

Relatório IC

Fernanda Buzza Alves Barros

___ de ___ de ___

INTRODUÇÃO

Problemas de dados faltantes em pesquisa são recorrentes em bancos de dados. Para a solução desses problemas existem vários métodos que podem ser utilizados. Entretanto, todos os métodos possuem uma questão principal: como inferir os valores não observados?

Para a resposta dessa pergunta, temos que o ideal seria ter os dados, porém na falta deles temos que utilizar o método que melhor se ajusta a distribuição dos dados.

Nessa pesquisa utilizaremos o método proposto por Rubin (1987), Van Buuren e Groothuis-Oudshoorn (2011), que é conhecido como Imputação Múltipla.

METODOLOGIA

A Imputação Múltipla consiste em gerar valores (m vezes) para os dados faltantes, ela cria uma matriz com todas as M imputações. Para gerar essas imputações existem alguns métodos, como por exemplo *Predictive Mean Matching (pmm)* e *Unconditional Mean Imputation (mean)*, que serão os métodos utilizados nesse estudo.

Predictive Mean Matching (pmm)

Unconditional Mean Imputation (mean)

RESULTADOS

Banco de dados

Para realizar a imputação dos dados utilizamos o banco de dados *US Term Life insurance* do pacote *CASdatasets* disponível no software R. As imputações e os resultados foram obtidos utilizando esse mesmo software estatístico. O banco de dados possui 18 variáveis com 500 observações, como pode ser visto abaixo.

```
## 'data.frame':   500 obs. of  18 variables:
##  $ Gender      : int  1 1 1 1 1 1 0 1 1 1 ...
##  $ Age         : int  30 50 39 43 61 34 75 29 35 70 ...
##  $ MarStat     : int  1 1 1 1 1 2 0 1 2 1 ...
##  $ Education   : int  16 9 16 17 15 11 8 16 4 17 ...
##  $ Ethnicity   : int  3 3 1 1 1 2 1 1 3 1 ...
##  $ SmarStat    : int  2 1 2 1 2 1 0 2 1 2 ...
##  $ Sgender     : int  2 2 2 2 2 2 0 2 2 2 ...
##  $ Sage       : int  27 47 38 35 59 31 0 31 45 74 ...
##  $ Seducation  : int  16 8 16 14 12 14 0 17 9 16 ...
##  $ NumHH      : int  3 3 5 4 2 4 1 3 2 2 ...
##  $ Income     : int  43000 12000 120000 40000 25000 28000 2500 100000 20000 101000 ...
##  $ TotIncome  : int  43000 0 90000 40000 1020000 0 0 84000 0 6510000 ...
```

```
## $ Charity      : int  0 0 500 0 500 0 0 0 0 284000 ...
## $ Face         : int  20000 130000 1500000 50000 0 220000 0 600000 0 0 ...
## $ FaceCVLifePol : int  0 0 0 75000 7000000 0 14000 0 0 2350000 ...
## $ CashCVLifePol : int  0 0 0 0 300000 0 5000 0 0 0 ...
## $ BorrowCVLifePol: int  0 0 0 5 5 0 5 0 0 5 ...
## $ NetValue      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Porém selecionamos as seguintes variáveis: Gênero (gênero do entrevistado); Idade (idade do entrevistado); Estado Civil (estado civil do entrevistado); Escolaridade (número de anos de escolaridade do entrevistado); Etnia (etnia); Renda (renda anual da família).

O banco de dados não possui dados faltantes, portanto para avaliar a Renda (variável de interesse) foi necessário gerar os dados faltantes. Sendo assim utilizamos uma distribuição binomial com probabilidade de sucesso de 0.2 para a criar dos dados faltantes na variável Renda, fixando a semente em `set.seed(0)`.

Primeiras observações do banco de dados original:

```
##   Gender Age MarStat Education Ethnicity Income
## 1      1  30      1      16          3  43000
## 2      1  50      1       9          3  12000
## 3      1  39      1      16          1 120000
## 4      1  43      1      17          1  40000
## 5      1  61      1      15          1  25000
## 6      1  34      2      11          2  28000
```

Primeiras observações do banco de dados com dados faltantes na variável Renda:

