Trabalho Final de Econometria I

Modelo de Trabalho para a Cadeira de Econometria I - Uma Análise Salarial Simples do Estado do Acre, Utilizando Dados da PNAD Contínua do ano de 2017.

Gustavo de Oliveira *- 14 de agosto de 2019*

O presente trabalho visa exemplificar um modelo de trabalho para a cadeira de Econometria I. Este foi feito, essencialmente, utilizando a linguagem LaTeXe a linguagem R, por meio da interface do pacote *knitr*. Todos os códigos de *chunks*, bem como a classe do trabalho ou o documento TeXestarão disponibilizados para uso. Além disso, ressalta-se que a base de dados utilizada para o trabalho também estará disponível. O objetivo é a possibilidade de replicação deste documento por quaisquer interessados.

1 A Parte Prática

O corte inicial foi referente ao estado do Acre, por meio dos dados da PNADC. Isso é, temos em mãos, inicialmente, um dataframe com 9287 observações. Além disso, vale ressaltar que o dataframe que temos já foi modificado: temos *17* variáveis. A primeira coisa que temos que fazer é ler os dados.

1.1 Preparação dos Dados

Primeiro, definiremos nosso diretório de trabalho (a pasta onde iremos trabalhar). Feito isso, só nos resta ler o arquivo que desejarmos - nesse caso o **pnadc_anual_ac_2017.csv**

```
setwd('/home/gustavo/Documents/EconometriaI/')
pnadc_2017 <- read.csv('pnadc_anual_ac_2017.csv')
```

O dataframe pnad_2017 é composto de inúmeras variáveis que podemos trabalhar. Vamos, entretanto, restringir nosso dataframe mais uma vez. Afim de obtermos uma maior homogeneidade nos nossos dados, trabalharemos com aquelas observações que correspondem à posição = 4. Isso é, empregados do setor público (inclusive às empresas de economia mista). Chamaremos nosso novo dataframe de “data”.

```
data <- subset(pnadc_2017, posicao == 4)
```

Assim, restringimos nosso dataframe a 626 observações, ainda com 17 variáveis. Podemos, agora, criar e manipular as variaveis.

1.2 Manipulação e Criação de Variáveis

A primeira coisa que faremos é criar algumas dummys. A primeira será seferente ao sexo. No nosso dataframe, sexo = 1 é homem, assim, seguiremos a mesma lógica. Utilizaremos uma estrutura de repetição e uma de controle para o fazermos. Caso, sexo = 1, a observação “será homem”; caso contrário (0), será mulher. Chamaremos nossa nova variável de genero.

```
genero <- c() # criamos um vetor vazio para genero.
for(i in 1:nrow(data)){ # se for homem, genero[i] = 1,
                        # c.c genero[i] = 0
  if(data$sexo[i] == 1){
```

```
    genero[i] <- 1
  } else{
    genero[i] <- 0
  }
}
```

Como esperado, essa variável terá o mesmo comprimento do nosso dataframe, 626 observações. Achamos prudente criar uma outra dummy. referente a cor do indivíduo. Nossa nova variável será "branco". Se a "observação for branca" essa será 1, 0 c.c:

```
branco <- c() # criamos um vetor vazio para genero.
for(i in 1:nrow(data)){ # se for branco, branco[i] = 1,
                        # c.c branco[i] = 0
  if(data$cor_raca[i] == 1){
    branco[i] <- 1
  } else{
    branco[i] <- 0
  }
}
```

Vamos, então, criar um novo dataframe - o que utilizaremos evetivamente para as estimações e estatísticas. Chamaremos esse novo dataframe de data2017

```
data2017 <- data.frame(
  data$idade,
  data$anos_estudo,
  data$rendimento_mensal,
  branco,
  genero
)
```

Afim de elhor organizar nossos dados, vamos renomear as variáveis:

```
names(data2017) <-
  c('idade', 'estudo', 'salmes', 'branco', 'genero')
```

Podemos agora, trabalhar melhor com as variáveis.

1.3 Estatísticas descritivas e Gráficos

Agora que temos nosso dataframe de trabalho, vamos melhor verificar as variáveis afim de, se possível, tratarmos de linearidades ou outros vatores que poderiam prejudicar ou inviabilizar nosso modelo futuro.

A primeira coisa que devemos fazer é uma análise descritiva das variáveis. para uma melhor visualização, utilizaremos o pacote *xtable* do R, presente no CRAN. Esse pacote exporta as tabelas já em formato LaTeXpara o nosso documento.

