Exercício Econométrico

Vinícius Ventura

Contents

Link para os scripts das 3 questões (STATA, R e Python):

Questão	Link
R Markdown	R Markdown
Questão 1 - STATA	Do-file
Questão 2 - R	R Script
Questão 3 - Python (Google Colab)	Google Colab

Questão 1 - Stata

a) Regressão

Importando os dados

```
* Definindo o diretório onde o stata irá trabalhar *
cd "C:\Users\Robson Ventura\OneDrive\Trabalhos\Econometria - Stata\Exe - AB2 -
Econometria 2\Resultados"

*Importando os dados*

use http://www.stata.com/data/jwooldridge/eacsap/fertil2.dta

save exereconometrico.dta, replace
```

Primeiramente, tanto para a questão 1 quanto 3.1 foram renomeadas algumas variáveis, resultando na relação a seguir:

Original	Usado	Descrição
children	cria	Número de filhos vivos
age	age	Idade da mãe em anos
agesq	agesq	Idade da mãe em anos ao quadrado

Original	Usado	Descrição
educ	educ	Anos de educação
electric	elet	Possui eletricidade
urban	urb	Vive em área urbana

Código de renomeação no STATA:

```
rename children cria
rename electric elet
rename urban urb

label variable cria "Número de filhos vivos"
label variable age "Idade da mãe em anos"
label variable agesq "Idade da mãe em anos ao quadrado"
label variable educ "Anos de educação"
label variable elet "Possui eletricidade"
label variable urb "Vive em área urbana"
```

Código de renomeação no Python:

```
df = df.rename(columns={'children': 'cria' ,'electric': 'elet' ,'urban': 'urb'})
```

Regressão

```
reg cria age agesq
```

Quadro 1 - Regressão com todas as variáveis do modelo

Source	SS	df	MS	Número.de.obs	X4358
Model	12333	5.000	2466	Prob>F	0
Residual	9176	4352	2.109	Adj R-squared	0,573
Total	21510	4357	4.937	Root MSE	1.452
cria	Coeficientes	Desvio Padrão	t	P>|t|	Intervalo de Confiança 0,95
age	0,341	0,017	20.650	0	0,309 a 0,373
agesq	-0,003	0	-10.090	0	-0,003 a -0,002
educ	-0,075	0,006	-11.950	0	-0,088 a -0,063
elet	-0,31	0,069	-4.490	0	-0,445 a -0,175
urb	-0,2	0,047	-4.300	0	-0,291 a -0,109
cons	-4.223	0,24	-17.580	0	-4.693 a -3.752

A partir dos resultados da regressão com 4358 observações, primeiramente, com base no R^2 ajustado, podemos afirmar que o modelo explica cerca de 57% da variável dependente *cria*. Já vendo a significância estatística das variáveis, vemos primeiramente que, de acordo com o valor-p do teste F, rejeitamos a hipótese nula que os coeficientes conjuntamente são estatisticamente iguais a 0. Analisando similarmente os testes t e valores-p de cada variável, notamos que, *a priori*, todos os coeficientes são estatisticamente significantes ao nível de significância de 1%, tendo valores dos testes t em módulo acima de 2 e valores-p abaixo de 0,01.

Analisando os coeficientes, podemos afirmar que as variáveis de educação, eletricidade, e residência em área urbana afetam negativamente a quantidade de filhos vivos, podendo ser explicada teoricamente com o seguinte raciocínio: quanto mais bem estruturada for a moradia da mãe, estiver em área urbana e tiver mais

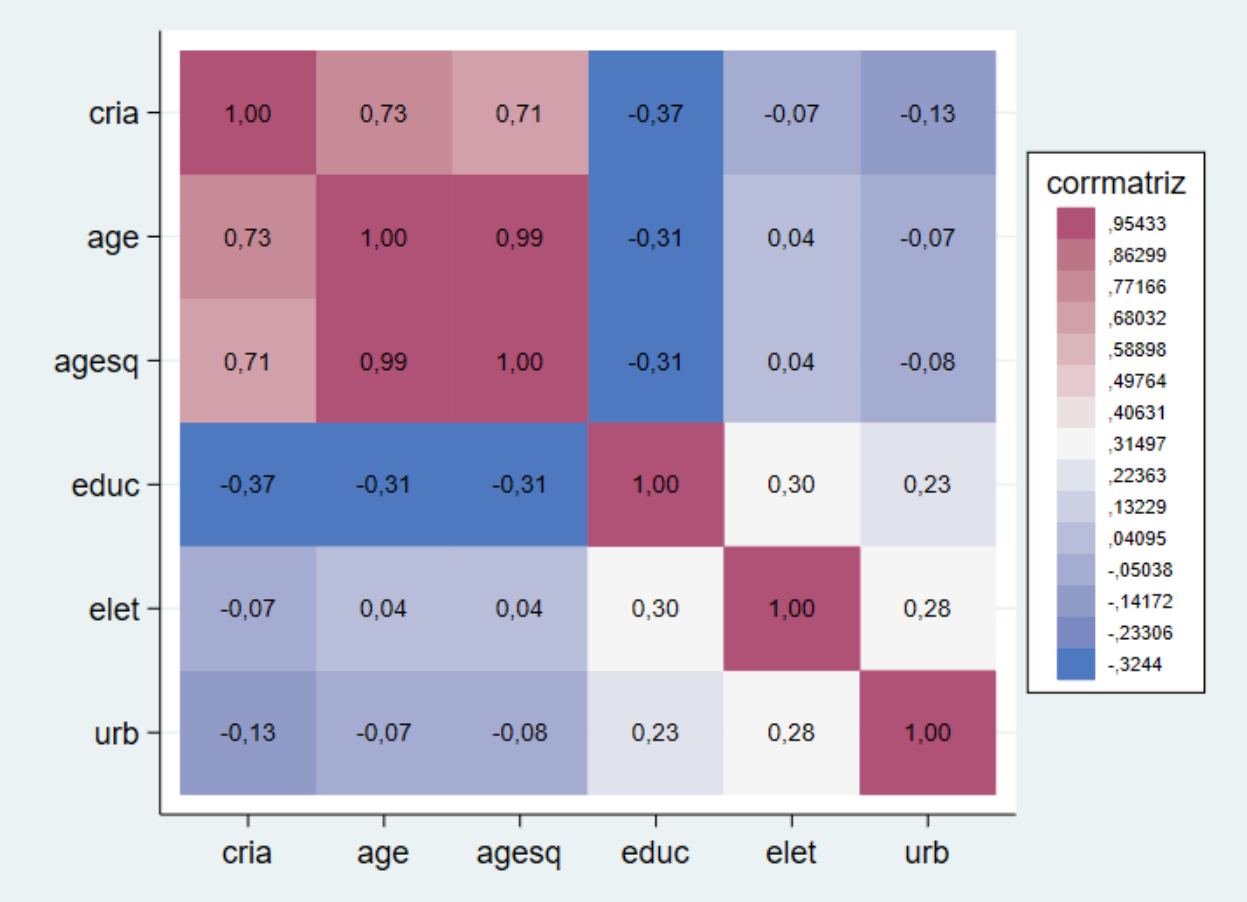
anos de educação, menos filhos em média ela terá. Isto *ceteris paribus*, pois, analisando por exemplo a idade da mãe, vemos uma relação em média positiva com a quantidade dos filhos mantendo constante as demais variáveis.

