Aplicação de métodos de simulação de dados sintéticos

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 para realizar a pesquisa: Gênero (gênero do entrevistado); Idade (idade do entrevistado); Estado Civil (estado civil do entrevistado); Escolaridade (número de anos de escolaridade do entrevistado); Etnia (etnia do entrevistado); Renda (renda anual da família do entrevistado).

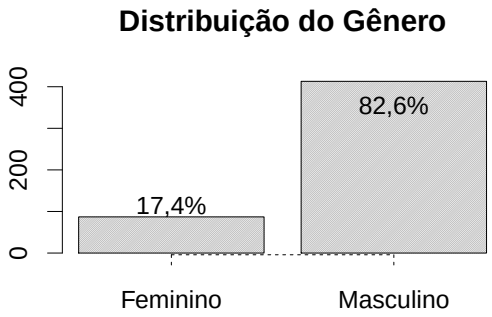
Primeiras observações do banco de dados original:

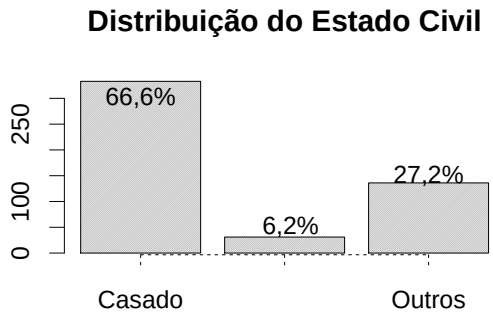
```
##   Gender Age MarStat Education Ethnicity Income
## 1      1  30       1        16         3  43000
## 2      1  50       1         9         3  12000
## 3      1  39       1        16         1 120000
## 4      1  43       1        17         1  40000
## 5      1  61       1        15        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

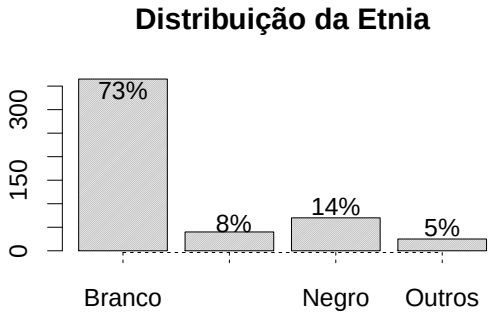
Análise Descritiva

Após a escolha das variáveis para esse estudo, iremos realizar uma análise descritiva de cada uma delas, e assim avaliar as relações existentes entre a variável resposta e as covariáveis do banco de dados. Ao final realizaremos um dos principais objetivos dessa pesquisa, que é verificar os possíveis questionamentos sobre a Renda a partir das outras variáveis.

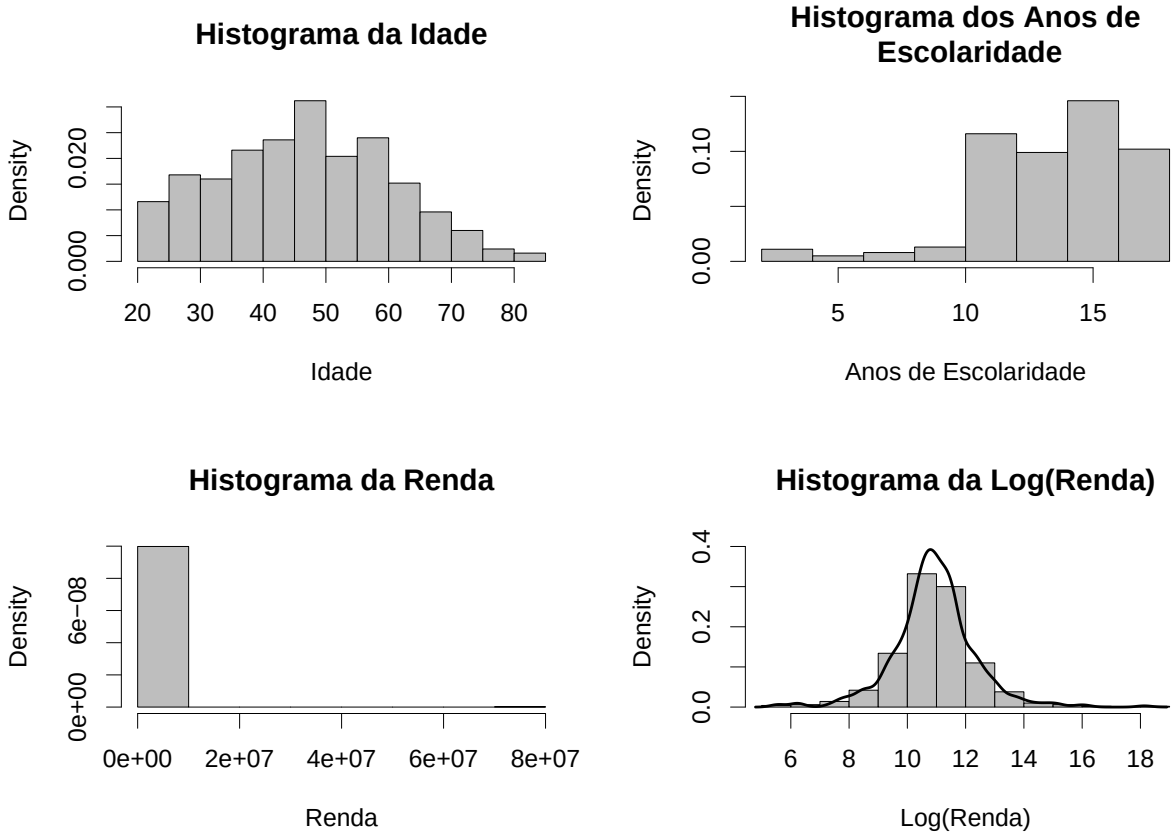
Primeiramente analisaremos as variáveis individualmente, com interesse em suas distribuições e comportamentos. Pelos dados observamos que as variáveis contínuas são: Renda, Idade e Escolaridade, e as variáveis discretas são: Gênero, Estado Civil e Etnia.







Para as variáveis discretas temos que o banco de dados possui uma quantidade maior de entrevistados do sexo masculino (413 entrevistados) do que do sexo feminino (87 entrevistadas); para o estado civil temos uma concentração maior de respostas para os entrevistados Casados (333 entrevistados) e a menor quantidade de entrevistados pertence ao estado civil de Morando Juntos (31 entrevistados), sendo que o estado civil Outros possui 136 entrevistados; por fim para a etnia temos uma maior quantidade de entrevistados que possuem etnia Branco (365 entrevistados), sendo que as outras etnias possuem valores menores de entrevistados no banco de dados: Hispânico (40 entrevistados), Negro (70 entrevistados) e Outros (25 entrevistados).

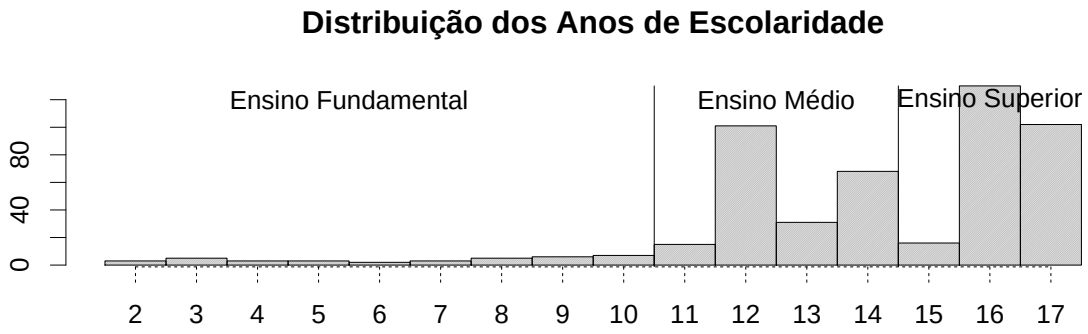


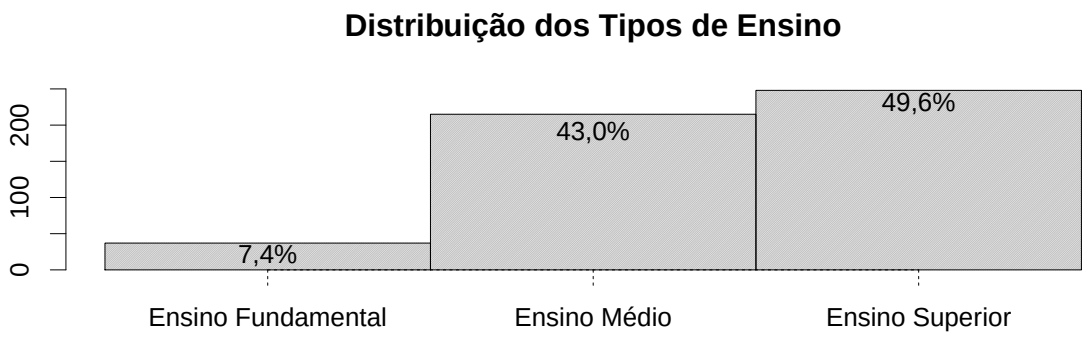
Para as variáveis contínuas temos que a variável Idade está bastante distribuída entre os 20 anos e 70 anos, após 70 anos vemos poucos entrevistados no banco de dados, sendo também que a idade máxima é 85 anos e a idade mínima é 20 anos. A média é 47.164 anos e a mediana 47 anos. O primeiro quantil é de 37 anos, representando a idade que deixa 25% das observações abaixo e 75% acima dessa idade. E o terceiro quantil é de 58 anos, representando a idade que possui 75% das observações abaixo dela e 25% das observações acima dela.

A distribuição da variável Escolaridade possui maior concentração de entrevistados após 10 anos de escolaridade.

E para a variável Renda temos que a renda mínima anual é 260 dólares e a renda máxima anual é 75000000 dólares. A mediana e a média são 54000 dólares e 321021 doláres, respectivamente. O primeiro quantil é de 28000 dólares, representando o valor de renda anual que deixa 25% das observações abaixo dela e 75% acima deela. E o terceiro quantil é de 106000 dólares, representando a renda anual que possui 75% das observações abaixo dela e 25% das observações acima dela. Aplicamos o logarítimo para melhor visualização da distribuição da renda anual através do histograma e percebemos uma aparência com a distribuição normal.