```
##   Gender Age MarStat Education Ethnicity Income
## 1      1  30      1      16          3    NA
## 2      1  50      1       9          3  12000
## 3      1  39      1      16          1 120000
## 4      1  43      1      17          1  40000
## 5      1  61      1      15          1    NA
## 6      1  34      2      11          2  28000
```

Análise Descritiva

Analisaremos as relações entre as variáveis selecionadas do banco de dados original. O principal objetivo é verificar os possíveis questionamentos sobre a Renda a partir das outras variáveis. Temos interesse em responder as seguintes perguntas:

- Como está distribuída a variável Renda.
- Qual é a relação entre a Renda e o Gênero.
- Com maiores anos de escolaridade há aumento da renda.
- O estado civil tem influência na renda.
- Avaliar a relação entre a etnia e a renda.

Primeiramente iremos analisar algumas das principais informações das variáveis, como: mínimo, máximo, média, mediana e quantis. Assim obtemos a tabela abaixo:

Para a variável Idade temos:

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  20.00  37.00   47.00   47.16  58.00   85.00
```

Assim para a pesquisa a idade máxima é 85 anos e a idade mínima é 20 anos. A média é 47,16 anos e a mediana 47 anos. O primeiro quantil é de 37 anos e o terceiro é de 58 anos.

Analisando a variável Anos de Escolaridade temos:

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	2.00	12.00	14.00	14.06	16.00	17.00

O mínimo de anos de escolaridade é 2 anos e o máximo é 17 anos. A mediana e a média são 14 anos e 14,06 anos, respectivamente. E o primeiro quantil é 12 anos e o terceiro quantil é 16 anos.

Analisando a variável Renda obtemos:

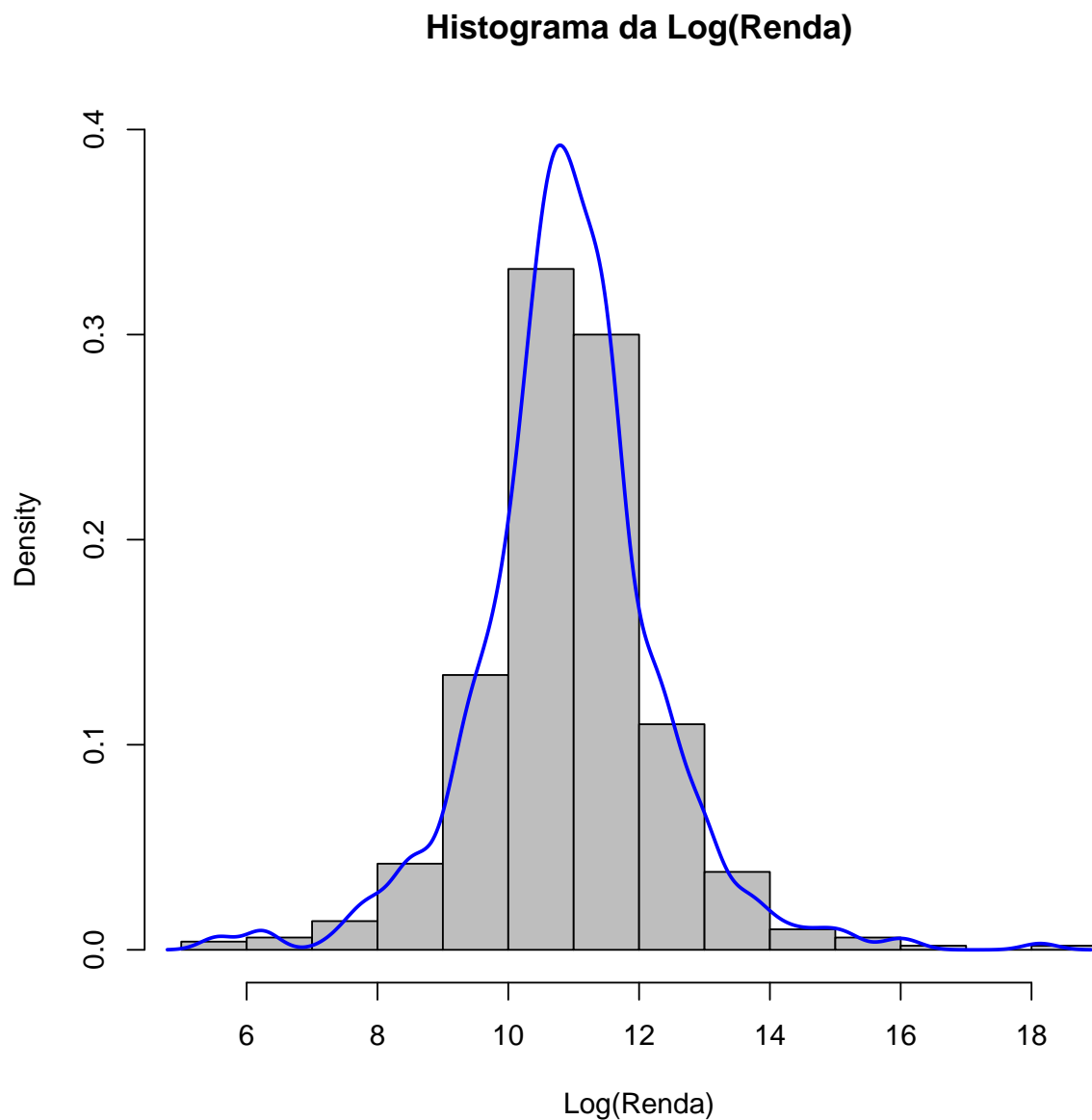
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	260	28000	54000	321022	106000	75000000