Ademais, vale dar destaque as variáveis *dummys*, as variáveis *elet* e *urb*, estas duas trazem análise categórica da variável *cria* com base em elementos de condições de vida e moradia, sendo, de acordo com os resultados, os coeficientes -0,31 e -0,2 respectivamente, portanto, ter eletricidade e viver em área urbana diminuem nestas respectivas proporções em média a quantidade de filhos vivos.

b) Matriz de Correlação e FIV

```
correlate cria age agesq educ elet urb
matrix corrmatriz = r(C)
heatplot corrmatriz, values(format(%4.2f)) color(hcl diverging, intensity(.7))
```

Gráfico de calor da matriz de correlação entre as variáveis



Pois bem, analisando o gráfico da matriz de correlação, podemos ver que as variáveis *age* e *agesq* possuem correlação positiva com a variáveis *cria*, como ressaltado anteriormente a partir dos resultados da regressão. Seguindo a lógica que quanto mais velha a mãe mais filhos em média ela tem mantendo tudo mais constante. Seguindo isto, com o comando *VIF* obtemos a análise do FIV.

```
VIF
```

Teste Fator de Inflação de Variância (FIV)

Variável	FIV	TOL
age	42.48	0.023538
agesq	42.38	0.023596
educ	1.26	0.791565
elet	1.19	0.842930
urb	1.12	0.895796
Média	17.69	0.515485

Como podemos ver, a média do FIV deu acima de 10, muito por causa das duas variáveis com multicolinearidade perfeita, portanto, foi estimado o modelo sem uma das duas, já que, relativamente, elas trazem interpretações semelhantes, assim, a variável escolhida para permanecer foi a *age* que representa a idade da mãe em anos. Rodando a regressão e o FIV novamente sem a variável *agesq* temos os seguintes resultados:

```
reg cria age educ elet urb
vif
```

Resolvendo Regressão sem a variável agesq

Source	SS	df	MS	Número.de.Obs	X4358
Model	12119,112	4.000	3029,778	Prob>F	0
Residual	9390,92	4.353	2,157	R-squared	0,563
Total	21510,032	4.357	4,936	Root MSE	1,468
Variáveis	Coeficientes	Desvio Padrão	t	P>|t|	Intervalo de Confiança 0,95
age	0,177	0,003	64,78	0	0,171 a 0,182
educ	-0,076	0,006	-11,87	0	-0,088 a -0,063
elet	-0,303	0,07	-4,35	0	-0,44 a -0,167
urb	-0,171	0,047	-3,65	0	-0,263 a -0,079
cons	-2	0,097	-20,7	0	-2,189 a -1,81