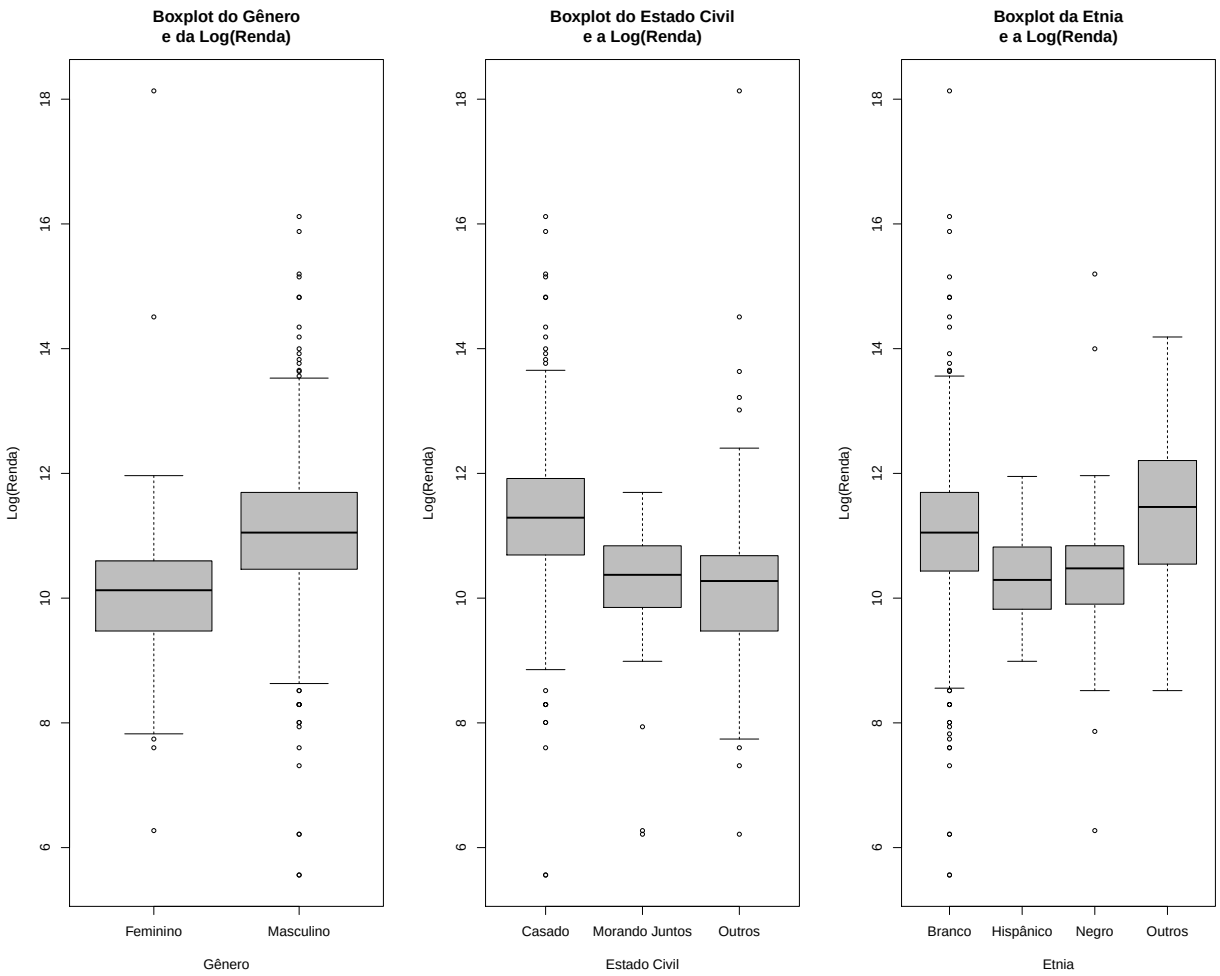
Além da análise da variável Escolaridade em anos, foi realizada também a análise dos anos de escolaridade divididos pelos tipos de ensino existentes, que são: 2-10 anos de escolaridade é o Ensino Fundamental, 11-14 anos de escolaridade é o Ensino Médio e de 15-17 anos de escolaridade é o Ensino Superior, assim obtemos:



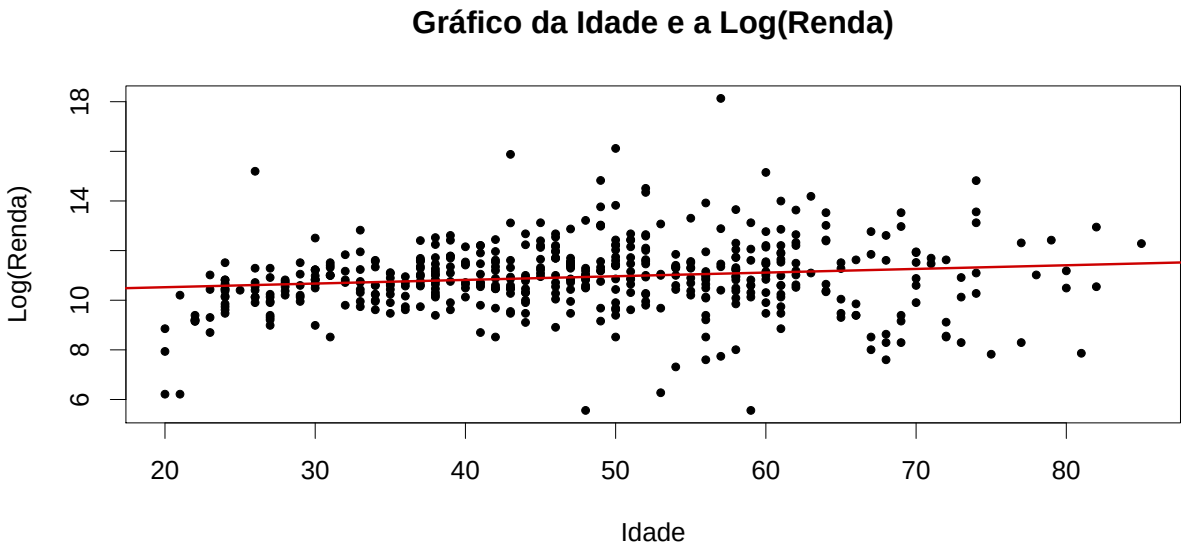


Avaliando os tipos de ensino percebemos uma maior concentração de entrevistados que possuem o ensino médio e o ensino superior, sendo que quase a metade dos entrevistados possuem ensino superior, esse valor corresponde a 248 entrevistados.

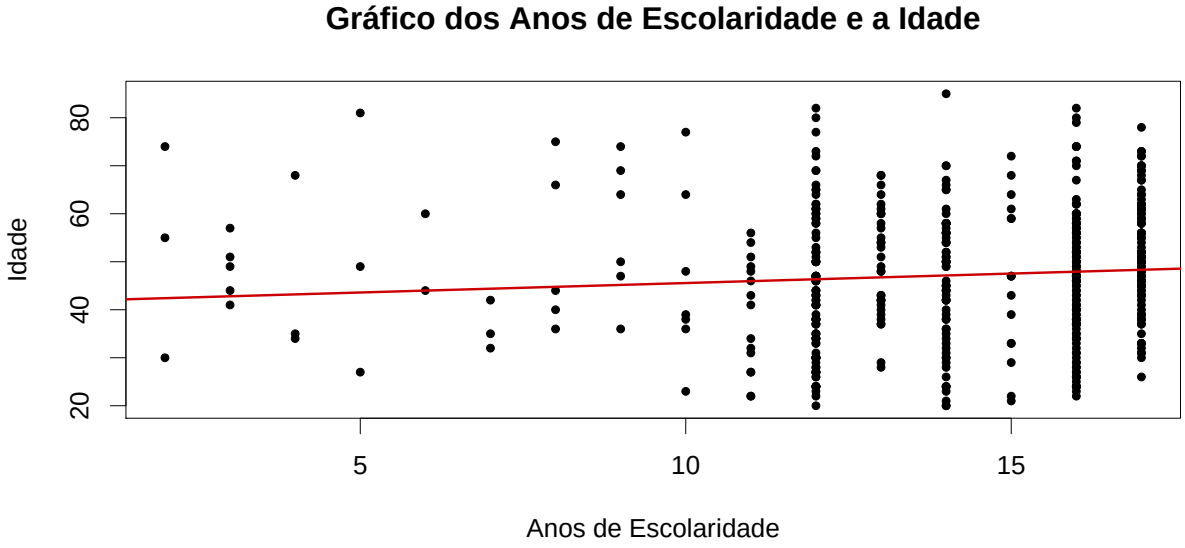
Analisamos também a relação entre a variável resposta (Renda) e as covariáveis presentes no banco de dados escolhido. Assim obtivemos os seguintes resultados:



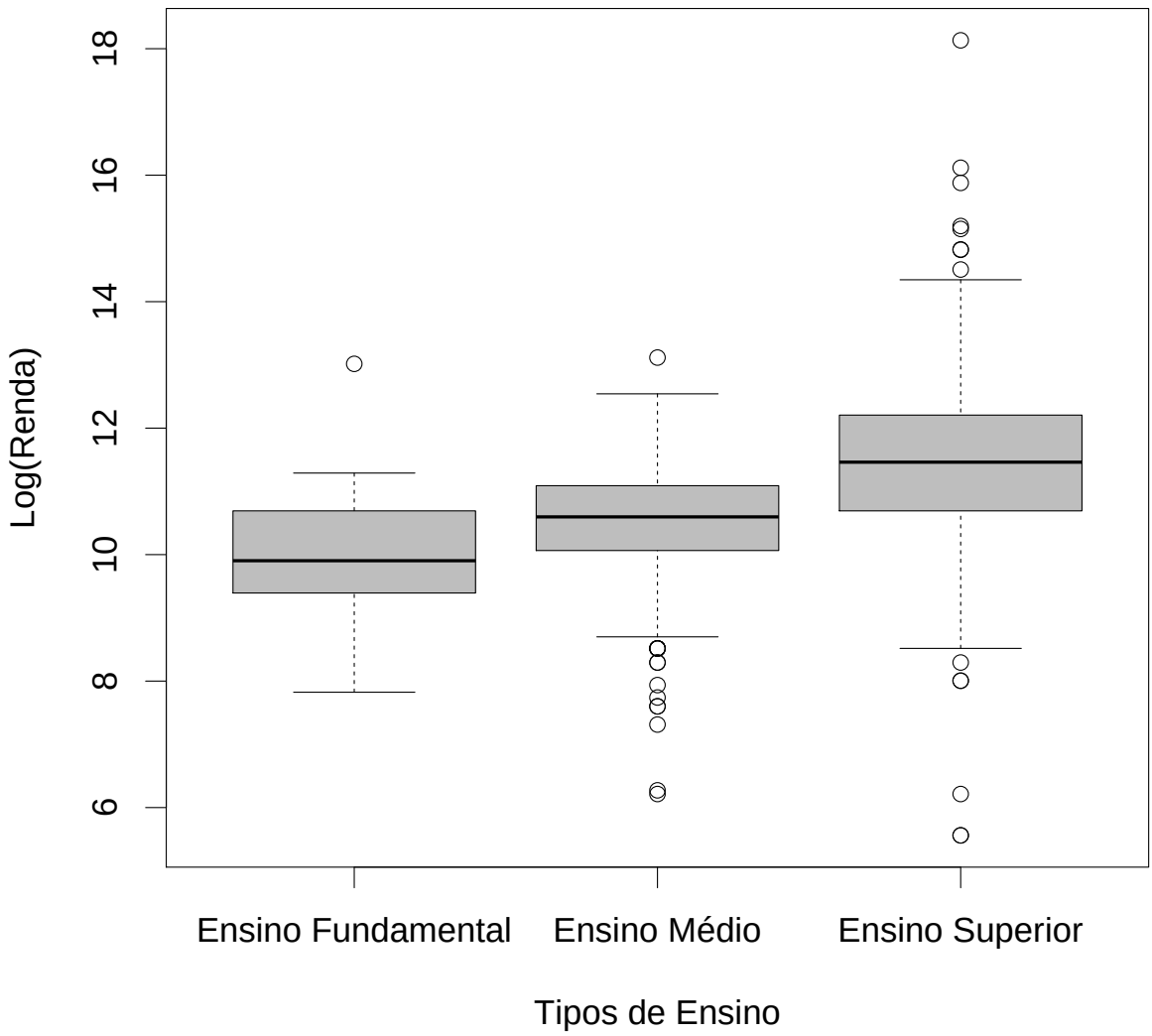
Para as variáveis discretas temos os boxplots da Log(Renda) com cada uma das variáveis separadamente. Para o gênero observamos um valor de renda maior para o sexo masculino, comparando com o sexo feminino; já a variável Estado Civil os entrevistados casados possuem uma renda maior, sendo que a mediana da renda dos que moram juntos com o parceiro(a) e outros estão bastante próximas, porém são inferiores aos valores de renda dos entrevistados casados. Na comparação entre a Log(Renda) e a Etnia percebemos uma amplitude da renda maior para as etnias Branco e Outros, entretanto, como foi observado anteriormente, essas etnias correspondem a 73% e 5%, respectivamente, do total do banco de dados.