Essa variável possui como renda mínima 260 dólares e renda máxima 75.000.000 dólares. A mediana e a média são 54.000 dólares e 321.022 dólares, respectivamente. E o primeiro e terceiro quantis são 28.000 dólares e 106.000 dólares.

Por fim, avaliando as variáveis Estado Civil, Gênero e Etnia temos:

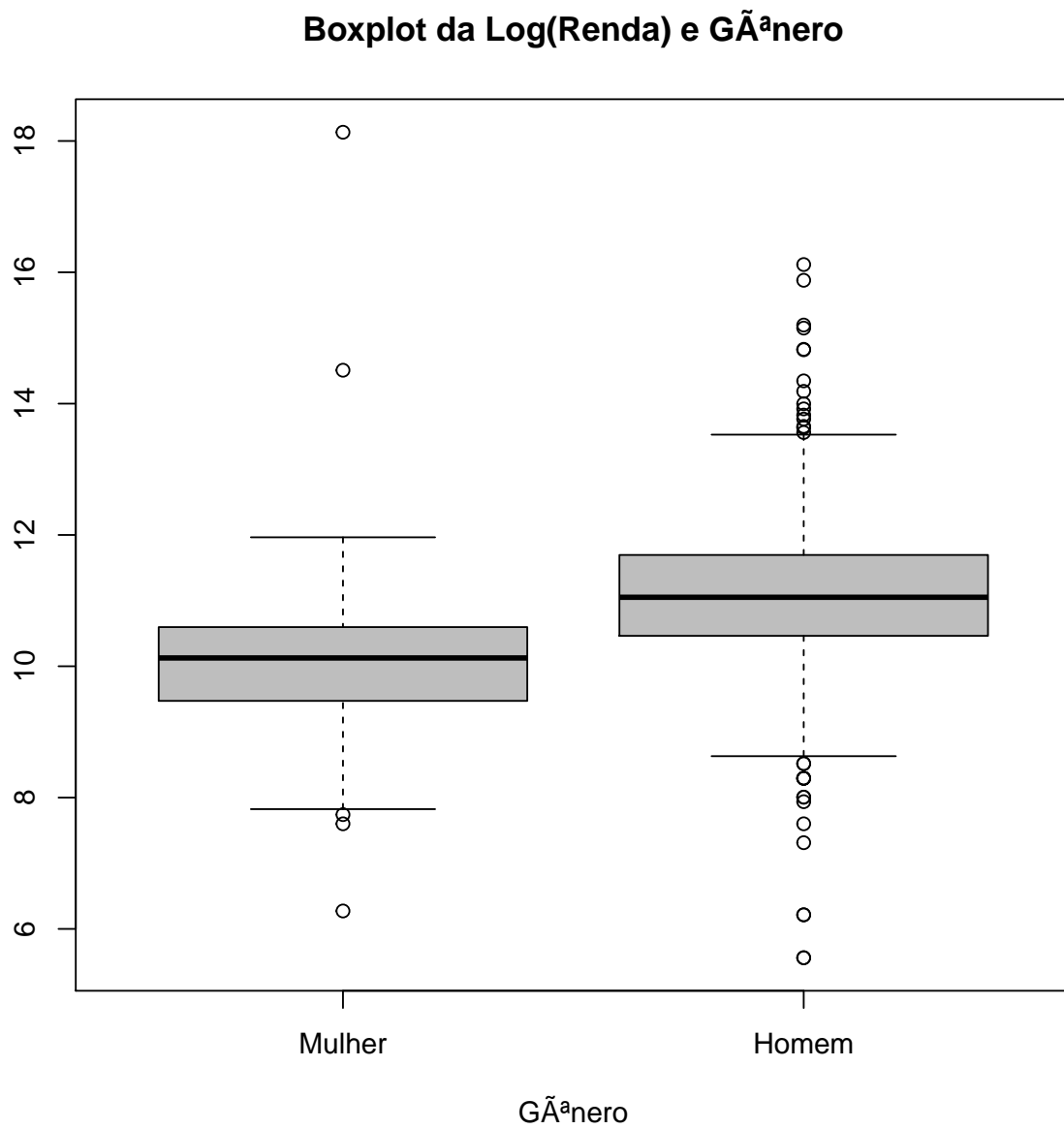
Sendo assim a pesquisa possui 87 respondentes do sexo feminino e 413 respondentes do sexo masculino.