Fator de Inflação de Variância (FIV) - sem *agesq*

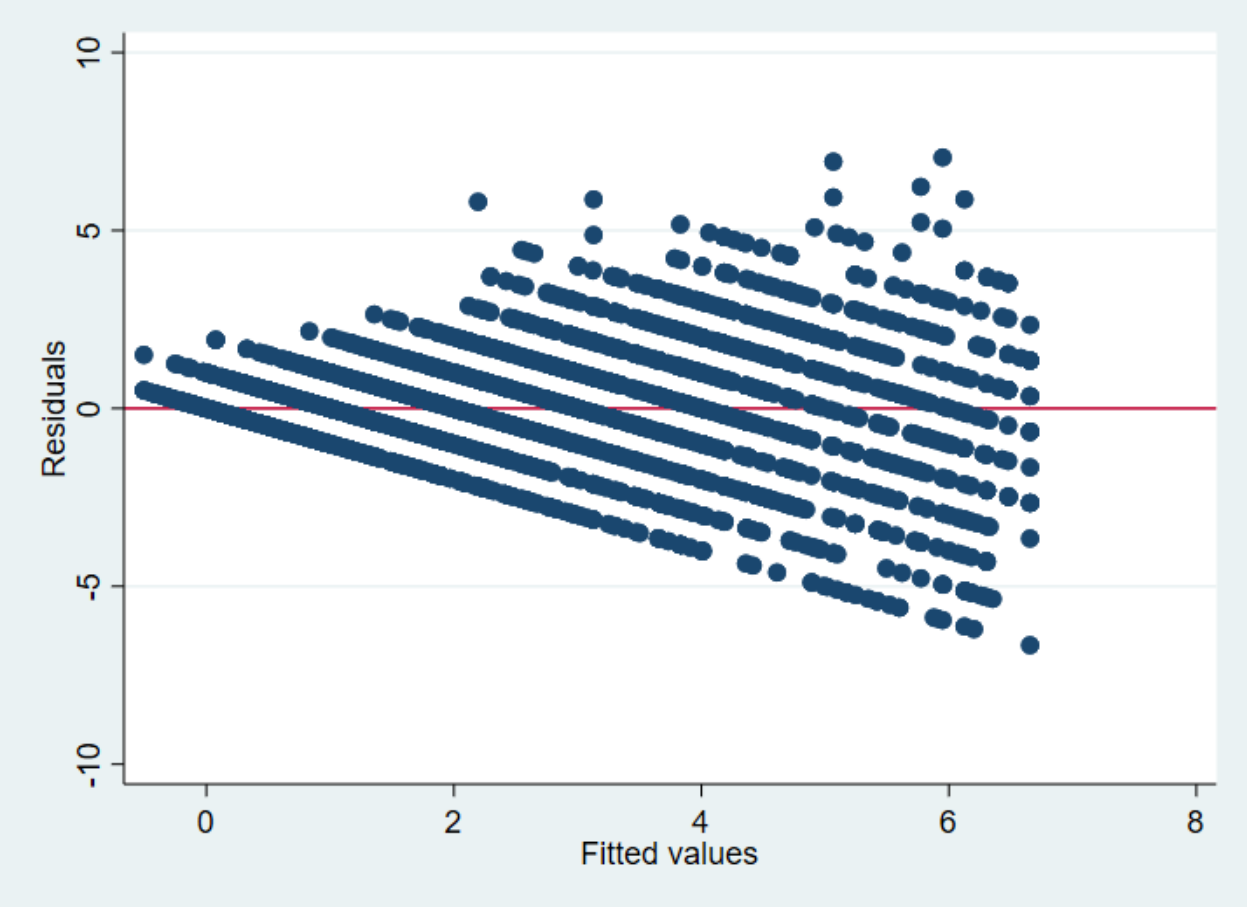
Variáveis	FIV	TOL
educ	1,26	0,79
elet	1,19	0,84
age	1,13	0,88
urb	1,11	0,9
Média	1,17	0,85

Pois bem, neste caso, é evidenciado uma diminuição no valor médio do FIV fazendo-nos aceitar a hipótese nula de não multicolinearidade.

c) Teste White e BPG para verificar a presença de heterocedasticidade

```
rvfplot, yline(0)
```

Verificação dos Resíduos Distribuição dos resíduos



Com base no gráfico anterior, é evidente que a variância dos erros não são constantes, aumentando sua dispersão ao longo do gráfico. Para irmos além da análise gráfica, segue os testes BP e White para heterocedasticidade.

```
hettest
```

Testes Teste de Breusch-Pagan

Análise	Resultado
chi2(4)	2140,57
Prob>chi2	0,00

```
imtest, white
```

Teste White

Fonte	chi2	df	p
Heteroskedasticity	1124,69	12.000	0,00

Com base nos resultados dos dois testes, rejeitamos a hipótese nula de homocedasticidade pelo baixo valor-p dos dois, ficando abaixo de 0,01. Deste modo, temos a evidências que temos problema de heterocedasticidade, ou seja, variância não constante dos resíduos.

Para correção do modelo, foi usado o comando *robust* referente a correção de White, a qual ajusta os desvios padrões com base na variância dos erros, trazendo estimativas mais precisas na presença de heterocedasticidade.

```
reg cria age educ elet urb, robust
```

Resolvendo Regressão com Robust para ajustamento com base nos erros

Variáveis	Coeficientes	Desvio.Padrão.Robust	t	P..t.	Intervalo.de.Confiança.95.
age	0,176	0,00342	51,65	0	0,169946 a 0,183356
educ	-0,075	0,0063825	-11,85	0	-0,08812 a -0,063098
elet	-0,303	0,064938	-4,67	0	-0,43068 a -0,176058
urb	-0,171	0,0458123	-3,74	0	-0,26112 a -0,08149
cons	-1,999	0,0949218	-21,07	0	-2,18575 a -1,81356

Questão 2 - R

a) Importando dados

a.1) Instalando e importando os pacotes usados

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("readxl")
install.packages("psych")
install.packages("openxlsx")
install.packages("gmodels")
install.packages("tseries")
library(tseries)
library(openxlsx)
library(dplyr)
library(psych)
library(ggplot2)
library(readxl)
library(gmodels)
```