Para a relação entre as variáveis Log(Renda) e a Idade temos o gráfico de dispersão acima, nele percebemos uma pequena inclinação no ajuste da curva quando ocorre o aumento das idades dos entrevistados, o que indica um possível ganho de renda anual maior para os entrevistados a medida que aumenta a idade.

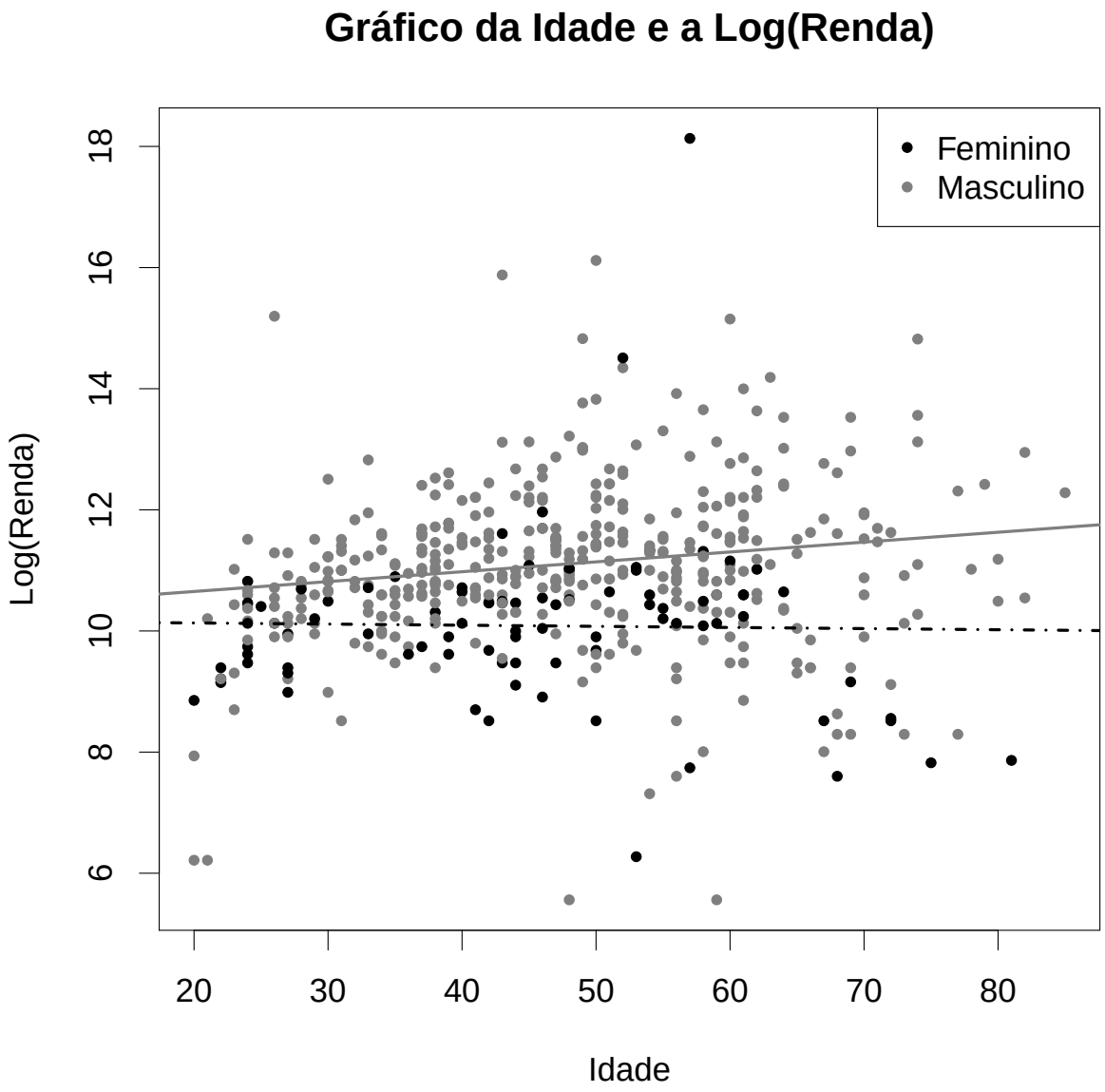


Analisando as variáveis Anos de Escolaridade e a Idade, percebemos uma inclinação quando aumenta os anos de escolaridade e as idades dos entrevistados, ou seja, os entrevistados possuem maior anos de escolaridade à medida que aumentam as idade, o que condiz com a realidade dos ensinos visto anteiormente. Embora quantidades de anos de escolaridade maior possuem grande concentração de pessoas em diversas idades, para os anos de escolaridade entre 2-10 percebemos a presença de variabilidade.

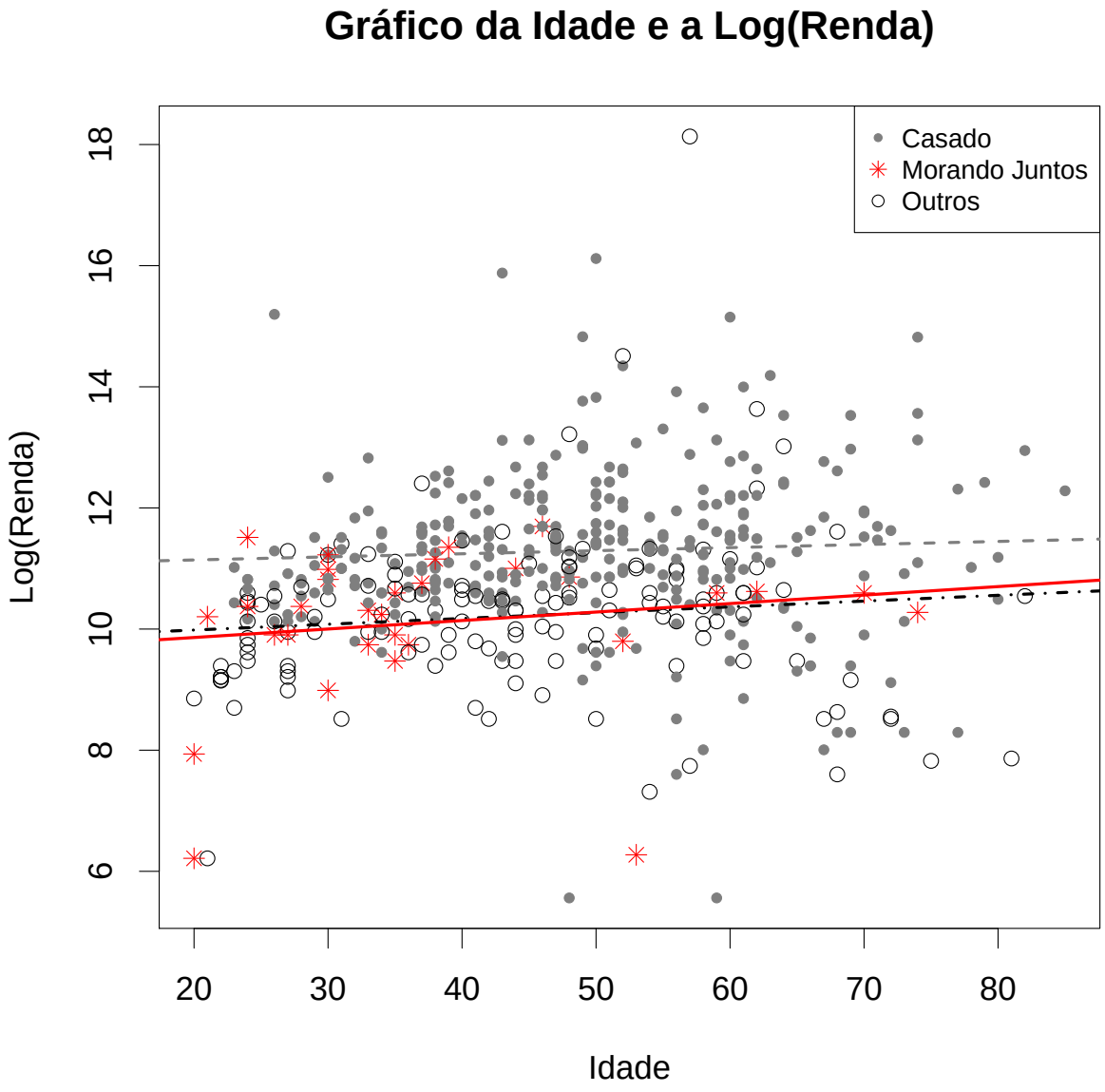


Avaliando a relação entre as variáveis Log(Renda) e a recodificação da variável Anos de Escolaridade, variável separada em tipos de ensino para melhor visualização da relação existente, temos que o tipo de ensino influencia na renda dos entrevistados. Assim observamos valores de renda maiores para o Ensino Superior, que possui de 15-17 anos de escolaridade.