1.3.1 Estatísticas Descritivas

Vamos melhor entender nossas variáveis, então:

```
library(xtable)
print(xtable(summary(data2017),
             caption = 'Estatísticas Descritivas de data2017'),
      include.rownames=FALSE, caption.placement = 'top')
```

Tabela 1 – Estatísticas Descritivas de data2017

idade	estudo	salmes	branco	genero
Min. :15.00	Min. : 0.00	Min. : 100	Min. :0.0000	Min. :0.0000
1st Qu.:32.00	1st Qu.:11.00	1st Qu.: 1200	1st Qu.:0.0000	1st Qu.:0.0000
Median :40.00	Median :14.00	Median : 2000	Median :0.0000	Median :0.0000
Mean :40.66	Mean :12.75	Mean : 3056	Mean :0.2332	Mean :0.4281
3rd Qu.:49.00	3rd Qu.:15.00	3rd Qu.: 3200	3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :70.00	Max. :15.00	Max. :30000	Max. :1.0000	Max. :1.0000
		NA's :1		

Podemos perceber, por exemplo, com a tabela de estatísticas descritivas, que homens representam cerca de 42,81% da amostra. Da mesma forma, percebemos que cerca de 23,32% da amostra é branca.

1.3.2 Análise Gráfica

Vamos agora, realizar uma análise gráfica afim de identificar, por exemplo, alguma anomalia no nosso conjunto de dados - um outlier por exemplo. Não o faremos para as dummys

```
par(mfrow = c(1,3))
for(i in 1:3){
  hist(data2017[, i], xlab = names(data2017)[i],
       main = NULL, probability = TRUE)
}
```

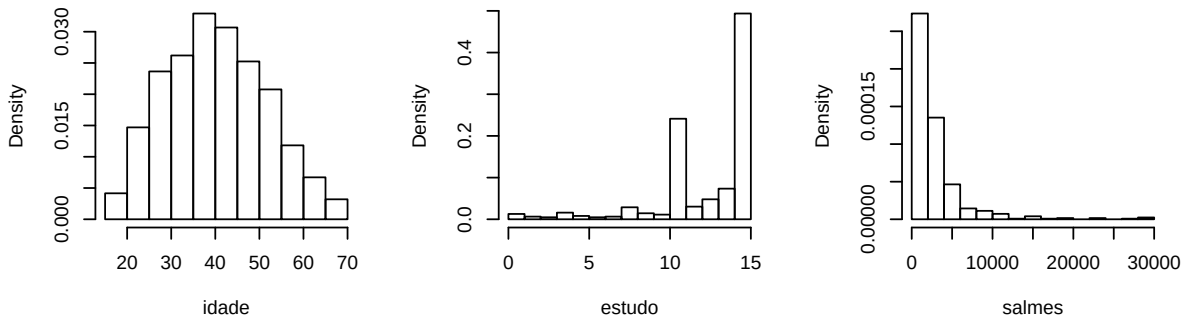

Figura 1 – Histogramas das Variáveis idade, estudo e salmes

Caso os histogramas não nos indique com exatidão uma possível inferência sobre nossos dados, podemos verificar a presença de possíveis outliers por meio dos boxplots.

```
par(mfrow = c(1,3))
for(i in 1:3){
```

```
  boxplot(data2017[, i], xlab = names(data2017)[i],
       main = NULL)
}
```

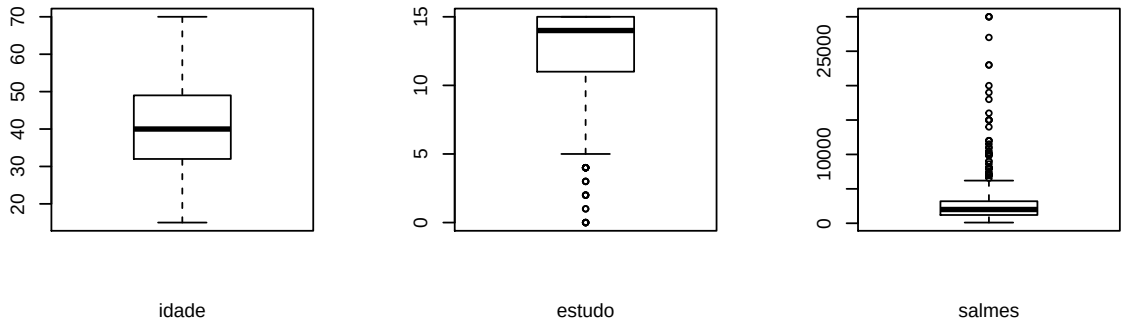

Figura 2 – Boxplot das Variáveis idade, estudo e salmes

Afim de melhorarmos nosso modelo, vamos realizar uma transformação linear na variável **salmes**. Assim, aplicaremos o log sobre essa e incluiremos no nosso data2017. Para isso, basta aplicarmos a função log() - chamaremos essa nova variável de lnsalmes:

```
data2017$lnsalmes <- log(data2017$salmes)
```