```
df <- read_excel("C:/Users/Robson Ventura/OneDrive/Trabalhos/Econometria - R/Atividade AB2/data_prova.x
View(df)
```

a.2) Importando os dados

ano	uf	codmun	munic	analf	pobre	rpc	agua	lixo	esgag_inad	poprur	poptot	popurb
2000	11	1100015	ALTA FLORESTA D'OESTE	15.84	35.59	371.15	57.13	70.33	27.52	14192	26533	12341
2000	11	1100023	ARIQUEMES	11.77	21.55	530.87	73.54	90.61	21.14	19385	74503	55118
2000	11	1100031	CABIXI	17.50	36.47	342.46	58.61	76.57	27.36	4846	7518	2672
2000	11	1100049	CACOAL	12.14	25.25	456.63	78.60	90.22	14.28	22170	73568	51398
2000	11	1100056	CEREJEIRAS	13.90	33.57	511.47	73.98	70.04	24.82	3361	18207	14846

Filtrando por Alagoas

```
al <- subset(df, df$uf == 27 & df$ano == 2010)
View(al)
```

ano	uf	codmun	munic	analf	pobre	rpc	agua	lixo	esgag_inad	poprur	poptot	popurb
2010	27	2700102	ÁGUA BRANCA	30.92	52.87	213.00	61.78	97.94	36.29	14276	19377	5101
2010	27	2700201	ANADIA	36.16	47.59	239.53	61.62	99.15	11.06	8475	17424	8949
2010	27	2700300	ARAPIRACA	22.45	26.65	423.28	94.68	97.38	12.38	32525	214006	181481
2010	27	2700409	ATALAIA	33.57	42.73	237.33	82.49	95.31	18.19	21865	44322	22457
2010	27	2700508	BARRA DE SANTO ANTÔNIO	27.94	44.24	247.81	91.16	93.66	2.74	988	14230	13242

b) Análise Descritiva

b.1) Análise geral

```
summary(al[c('analf','pobre', 'poptot', 'popurb', 'rpc', 'esgag_inad')])
```

	analf	pobre	poptot	popurb	rpc	esgag_inad
	Min. :11.86	Min. :15.57	Min. : 2866	Min. : 1171	Min. :151.6	Min. : 1.330
	1st Qu.:29.43	1st Qu.:40.60	1st Qu.: 8444	1st Qu.: 4068	1st Qu.:204.6	1st Qu.: 8.102
	Median :32.34	Median :46.94	Median : 17077	Median : 7054	Median :233.4	Median :15.010
	Mean :32.56	Mean :45.94	Mean : 30593	Mean : 22528	Mean :251.2	Mean :20.249
	3rd Qu.:37.32	3rd Qu.:51.48	3rd Qu.: 25352	3rd Qu.: 14424	3rd Qu.:269.0	3rd Qu.:31.320
	Max. :43.89	Max. :67.57	Max. :932748	Max. :932129	Max. :792.5	Max. :60.860

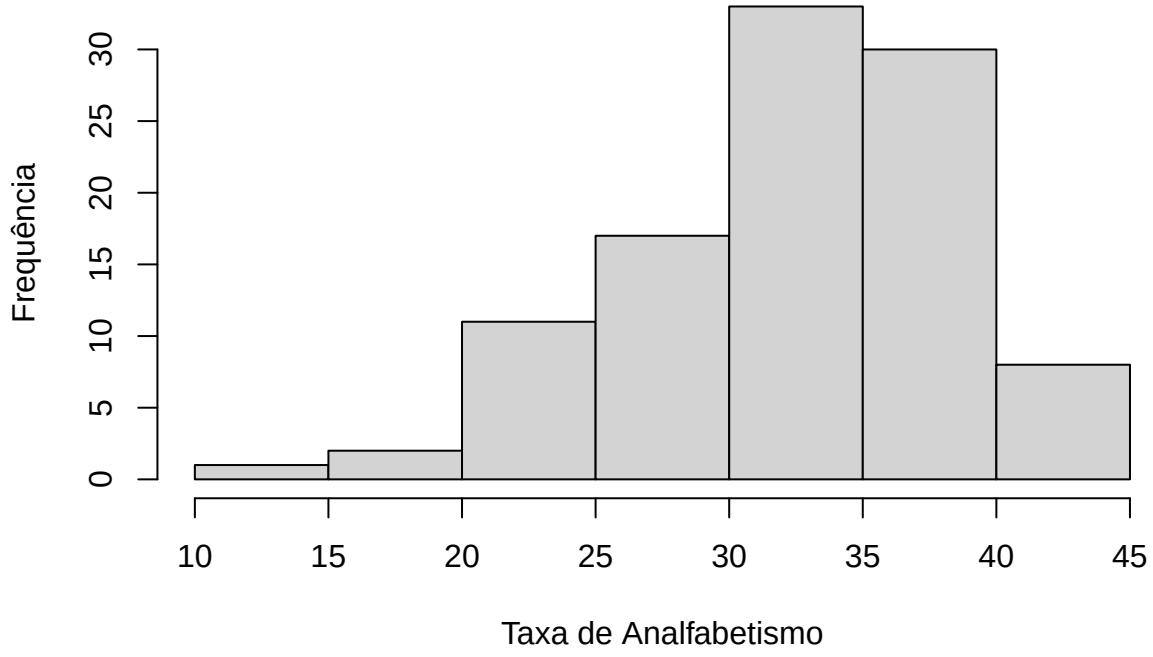
Analisando as variáveis em geral, podemos ver que o estado possui médias, tanto das variáveis de analfabetismo quanto de pobreza, bastante altas, visto que, em média, a taxa de pobreza dos municípios do estado

é 46,94%. Tal dado tem bastante correlação com a renda per capita da população, a qual em média é R$ 251,20, tendo o mínimo em R$ 151,60, referente ao município de Olho d'Água Grande, valor substancialmente baixo. Já analisando as demais variáveis, vemos que em média, 20,24% da população alagoana não possui tratamento de esgoto adequado, isto sendo que cerca de 25% dos municípios de Alagoas possuem uma porcentagem de esgoto inadequado igual ou inferior a 8,102.

De acordo com a tabela, a variável *analf* pode ser explicada através da relação pobreza e renda per capita. De modo que, quanto maior for a renda per capita menor será a taxa de pobreza e analfabetismo, respectivamente. Em Alagoas a taxa média de pobreza corresponde a 45,94%, ou seja, cerca da metade dos municípios apresentam condições de vulnerabilidade social, logo esse dado é um reflexo do índice de analfabetismo que alcança uma média 32,56% para o ano em análise, e renda média de R$251,20 por pessoa.Podemos visualizar como a variável se distribui de acordo com sua frequência com o gráfico a seguir.

Gráfico de frequência de analfabetismo - Alagoas em 2010