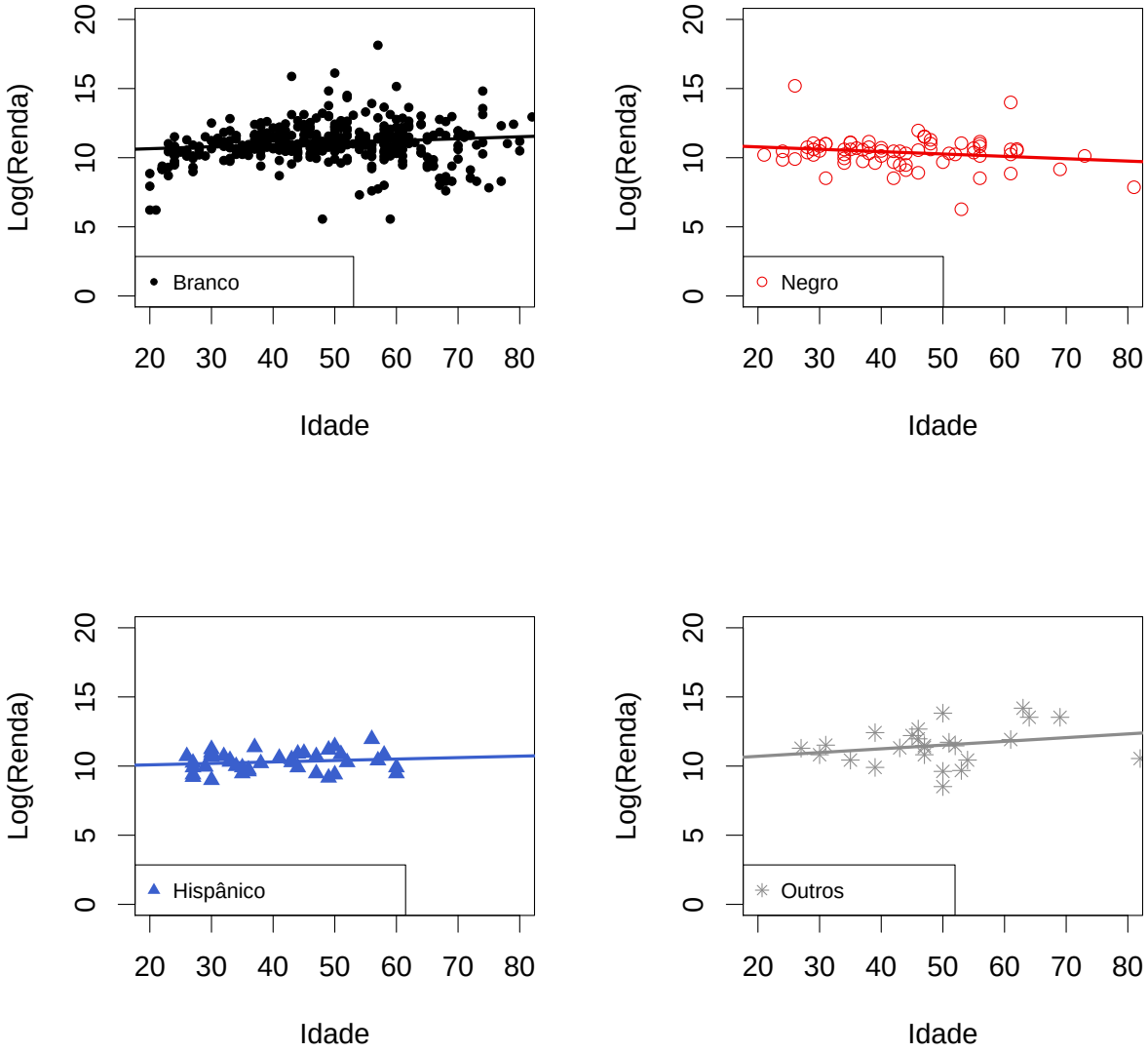
Após a apresentação das variáveis individualmente e em pares com a variável de interesse renda anual, realizamos a análise das variáveis em trios, como por exemplo a Log(Renda), Idade e o Gênero. As análises da relação dessas variáveis permitem fazer suposições sobre os modelos a serem estudados e verificar a influência de cada covariável na variável resposta.



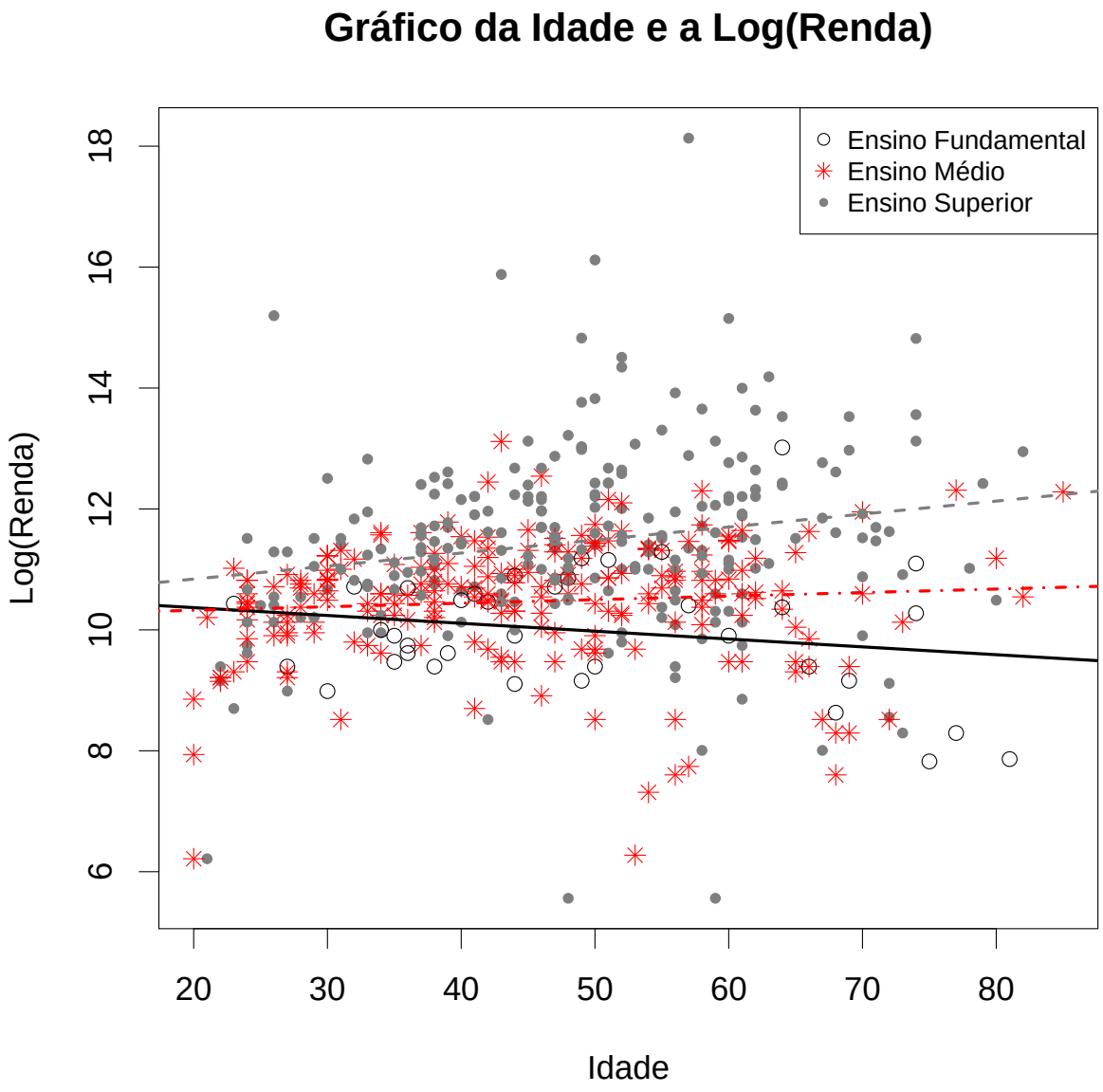
Analisando a relação entre as variáveis Log(Renda), Idade e Gênero percebemos o quanto a idade e o gênero influenciam nos valores da renda anual. O gráfico possui o ajuste das retas e mostra claramente o que foi discutido anteriormente, sobre os valores de renda anual para o sexo masculino serem maiores que o do sexo feminino; podemos perceber também uma inclinação na reta, à medida que aumenta a idade, para o sexo masculino, entretanto para as mulheres essa inclinação é muito pequena ou inexistente.



Para a análise entre as variáveis Log(Renda), Idade e Estado Civil, podemos observar que os entrevistados casados possuem uma quantidade maior de renda e através da inclinação da reta podemos concluir que a renda aumenta através da idade, já para os entrevistados que estão morando junto e os outros tipos de estado civil as retas e a inclinação estão quase juntas, sendo que possuem pequenas diferenças em algumas idades; o gráfico acima também mostra como está a distribuição por estado civil dos entrevistados.