Distribuição da Renda



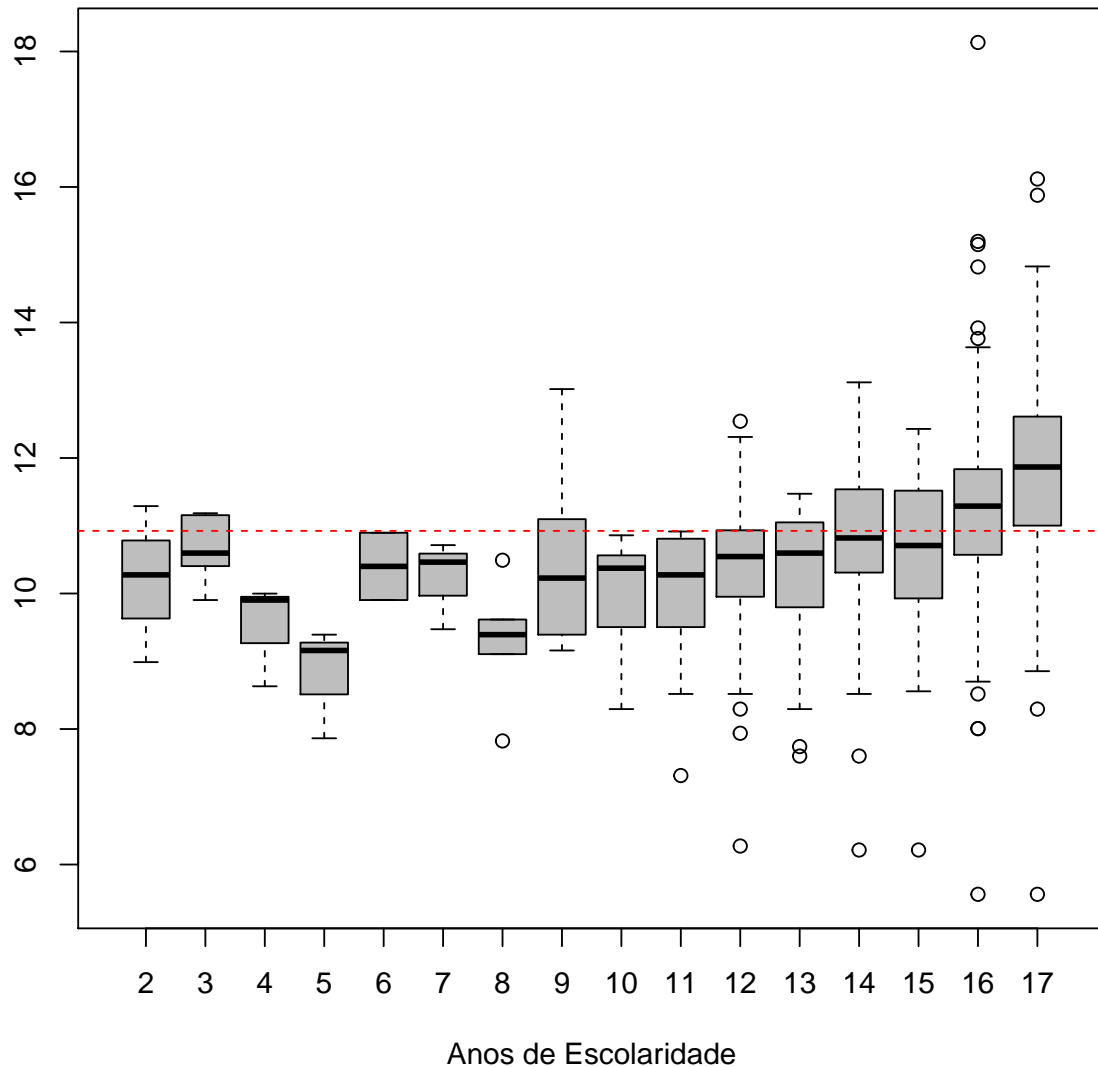
Pelo histograma podemos avaliar a distribuição da variável Renda, a partir dos dados retirados do banco de dados original. Observamos uma maior concentração de valores entre o Log(Renda) de 10 a 12. Nas caldas podemos perceber reduções de valores da renda familiar.

Renda e Gênero



Ao plotarmos os boxplots da Log(Renda) e o Gênero vemos a relação entre os valores da renda dos homens comparados com os das mulheres. Nesse caso os homens possuem maiores valores de renda do que as mulheres.