```
hist(al$analf,
     main = "",
     xlab = "Taxa de Analfabetismo",
     ylab = "Frequência")
```



Conforme demonstrado, é evidente uma maior concentração a partir da faixa dos 30%.

c) Criação de variável categórica para analfabetismo

```
cortes <- quantile(al$analf, probs = c(0, 0.05, 0.5, 1))
cortes
```

```
al$sitanalf <- cut(al$analf,
                   cortes, labels=c("bom", "regular", "ruim"),
                   right=FALSE)
head(al,5)
```

ano	uf	codmun	munic	analf	pobre	rpc	agua	lixo	esgag_inad	popurb	poptot	popurb	sitanalf
2010	27	2700102	ÁGUA BRANCA	30.92	52.87	213.00	61.78	97.94	36.29	14276	19377	5101	regular
2010	27	2700201	ANADIA	36.16	47.59	239.53	61.62	99.15	11.06	8475	17424	8949	ruim
2010	27	2700300	ARAPIRACA	22.45	26.65	423.28	94.68	97.38	12.38	32525	214006	181481	bom
2010	27	2700409	ATALAIA	33.57	42.73	237.33	82.49	95.31	18.19	21865	44322	22457	ruim
2010	27	2700508	BARRA DE SANTO ANTÔNIO	27.94	44.24	247.81	91.16	93.66	2.74	988	14230	13242	regular

d) Análise da sitanalf

d.1) Criação da variável filtrada para os municípios com a situação "bom"

```
analfbom <- subset(al, al$sitanalf==’bom’)
head(analfbom,5)
```

ano	uf	codmun	munic	analf	pobre	rpc	agua	lixo	esgag_inad	popurb	poptot	popurb	sitanalf
2010	27	2700300	ARAPIRACA	22.45	26.65	423.28	94.68	97.38	12.38	32525	214006	181481	bom
2010	27	2704302	MACEIÓ	11.86	15.57	792.54	80.17	97.74	2.32	619	932748	932129	bom
2010	27	2704708	MARECHAL DEODORO	21.90	32.01	431.43	97.25	94.74	4.89	2585	45977	43392	bom
2010	27	2707701	RIO LARGO	18.28	24.26	369.11	77.58	88.92	5.50	12534	68481	55947	bom
2010	27	2708600	SÃO MIGUEL DOS CAMPOS	21.28	28.11	360.82	96.16	99.31	1.33	2011	54577	52566	bom

d.2) Criação da variável filtrada para os municípios com a situação "ruim"

```
analfruim <- subset(al, al$sitanalf==’ruim’)
head(analfruim,5)
```

ano	uf	codmun	munic	analf	pobre	rpc	agua	lixo	esgag_inad	popurb	poptot	popurb	sitanalf
2010	27	2700201	ANADIA	36.16	47.59	239.53	61.62	99.15	11.06	8475	17424	8949	ruim
2010	27	2700409	ATALAIA	33.57	42.73	237.33	82.49	95.31	18.19	21865	44322	22457	ruim
2010	27	2700805	BELÉM	33.33	37.05	309.50	62.16	92.54	30.81	2679	4551	1872	ruim
2010	27	2700904	BELO MONTE	38.11	60.98	187.77	40.13	99.16	17.30	5859	7030	1171	ruim
2010	27	2701100	BRANQUINHA	41.53	40.52	221.93	81.68	97.83	17.15	3910	10583	6673	ruim

Analisando a variável categórica, vemos que 51 municípios são categorizados como ruim, 45 como regular e apenas 6 como bom. A tabela nos da os top 5 municípios em cada categoria, porém é bem relevante a pouca quantidade na categoria bom, ressaltando ainda mais a situação educacional do estado.

e) Criando variável de faixa populacional

```
cortespop <- c(-Inf, 5000, 20000, 50000, 100000, Inf)
al$fxpop <- cut(al$poptot, cortespop, labels=c(1,2,3,4,5), right=FALSE)
head(al[, c("poptot", "fxpop")],5)
```

poptot	fxpop
19377	2
17424	2
214006	5
44322	3
14230	2

f) Tabulação cruzada entre sitanalf e fxpop

```
cross <- xtabs(~al$sitanalf + al$fxpop)
popcross <- prop.table(cross, margin=1)*100
round(popcross,2)
```

	1	2	3	4	5
bom	0.00	16.67	16.67	33.33	33.33
regular	2.22	55.56	31.11	11.11	0.00
ruim	10.00	58.00	32.00	0.00	0.00

Já ao analisar as duas variáveis categóricas, vemos que, como já visto nas últimas demonstrações, estão inseridos, em maior quantidade, no grupo categorizado com taxa de analfabetismo "bom", os grupos 4 e 5 de faixa populacional, referentes a quantidades mais altas. Por outro lado, podemos ver que no grupo "ruim" temos maior parte no grupo de faixa populacional 2 e nenhuma porcentagem com os grupos 4 e 5. Esta análise facilita a visualização de que os municípios mais populosos possuem melhores condições educacionais.

g) Estatística descritiva da variável pobre por faixa populacional