Analisando a relação entre as variáveis Log(Renda), Idade e a Etnia, vemos que as retas pertencentes as etnias Branco, Negro e Outros partem quase do mesmo valor de renda, entretanto, ao longo das idades, possuem comportamentos diferentes para a inclinação da reta, sendo que para a etnia Negro a renda decresce a medida que aumenta a idade. Observamos maior renda para a etnia Outros, os Hispânicos, que começam a reta abaixo das outras etnias, intercepta a etnia Negro entre as idades 40-50 anos. Pelo gráfico da distribuição percebemos que a etnia Hispânico estão concentrados em torno do Log(Renda) igual a 10, e que a etnia Negro possue valores de renda bastante dispersos a medida que aumenta a idade.



Para a análise entre as variáveis Log(Renda), Idade e os tipos de Ensino, confirmamos a conlusão anterior sobre o Ensino Superior possuir renda maior que os outros tipos de ensino. Sendo que nas idades mais jovens a renda para o Ensino Fundamental e o Ensino Médio estão muito próximas quando analisamos a inclinação da reta, entretanto a partir dos 30 anos as retas desses dois tipos de ensino começam a se distanciar; assim há um aumento de renda para o Ensino Médio enquanto o Ensino Fundamental apresenta declínio.

Imputação

O banco de dados escolhido para aplicação da imputação não possui dados faltantes, portanto para avaliar os casos de imputação foi necessário gerar os dados faltantes. Os casos de imputação múltipla existentes são: perda completamente aleatória, perda aleatória e perda não aleatória. Sendo que na perda completamente aleatória o motivo pelo qual os dados estão ausentes não está relacionado às variáveis do estudo; já na perda aleatória a razão para um valor estar ausente está relacionada às outras variáveis observadas, mas não está

relacionada à variável em que há valores ausentes; e por fim na perda não aleatória o motivo pelo qual os dados estão ausentes está diretamente relacionado aos valores não observados da variável de interesse.

Para realização da 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* para comparar os resultados.

Perda Completamente Aleatória

```
##
##  iter imp variable
##   1   1  Income
##   1   2  Income
##   1   3  Income
##   1   4  Income
##   1   5  Income
##   2   1  Income
##   2   2  Income
##   2   3  Income
##   2   4  Income
##   2   5  Income
##   3   1  Income
##   3   2  Income
##   3   3  Income
##   3   4  Income
##   3   5  Income
##   4   1  Income
##   4   2  Income
##   4   3  Income
##   4   4  Income
##   4   5  Income
##   5   1  Income
##   5   2  Income
##   5   3  Income
##   5   4  Income
##   5   5  Income
##   6   1  Income
##   6   2  Income
##   6   3  Income
##   6   4  Income
##   6   5  Income
##   7   1  Income
##   7   2  Income
##   7   3  Income
##   7   4  Income
##   7   5  Income
##   8   1  Income
##   8   2  Income
##   8   3  Income
##   8   4  Income
##   8   5  Income
##   9   1  Income
##   9   2  Income
##   9   3  Income
```

```
##   9   4  Income
##   9   5  Income
##   10   1  Income
##   10   2  Income
##   10   3  Income
##   10   4  Income
##   10   5  Income
##   11   1  Income
##   11   2  Income
##   11   3  Income
##   11   4  Income
##   11   5  Income
##   12   1  Income
##   12   2  Income
##   12   3  Income
##   12   4  Income
##   12   5  Income
##   13   1  Income
##   13   2  Income
##   13   3  Income
##   13   4  Income
##   13   5  Income
##   14   1  Income
##   14   2  Income
##   14   3  Income
##   14   4  Income
##   14   5  Income
##   15   1  Income
##   15   2  Income
##   15   3  Income
##   15   4  Income
##   15   5  Income
##   16   1  Income
##   16   2  Income
##   16   3  Income
##   16   4  Income
##   16   5  Income
##   17   1  Income
##   17   2  Income
##   17   3  Income
##   17   4  Income
##   17   5  Income
##   18   1  Income
##   18   2  Income
##   18   3  Income
##   18   4  Income
##   18   5  Income
##   19   1  Income
##   19   2  Income
##   19   3  Income
##   19   4  Income
##   19   5  Income
##   20   1  Income
##   20   2  Income
```

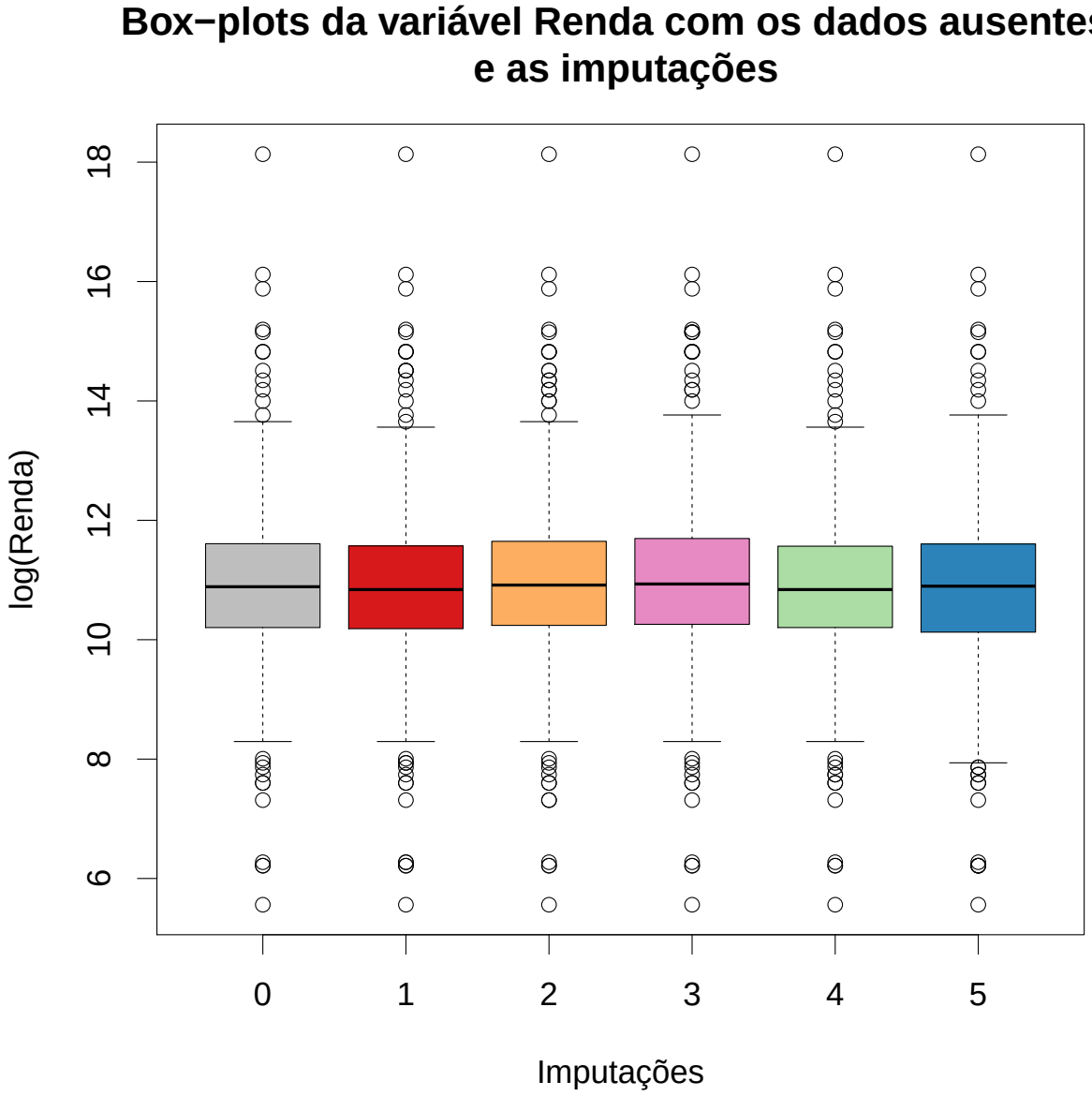
```
##   20   3  Income
##   20   4  Income
##   20   5  Income
```

Como dito anteriormente o banco de dados utilizado nesse estudo não possui dados ausentes. Para tanto o mecanismo de Perda Completamente Aleatória consiste em que o motivo pelo qual os dados estão ausentes não está relacionado as variáveis do estudo. Sendo assim para esse mecanismo utilizamos uma distribuição bernoulli com probabilidade de sucesso de 0,20 para gerar os dados ausentes na variável Renda, e fixamos uma semente ao gerar os números aleatórios.