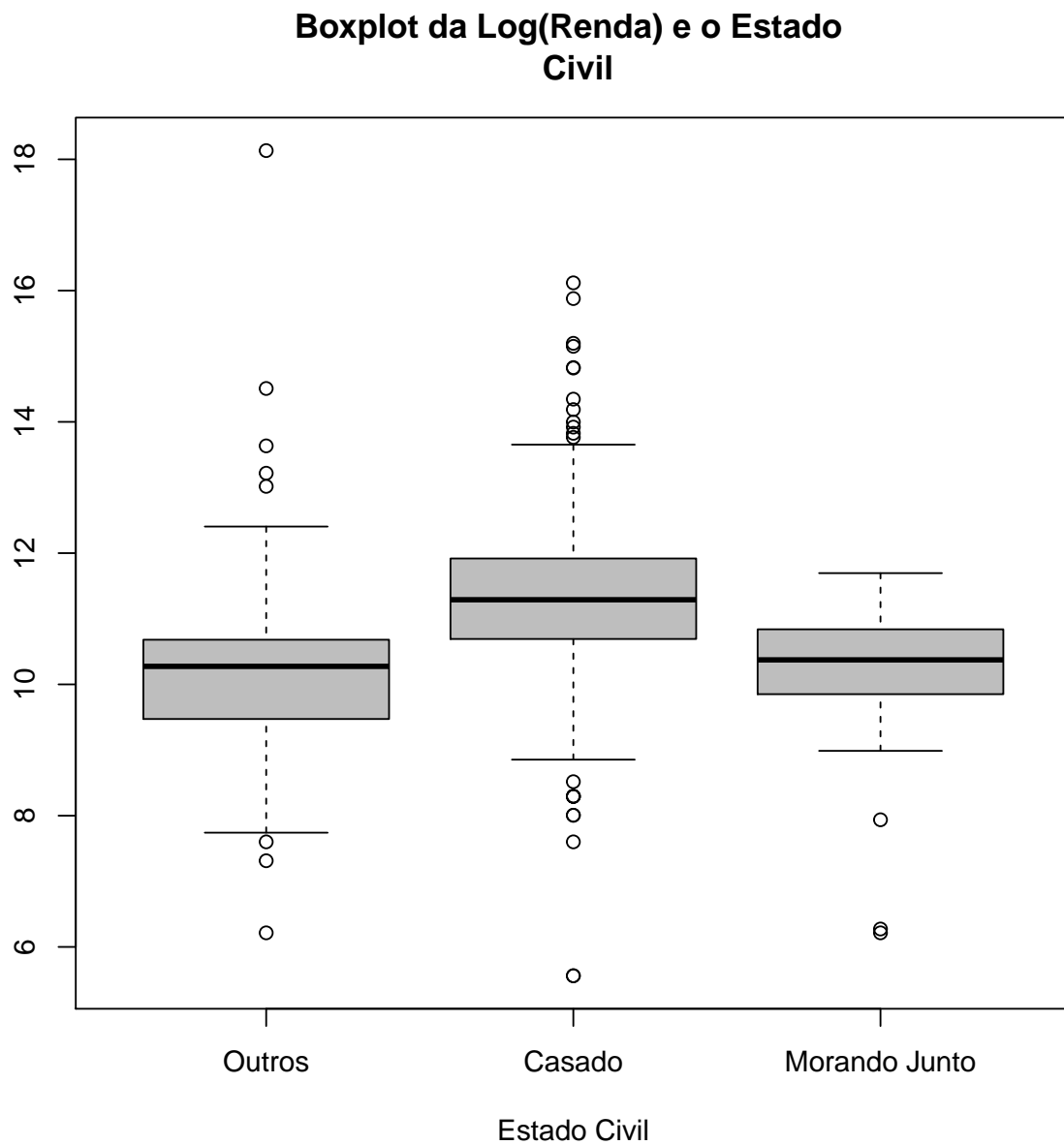
Boxplot da Log(Renda) e Anos de Escolaridade