```
describeBy(al$pobre, al$fxpop, digits = 2, skew=FALSE, ranges=FALSE)
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars n  mean    sd   se
## X1    1 6 45.33 10.98 4.48
## ------------------------------------------------------------
## group: 2
##    vars  n  mean   sd   se
## X1    1 56 48.28 9.06 1.21
## ------------------------------------------------------------
## group: 3
##    vars  n  mean   sd  se
## X1    1 31 46.24 7.81 1.4
## ------------------------------------------------------------
```

```
## group: 4
##    vars n  mean   sd   se
## X1    1 7 33.46 6.43 2.43
## ------------------------------------------------------------
## group: 5
##    vars n  mean   sd   se
## X1    1 2 21.11 7.83 5.54
```

A partir dessa análise, vemos que a média da taxa de pobreza do grupo com maiores populações é bem abaixo dos demais, dando destaque ao grupo 5 com, em média 21,11% e o grupo 1 com média de 45,33%, valores bastantes descrepantes. Além de que, vemos com os dados de desvio padrão que os dados também são mais dispersos do que dos grupos mais altos, porém, podemos verificar isto com melhor detalhes com coeficiente de variação.

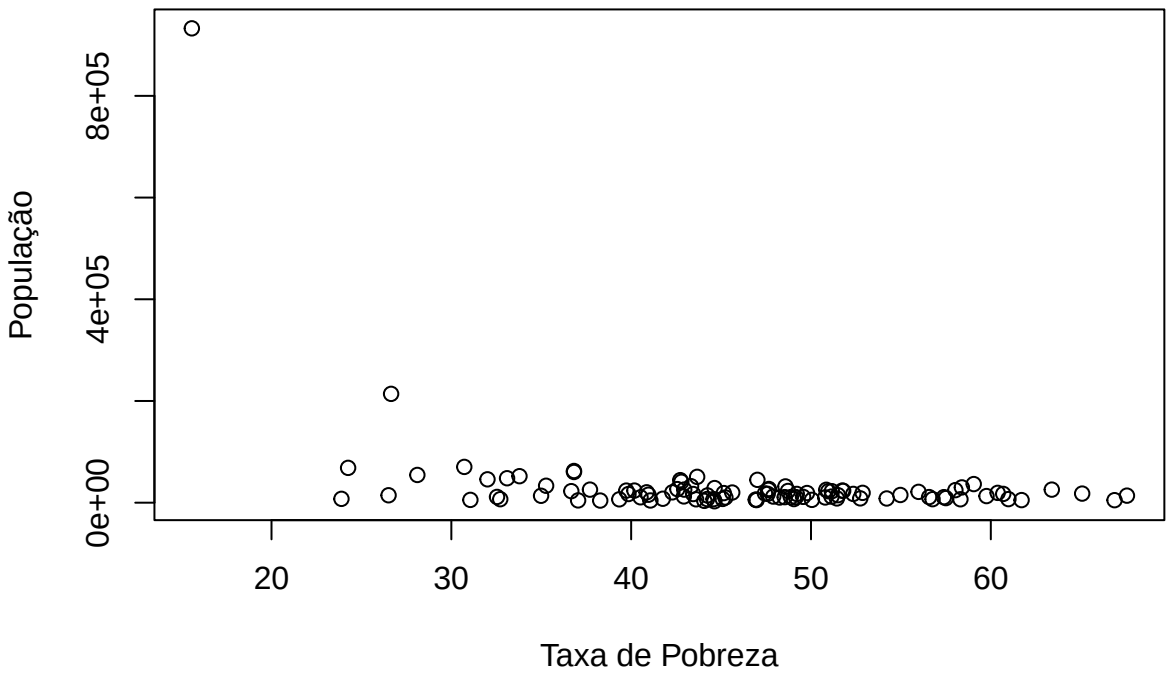
```
cvpobre <- al %>%
  group_by(fxpop) %>%
  CV <- sd(pobre)/mean(pobre)*100
  summarise(CV)
  cvpobre
cvpobre
```

fxpop	sd(pobre)/mean(pobre) * 100
1	24.22411
2	18.76594
3	16.88239
4	19.20521
5	37.11389

Vendo o coeficiente de variação, observamos que embora o grupo 1 realmente tenha um coeficiente de 24,22%, o grupo 5 ultrapassa com um coeficiente maior, portanto, sendo mais dispersos em relação a média de 21,11%. Para melhorar a visualização dos dados, segue o gráfico a seguir da população por taxa de pobreza.

Plotagem

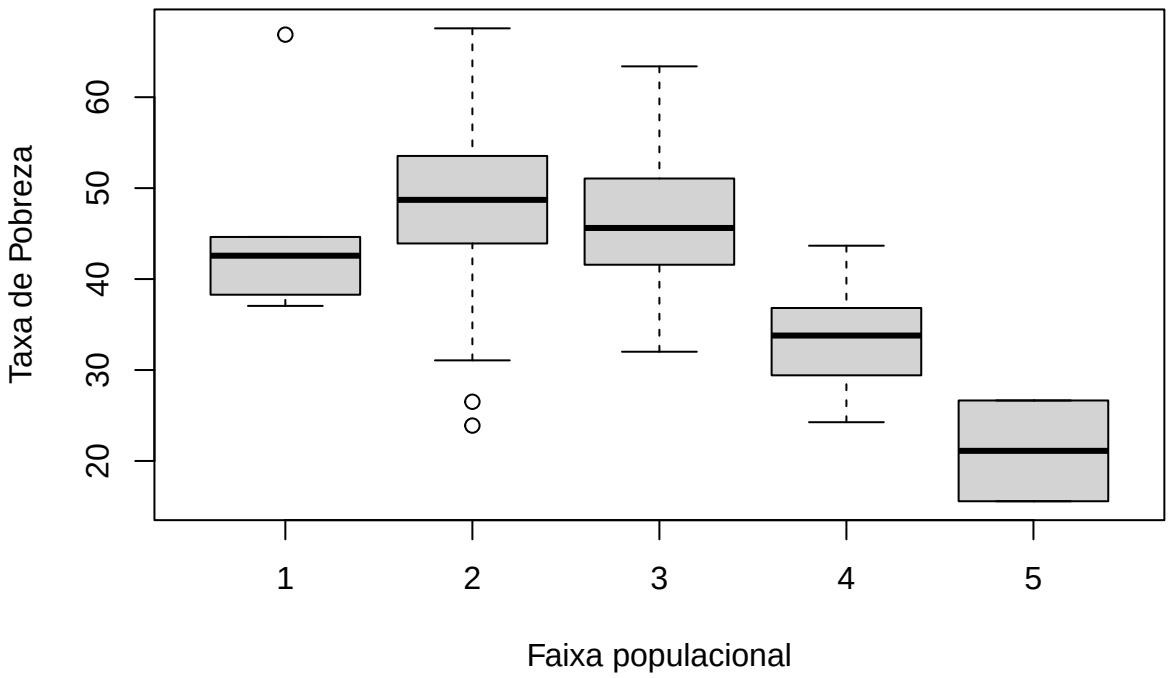
```
plot(al$pobre, al$poptot, ylab="População", xlab="Taxa de Pobreza")
```



A partir deste gráfico, conseguimos ver com certeza a presença de um outlier com população muito elevada e taxa de pobreza mais baixa, o qual se refere ao município de Maceió. Ao analisarmos os demais municípios, vemos uma concentração da taxa de pobreza em municípios com população baixa, os quais são referentes a municípios menores com menor comércio, renda per capita, além de menos arrecadação, investimento e, consequentemente, políticas públicas e geração de emprego.

h) Box plot de pobreza por faixa populacional