E temos os gráficos a seguir:

```
par(mfrow = c(1, 2))
boxplot(data2017$lnsalmes, xlab = 'lnsalmes')
hist(data2017$lnsalmes, xlab = 'lnsalmes', main = NULL)
```

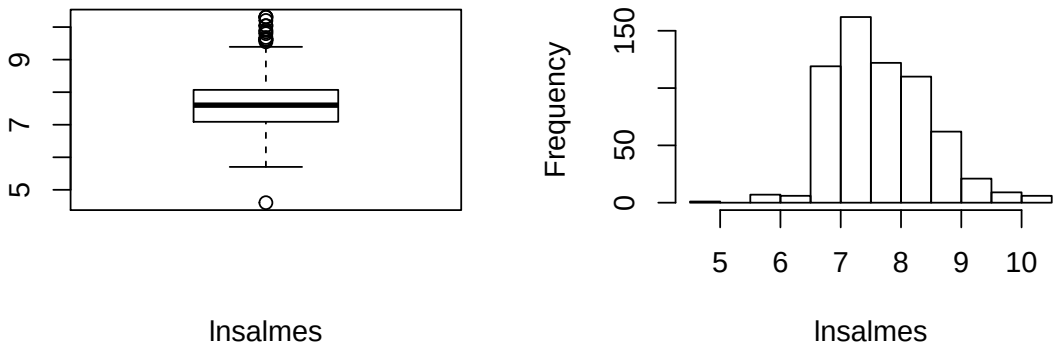

Figura 3 – Boxplot da Variável lnsalmes

1.4 Estimando nosso Modelo

Agora que já fizemos uma análise descritiva das nossas variáveis, vamos rodar a nossa primeira regressão. Esta terá a seguinte forma teórica:

$$lnsalmes = \beta_0 + \beta_1 genero + \beta_2 idade + \beta_4 estudo + \beta_5 branco + u$$

Vale lembrar que temos que levar em consideração os missings values, utilizaremos o na.action. Assim, estimamos nosso modelo, fazendo referência ao dataframe data2017:

```
modelo2017 <- lm(lnsalmes ~ genero + idade + estudo +
                   branco, na.action = na.omit, data = data2017)
```

Vamos visualizar a tabela de coeficientes utilizando o pacote stargazer. Este nos permite uma saida elegante para as regressões:

```
library(stargazer)
stargazer(modelo2017, title = 'Primeiro Modelo Estimado')
```

Tabela 2 – Primeiro Modelo Estimado

	Dependent variable:
	lnsalmes
genero	0.350***
	(0.054)
idade	0.020***
	(0.002)
estudo	0.111***
	(0.009)
branco	0.256***
	(0.063)
Constant	5.213***
	(0.170)
Observations	625
R^2	0.279
Adjusted R^2	0.275
Residual Std. Error	0.666 (df = 620)
F Statistic	60.112*** (df = 4; 620)
Note:	*p<0.1; **p<0.05; ***p<0.01

Alternativamente podemos criar um modelo que considere idade ao quadrado como variável explicativa:

$$lnsalmes = \beta_0 + \beta_1 genero + \beta_2 idade + \beta_3 idade^2 + \beta_4 estudo + \beta_5 branco + u$$

Para criarmos $idade^2$ na nossa regressão, basta utilizarmos o operador I, na "formula".

```
modelo2017.2 <- lm(lnsalmes ~ genero + idade + I(idade^2) + estudo +
                   branco, na.action = na.omit, data = data2017)
```

Comparando os modelos:

```
stargazer(modelo2017, modelo2017.2, title = 'Comparando os Modelos')
```

Tabela 3 – Comparando os Modelos

	Dependent variable:	
	lnsalmes	
	(1)	(2)
genero	0.350***	0.348***
	(0.054)	(0.054)
idade	0.020***	0.054***
	(0.002)	(0.015)
I(idadeˆ2)		−0.0004**
		(0.0002)
estudo	0.111***	0.105***
	(0.009)	(0.009)
branco	0.256***	0.259***
	(0.063)	(0.063)
Constant	5.213***	4.634***
	(0.170)	(0.306)
Observations	625	625
R^2	0.279	0.285
Adjusted R^2	0.275	0.280
Residual Std. Error	0.666 (df = 620)	0.664 (df = 619)
F Statistic	60.112*** (df = 4; 620)	49.450*** (df = 5; 619)
Note:		*p<0.1; **p<0.05; ***p<0.01