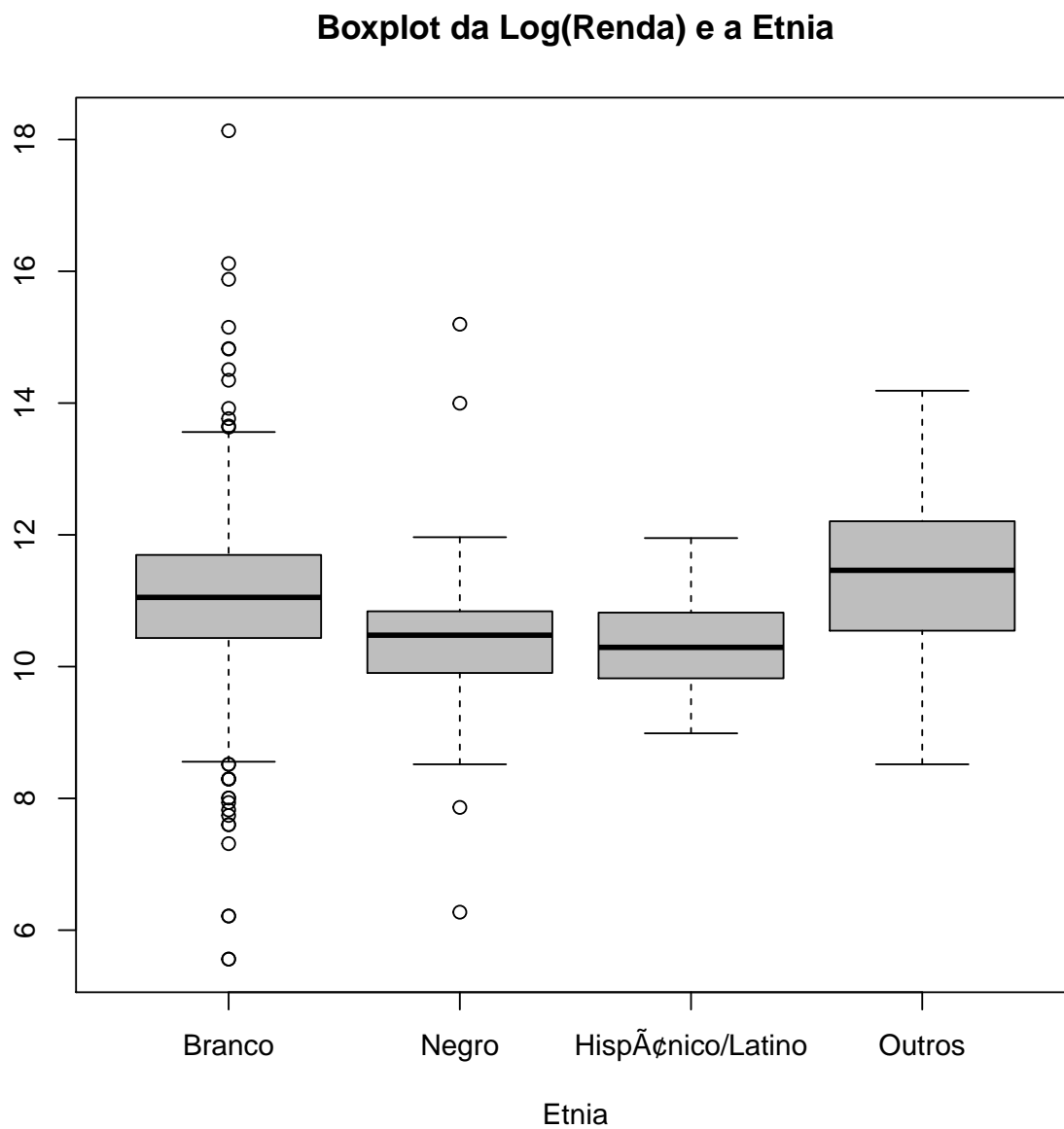
Ao analisar o efeito na quantidade de Anos de Escolaridade e a Renda, percebemos um crescimento na quantidade ganha de renda de acordo com os anos de escolaridade. Observamos que com a inclusão da reta pontilhada em vermelho, que representa a média da variável Log(Renda), a possibilidade de determinar os anos de escolaridade que estão acima da média de valores ganhos de renda. Entre os anos de escolaridade de 2 a 8 anos não percebemos uma relação crescente, sendo que há uma queda em 4, 5 e 8 anos de escolaridade, que pode ser devido a quantidade de entrevistados desses grupos representados no banco de dados; o que pode ser verificado na tabela abaixo:

##																	
##	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Sum
##	3	5	3	3	2	3	5	6	7	15	101	31	68	16	130	102	500