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

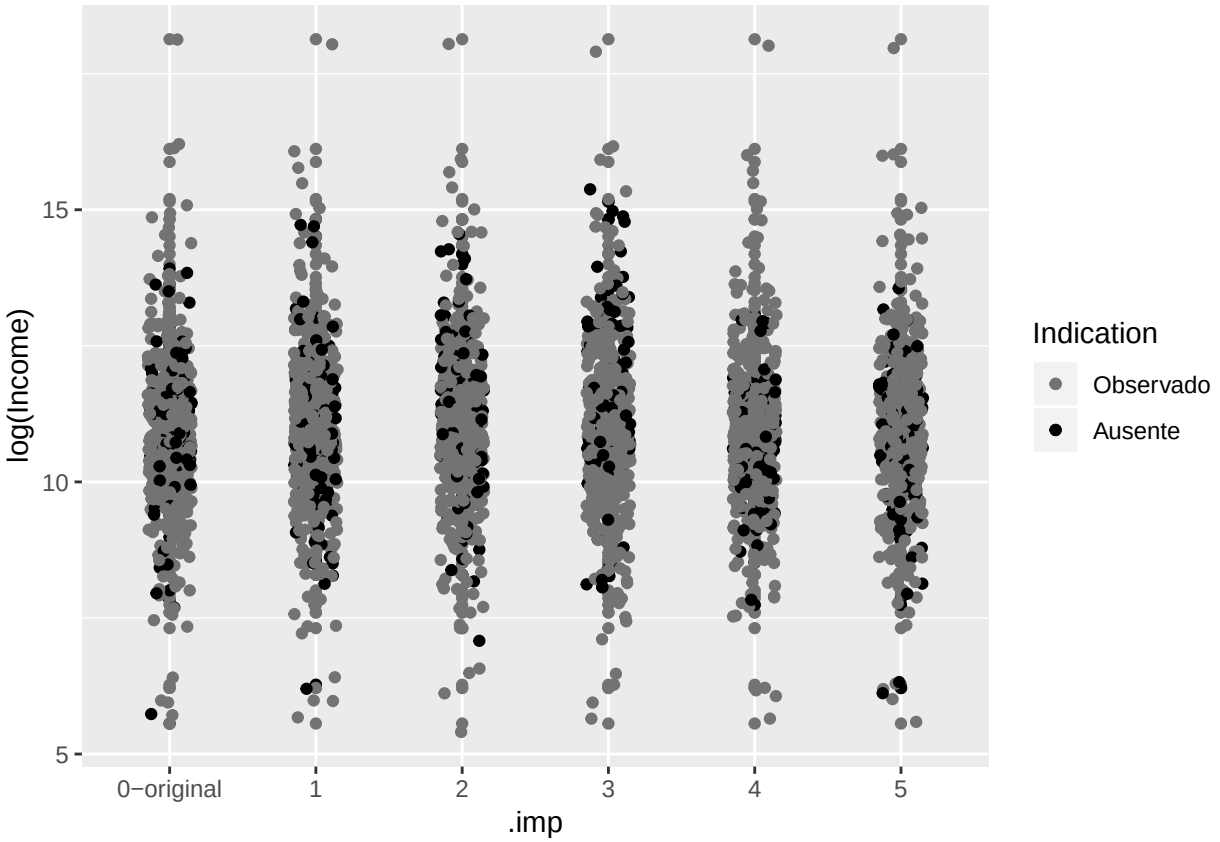
```
##      Gender Age        MarStat Education Ethnicity Income
## 1 Masculino  30         Casado        16 Hispânico     NA
## 2 Masculino  50         Casado         9 Hispânico  12000
## 3 Masculino  39         Casado        16    Branco 120000
## 4 Masculino  43         Casado        17    Branco  40000
## 5 Masculino  61         Casado        15    Branco     NA
## 6 Masculino  34 Morando Juntos        11     Negro  28000
##           Education2
## 1    Ensino Superior
## 2 Ensino Fundamental
## 3    Ensino Superior
## 4    Ensino Superior
## 5    Ensino Superior
## 6       Ensino Médio
```

Pelos box-plots abaixo podemos verificar os valores da Renda com os dados ausentes e as imputações geradas.

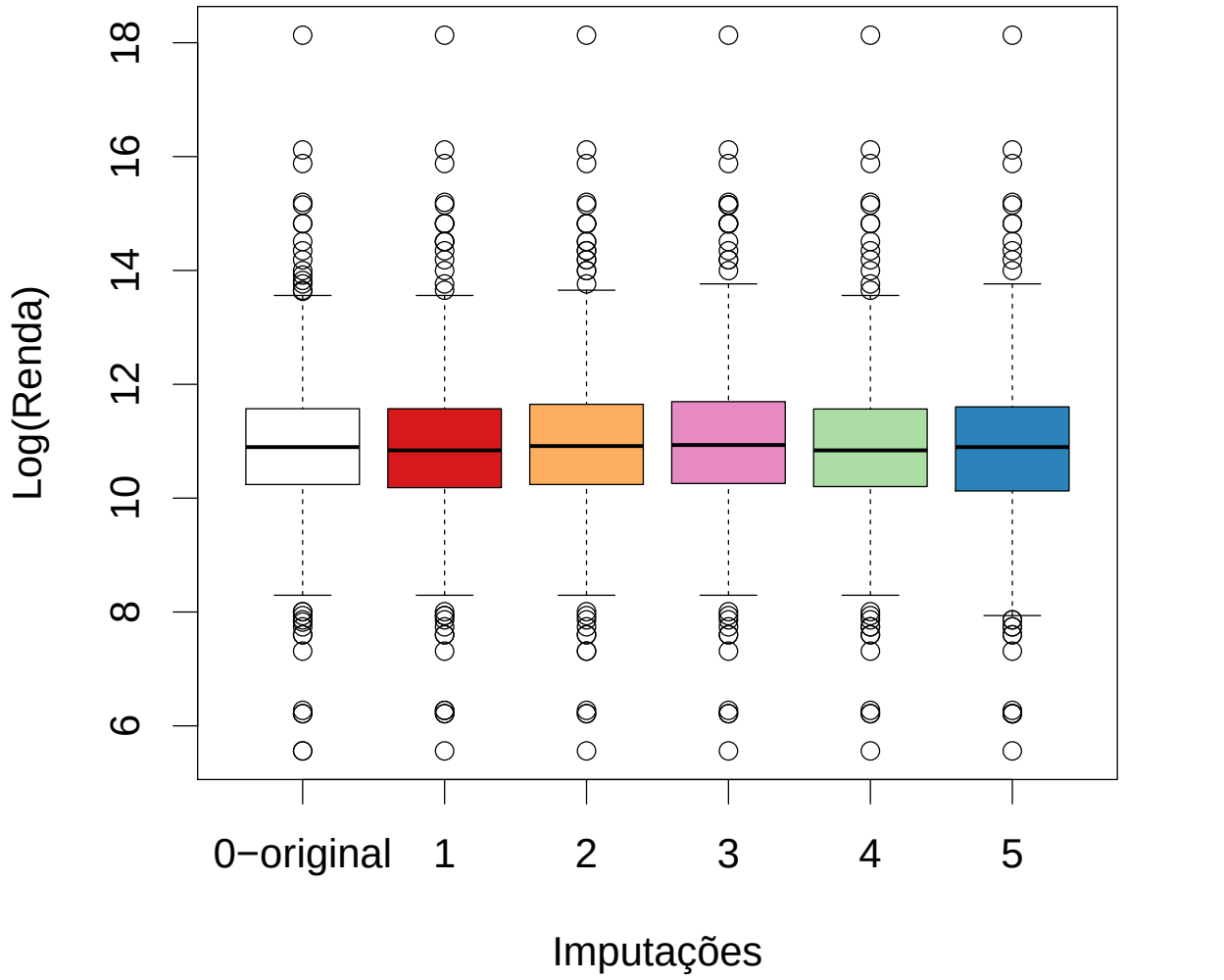


Nos box-plots acima percebemos que as 5 imputações geradas seguem de modo semelhante a distribuição do banco com os valores ausentes.

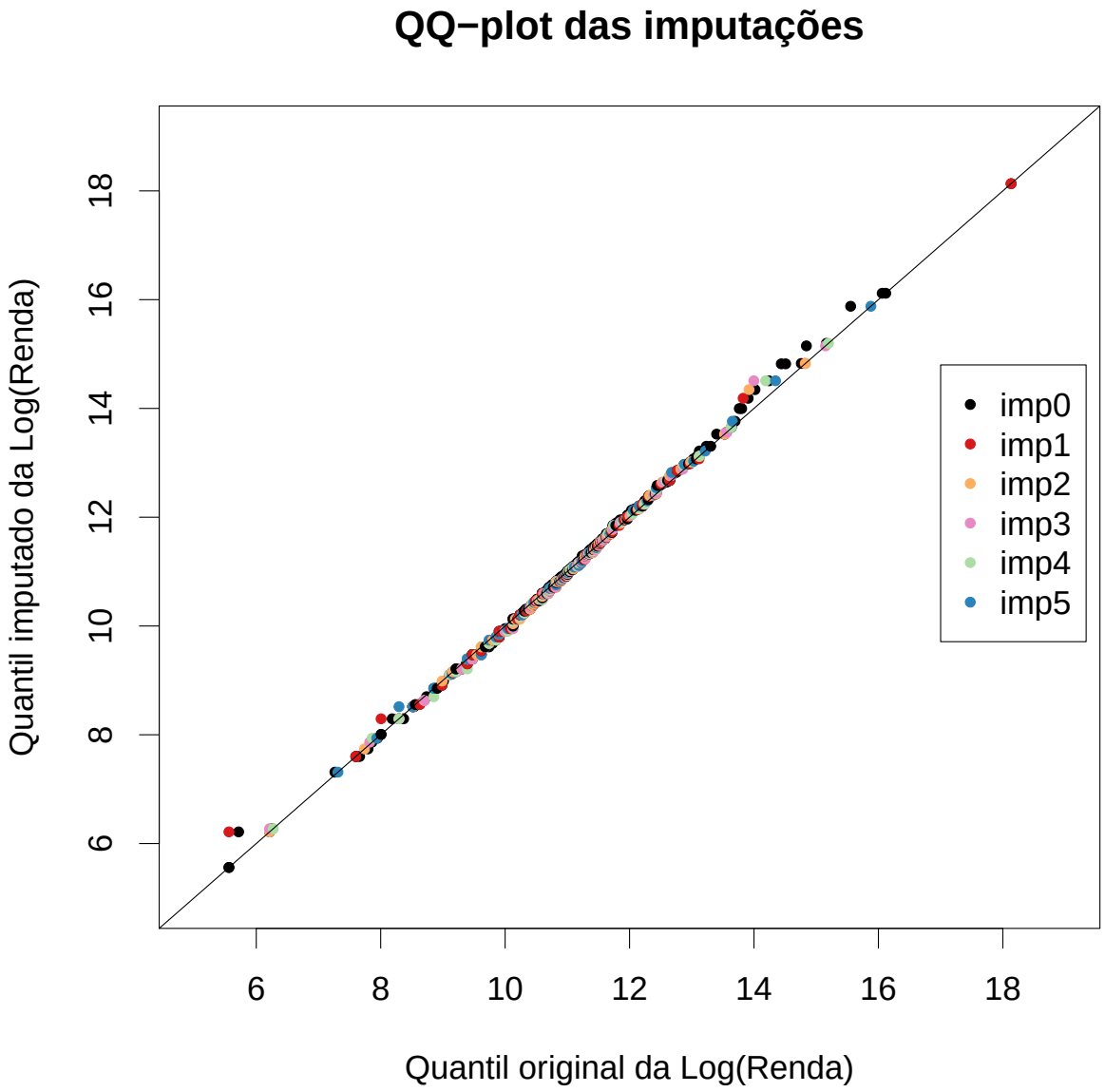
Como possuímos o banco de dados original realizamos a análise dos valores da renda no banco original com os valores da renda das imputações, e assim obtemos a seguinte distribuição:



Pelo gráfico de dispersão acima, temos que os valores orignais e os valores imputados estão bastante próximos, o que pode ser verificado também pelos box-plots abaixo, que analisa o banco de dados original e as imputações:

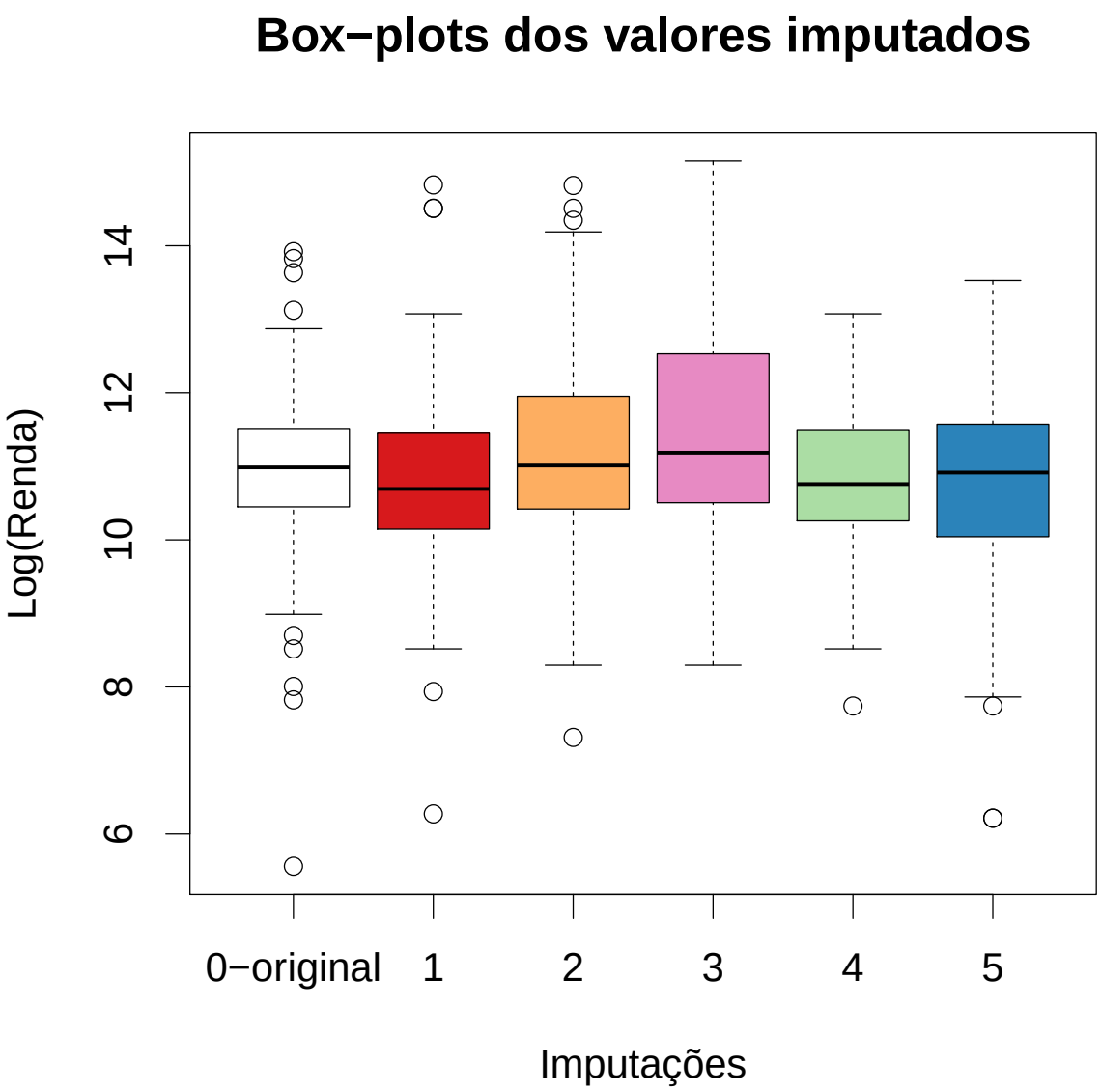