1.4.1 Análise de Resíduos

Utilizaremos o modelo "modelo2017.2" por compreendermos que $idade^2$ é teoricamente relevante para o modelo, seja pela sua significância individual ou conjunta. Extrairemos os resíduos do modelo 2 e chamaremos de "residuos":

```
residuos <- resid(modelo2017.2)
```

Uma primeira análise gráfica é apresentada abaixo:

```
par(mfrow = c(1,1))
plot(residuos, type = 'l')
abline(h=0, col = 'red', lty = 2)
```

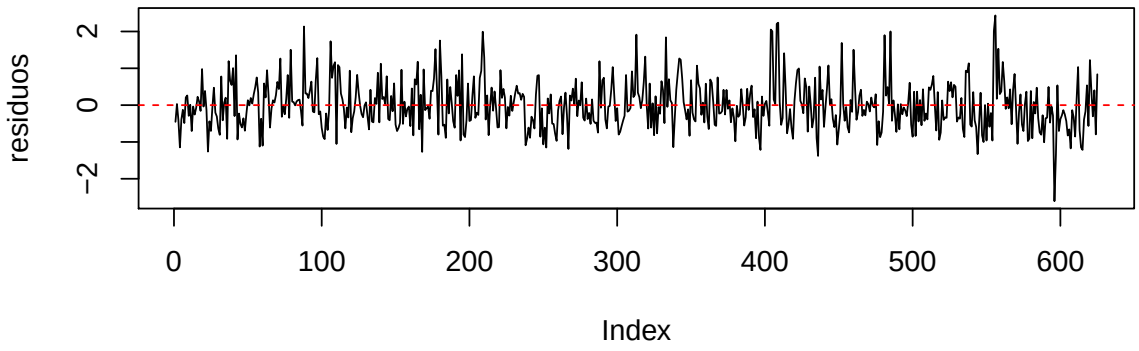

Figura 4 – Resíduos do modelo 2

Podemos ter uma visão mais geral dos gráficos relacionados ao modelo com o próprio comando plot. Este nos dará informações sobre leverages ou mesmo o qqplot:

```
par(mfrow = c(2,2))
plot(modelo2017.2)
```

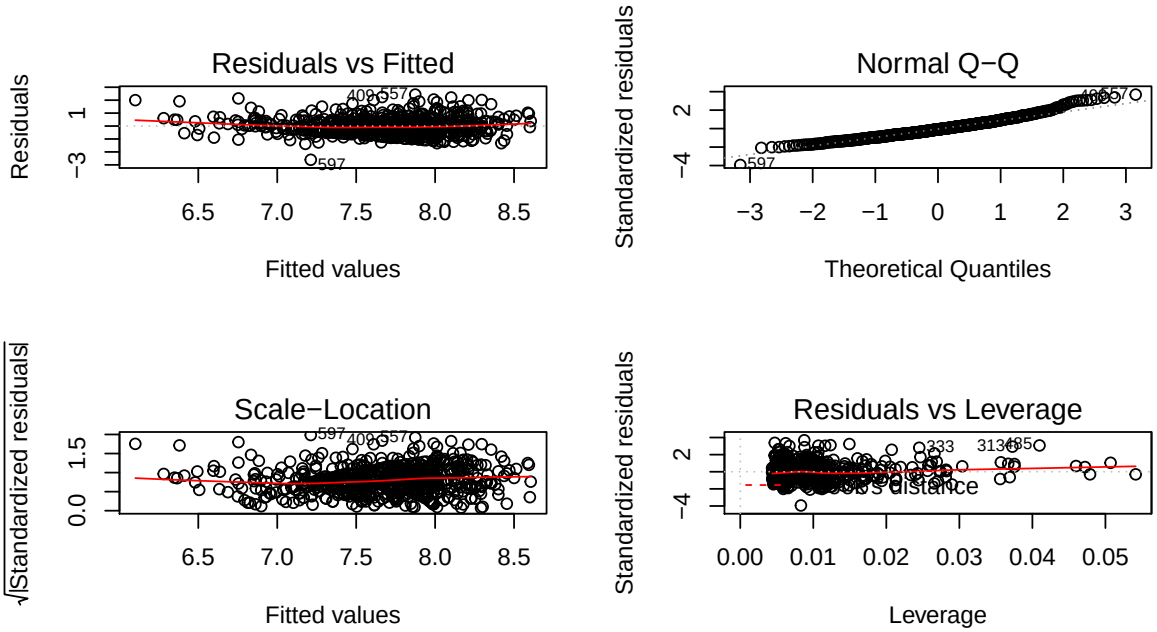

Figura 5 – Resíduos do modelo 2

Referências

MARTINE, G. et al. A pnad: notas para uma avaliação. *Livros*, p. 281–310, 2015.