**BIC:**	7.096e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	489.0066	3.158	154.862	0.000	482.817	495.196
C(Income)[T.2]	32.8821	1.020	32.237	0.000	30.883	34.881
C(Income)[T.3]	58.6997	1.397	42.018	0.000	55.962	61.438
C(Income)[T.4]	94.6772	2.383	39.728	0.000	90.006	99.348
C(Income)[T.5]	113.2481	3.665	30.898	0.000	106.064	120.432
C(Public)[T.1]	-52.4187	1.219	-42.989	0.000	-54.809	-50.029
C(MotherFormation)[T.1]	13.6199	2.886	4.719	0.000	7.963	19.277
C(MotherFormation)[T.2]	23.7520	2.870	8.276	0.000	18.127	29.377
C(MotherFormation)[T.3]	28.1563	2.842	9.906	0.000	22.586	33.727
C(MotherFormation)[T.4]	39.1893	2.748	14.259	0.000	33.803	44.576
C(MotherFormation)[T.5]	56.9279	2.948	19.308	0.000	51.149	62.707
C(MotherFormation)[T.6]	58.2754	3.059	19.053	0.000	52.281	64.270
C(MotherFormation)[T.7]	10.2443	3.381	3.030	0.002	3.617	16.871
House	6.2067	0.352	17.619	0.000	5.516	6.897

Omnibus:	1492.687	**Durbin-Watson:**	1.998
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1606.411
Skew:	0.400	**Prob(JB):**	0.00
Kurtosis:	3.080	**Cond. No.**	85.8

Potential Limitations

- Inaccurate answers to the questionnaire;
- The variable House can be a bad control variable in some cases;
- The variable Region englobes too much states and therefore can be too generalistic to draw more detail conclusions;
- Some public schools have quality equal or even superior than private schools, which can prejudice the general interpretation (such as military schools);

Thanks for watching!

Questions?