```
boxplot(pobre~fxpop, al, xlab='Faixa populacional', ylab='Taxa de Pobreza')
```



Ao analisarmos o gráfico, podemos confirmar a grande variação dos dados na faixa de população 5, além de vermos o movimento de queda da taxa de pobreza ao percorrermos municípios com cada vez maior população. Ao analisarmos os outliers, vemos que é detectado 1 no grupo de faixa populacional 1 e 2 no grupo 2, o qual também possui maior intervalo entre mínimos e máximos.

i) Criação de variável de urbanização

```
al$urb <- (al$popurb/al$poptot)*100
al$rur <- (al$poprur/al$poptot)*100
head(al[, c("poptot", "fxpop", "urb", "rur")],5)
```

poptot	fxpop	urb	rur
19377	2	26.32502	73.674976
17424	2	51.36019	48.639807
214006	5	84.80183	15.198172
44322	3	50.66784	49.332160
14230	2	93.05692	6.943078

Com a variável de urbanização, podemos ver analisar o quanto que o município conter uma área urbana influencia as demais variáveis socieconômicas, na teoria, em áreas urbanas é mais provavél de ter municípios mais desenvolvidos e com melhores dados tanto econômicos como educacionais.

j) Regressão com pobre como variável dependente

```
reg <- lm(pobre~urb, al)
print(summary(reg))
```

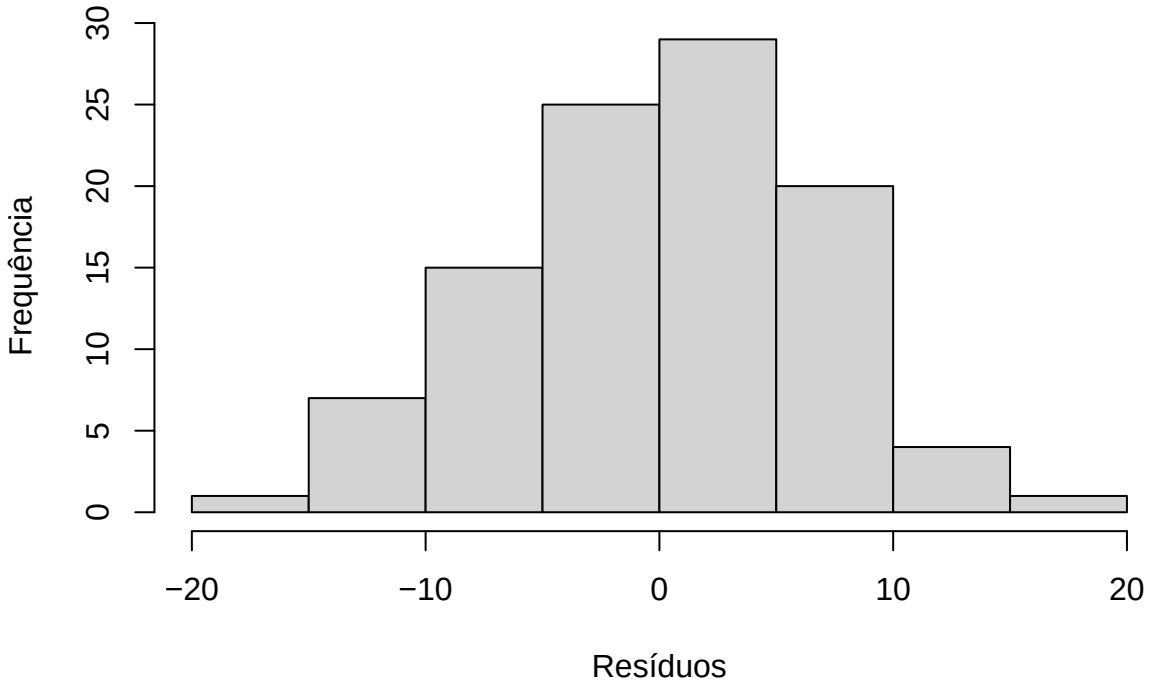
```
##
## Call:
## lm(formula = pobre ~ urb, data = al)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.4144  -4.1216   0.6515   4.7128  17.8903
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.28917    1.78077    36.1   <2e-16 ***
## urb         -0.32326    0.02914   -11.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.658 on 100 degrees of freedom
## Multiple R-squared:  0.5518, Adjusted R-squared:  0.5473
## F-statistic: 123.1 on 1 and 100 DF,  p-value: < 2.2e-16
```