Para realizar a adequação do modelo temos o gráfico QQ-plot abaixo:



Pelo QQ-plot podemos chegar a adequação do modelo dos dados com as observações faltantes com as imputações realizadas. Nele vemos que os valores estão bastante concentrados indicando que os valores imputados são adequados.

Como dito anteriormente, para gerar o mecanismo de perda completamente aleatória, criamos os valores ausentes na variável renda no banco de dados original, das 500 observações presentes no banco de dados o mecanismo gerou 96 observações ausentes. Portanto analisamos somente essas observações originais codificados como ausentes com os valores que foram imputados para elas. Assim temos abaixo os box-plots dessas 96 observações:



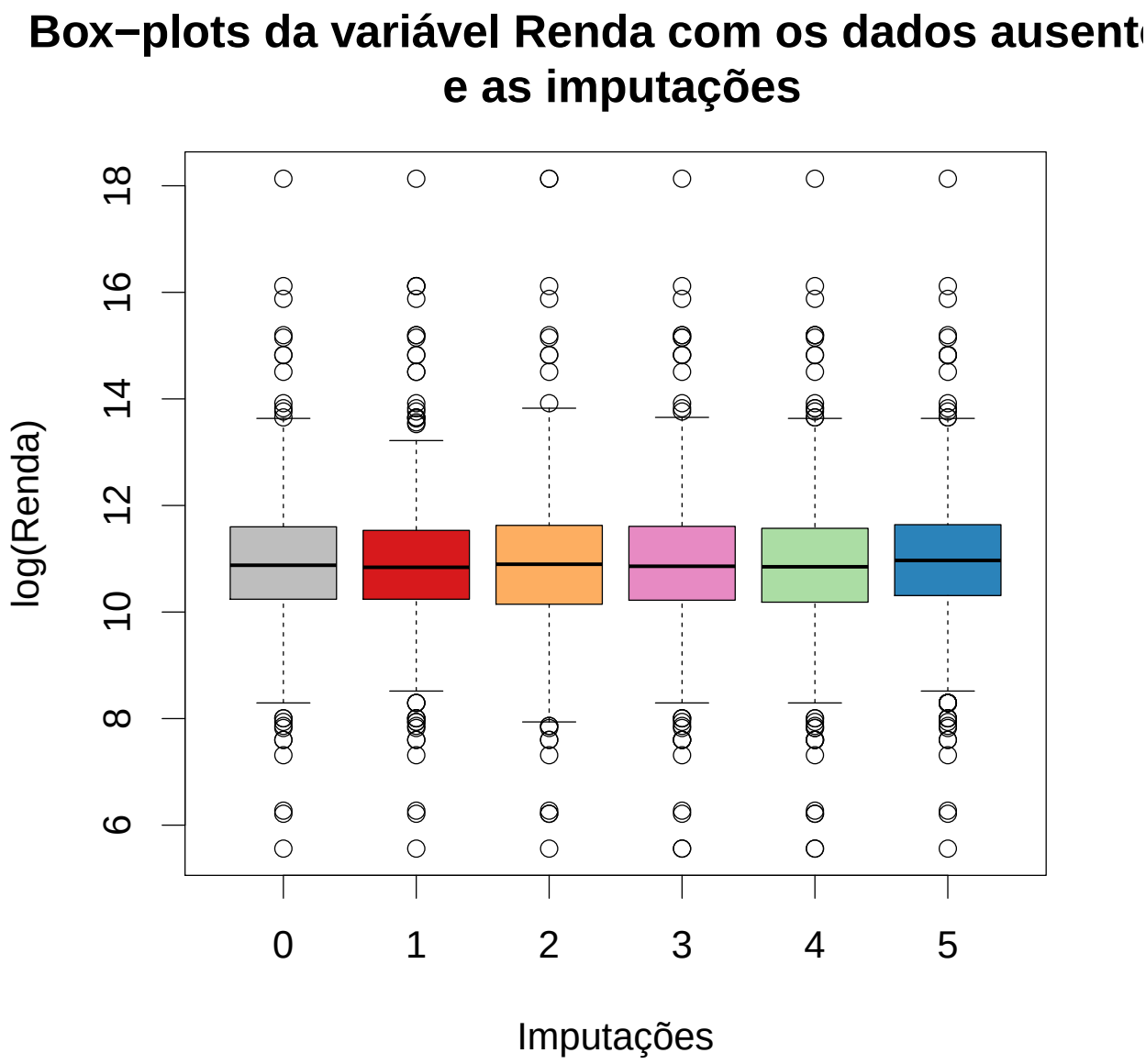
Perda Aleatória

O mecanismo de perda aleatória consiste em que a razão para um valor estar ausente está relacionada às outras variáveis do observadas, mas não está relacionada à variável em que há valores ausentes. Para gerar os dados ausentes desse mecanismo avaliamos a perda da variável Renda conjuntamente com as variáveis Gênero e Tipo de Ensino.

Gênero

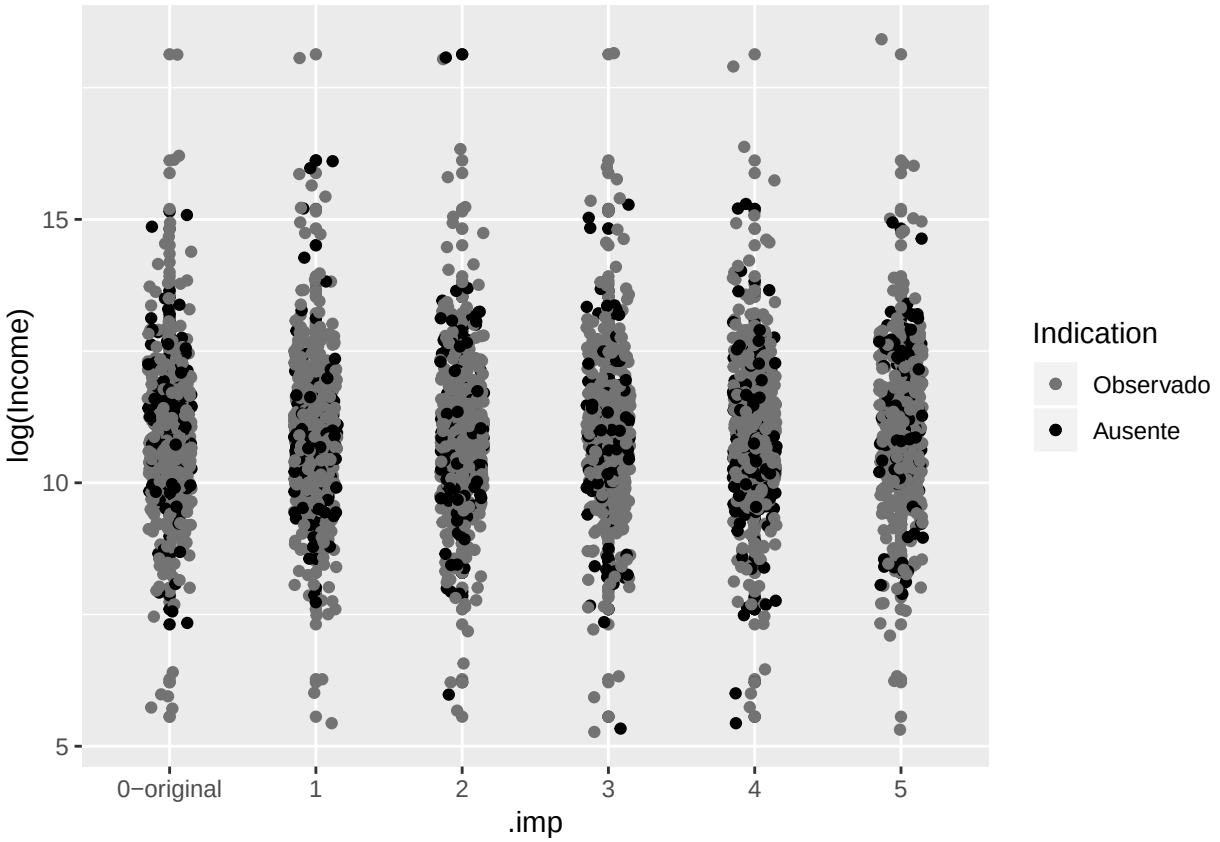
Para gerar os dados ausentes na renda, pelo caso de perda aleatória conjuntamente com o gênero, utilizamos uma distribuição bernoulli com probabilidade de sucesso de 0,10 para o sexo feminino e 0,30 para o sexo masculino, e fixamos uma semente ao gerar os números aleatórios.