Renda e Estado Civil

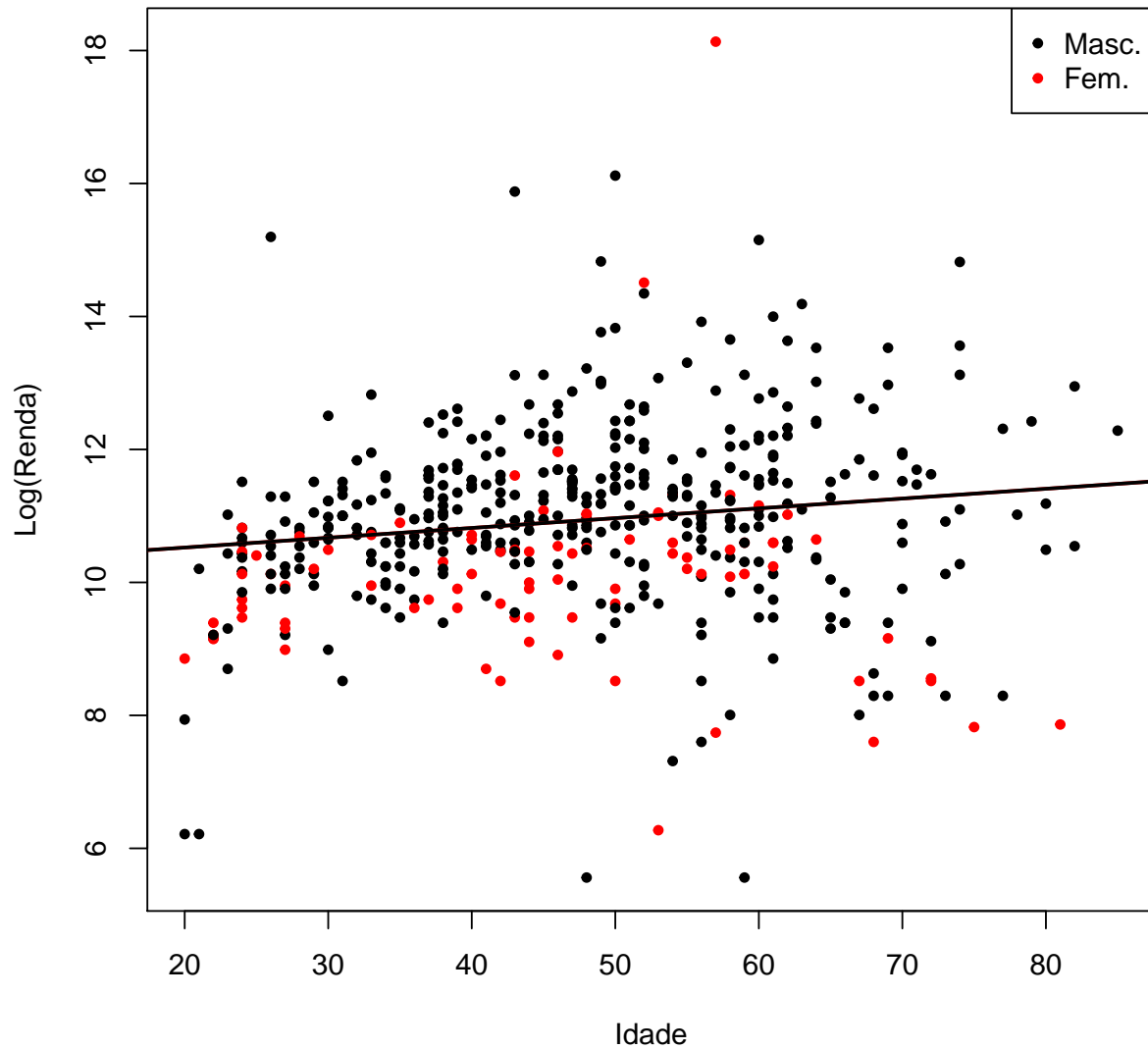


Para analisar a relação entre o Estado Civil e a Log(Renda) percebemos que as pessoas casadas possuem uma renda maior, quando comparado com os outros grupos apresentados pelo banco de dados.



Para avaliar a relação entre a Etnia e a Log(Renda) observamos maiores valores de renda para o grupo white e o others, sendo que os grupos black e hispanic apresentam similaridades nos valores de renda.

Gráfico da Log(Renda) e Idade



Imputação

Para realizar a imputação utilizamos o pacote *Multivariate Imputation With Chained Equations (MICE)*. A função que realiza a imputação chama-se *mice*, e nesse estudo realizamos a imputação 5 vezes ($m=5$) tanto para o método da *pmm* e da *mean* da função para comparar os resultados.

Abaixo temos o output da função de imputação com as principais informações.

CONCLUSÃO

REFERÊNCIAS BIBLIOGRÁFICAS

Rubin (1987)

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. [linked phrase](#)

Morris TP, White IR, Royston P (2015). Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. ;14:75.

Frees, E.W. (2011). Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.