Ao rodar a regressão linear pelo método dos mínimos quadrados ordinários, podemos inferir, ao analisar a variável de urbanização, que uma variação de 1 ponto percentual diminui, em média, a pobreza em 0,32 (valores absolutos). Ao analisar o R^2, vemos que o modelo explica 55,18% dos dados, visto que temos apenas uma variável no modelo, trazendo um modelo mal especificado. Ao analisar a significância estatística da variável, vemos que *urb* é estatisticamente significante ao nível de 1%.

k) Gerando os resíduos e fazendo o teste de Jarque-Bera

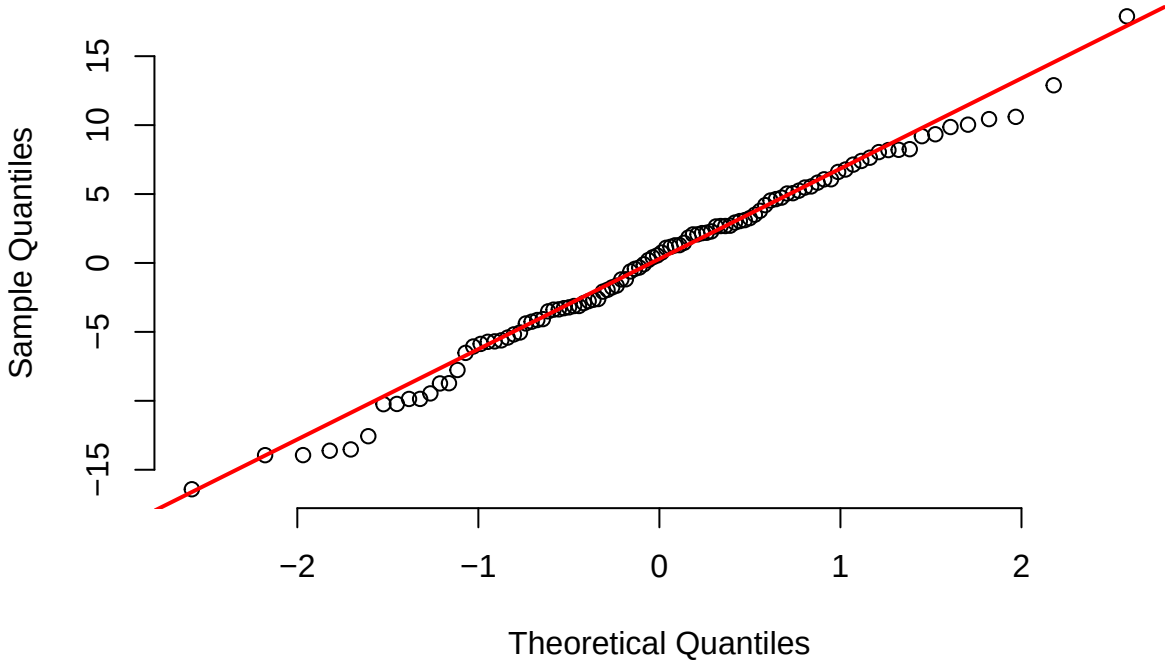
k.1) Gerando os resíduos

```
uhat <- reg$residuals
hist(reg$residuals, xlab='Resíduos', ylab='Frequência', main='')
```



Ao fazermos a análise gráfica da distribuição dos resíduos, vemos que eles, aparententemente, possuem uma distribuição normal, tanto visto o histograma anterior quanto ao analisarmos o gráfico Q-Q, o qual nos traz a análise da normalidade dos resíduos a partir de uma linha a 45º da origem, quanto mais alinhado mais normal é a distribuição.

```
qqnorm(reg$residuals, pch = 1, frame = FALSE, main='')
qqline(reg$residuals, col = "red", lwd = 2)
```



Portanto, para fazer um teste acertivo, segue o teste de Jarque-Bera de normalidade dos resíduos.

```
testejb <- jarque.bera.test(uhat)
print(testejb)
```

k.2) Teste de Jarque-Bera

```
##
##  Jarque Bera Test
##
## data:  uhat
## X-squared = 0.8252, df = 2, p-value = 0.6619
```

Conforme o resultado do teste (valor-p acima de 0,1), aceitamos a hipótese nula de que os resíduos são normalmente distribuídos.

Questão 3 - Pýthon

Link Google Colab