Os box-plots abaixo indicam como estão a distribuição dos valores da renda para o banco de dados com valores ausentes na renda e as imputações:



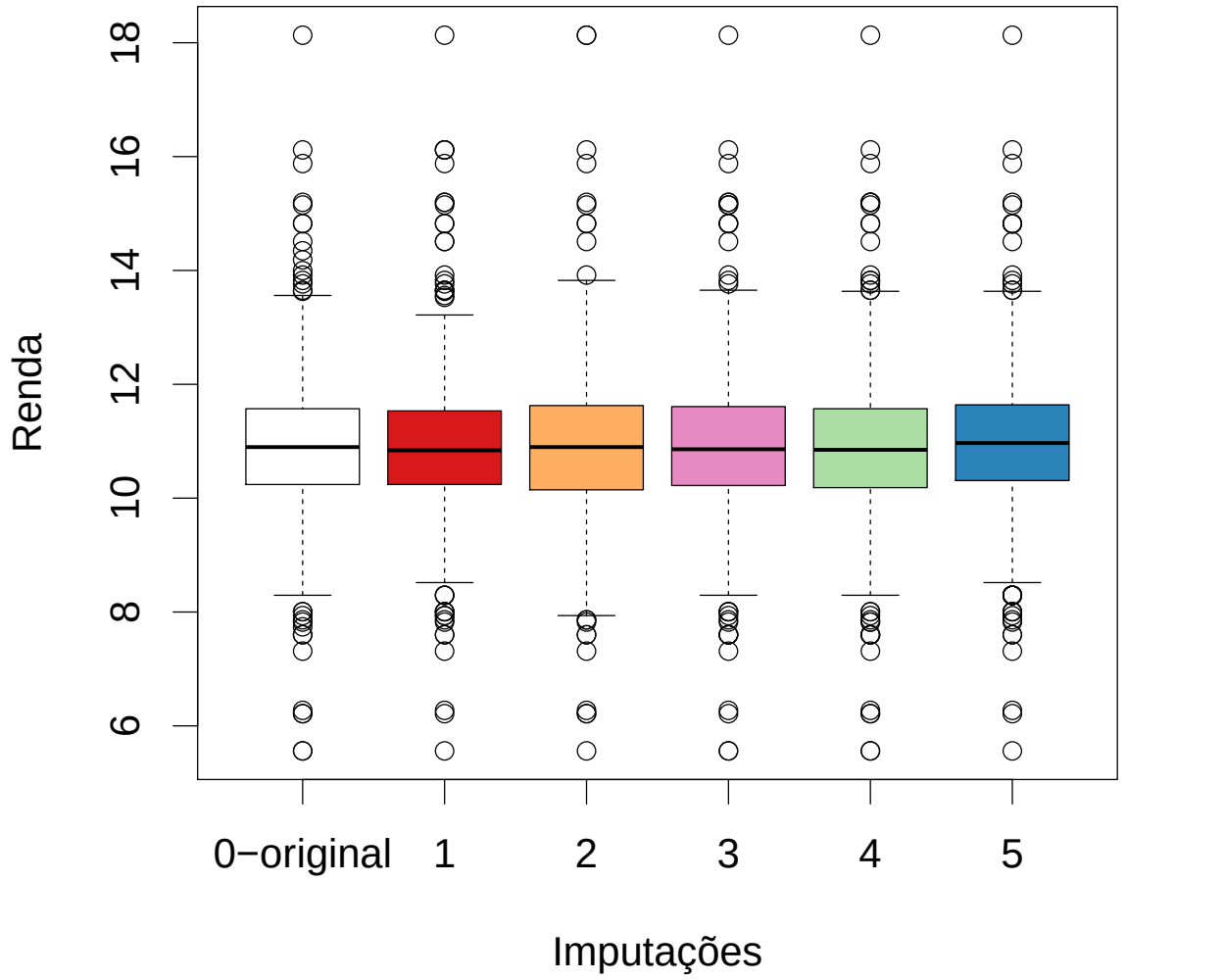
Pelos box-plots podemos verificar que há pequenos desvios entre eles, porém as imputações possuem bastante proximidade com o banco de dados com valores faltantes.

Como possuímos o banco de dados original realizamos a análise dos valores da renda no banco original com os valores da renda das imputações, e assim obtemos a seguinte distribuição:

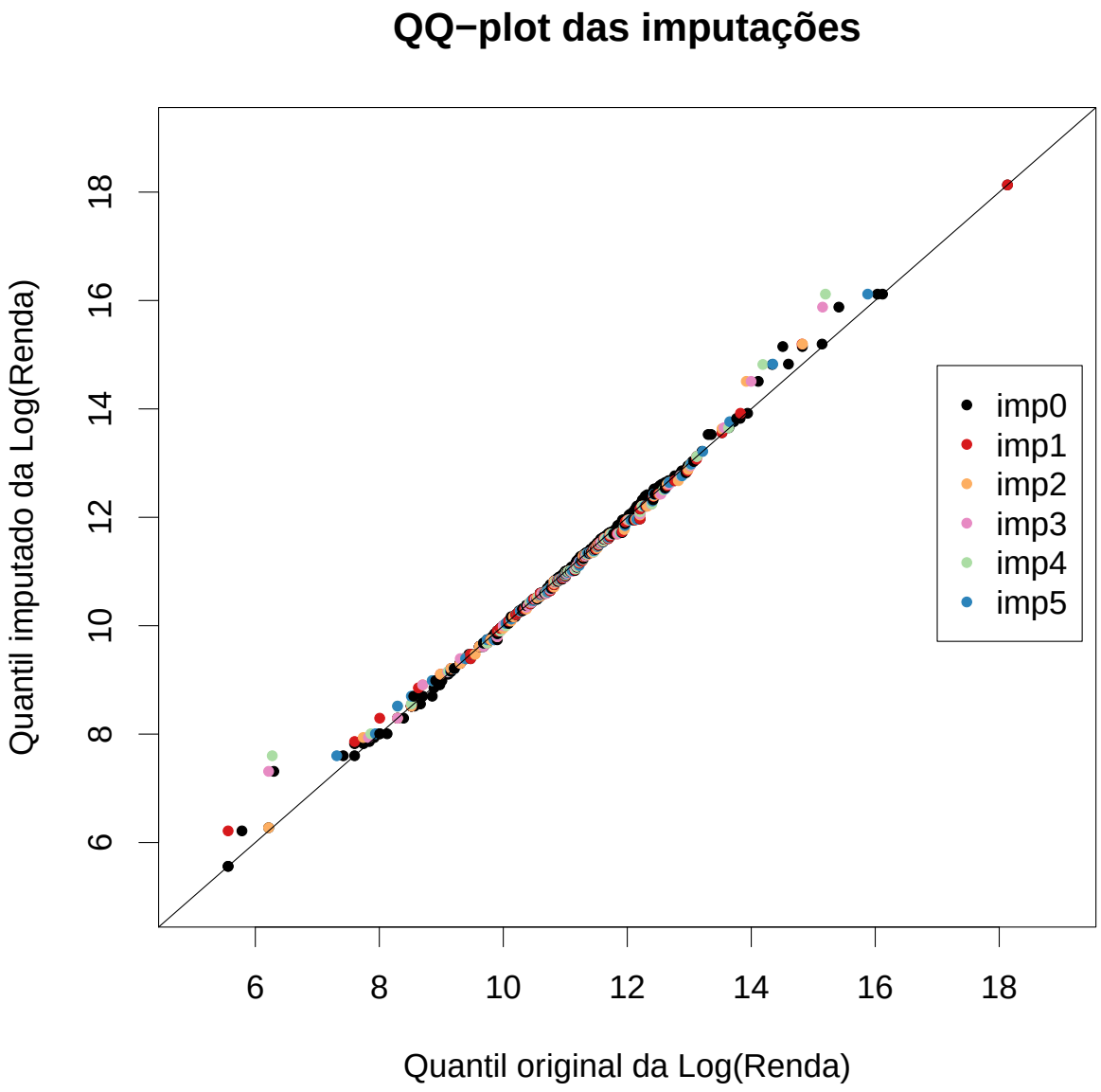


Pelo gráfico de dispersão acima percebemos melhor a distribuição verificada no gráfico dos box-plots anterior, novamente observamos indícios de pequenos desvios entre o banco de dados original e o banco de dados imputação. Nas imputações 2, 3 e 4 percebemos imputações de valores da log(renda) próximos de 5 sendo que no banco de dados original não encontramos tais valores ausentes ao redor do log(renda) igual a 5.

Para verificar esses pequenos desvios nas imputações com o banco original temos os box-plot abaixo:

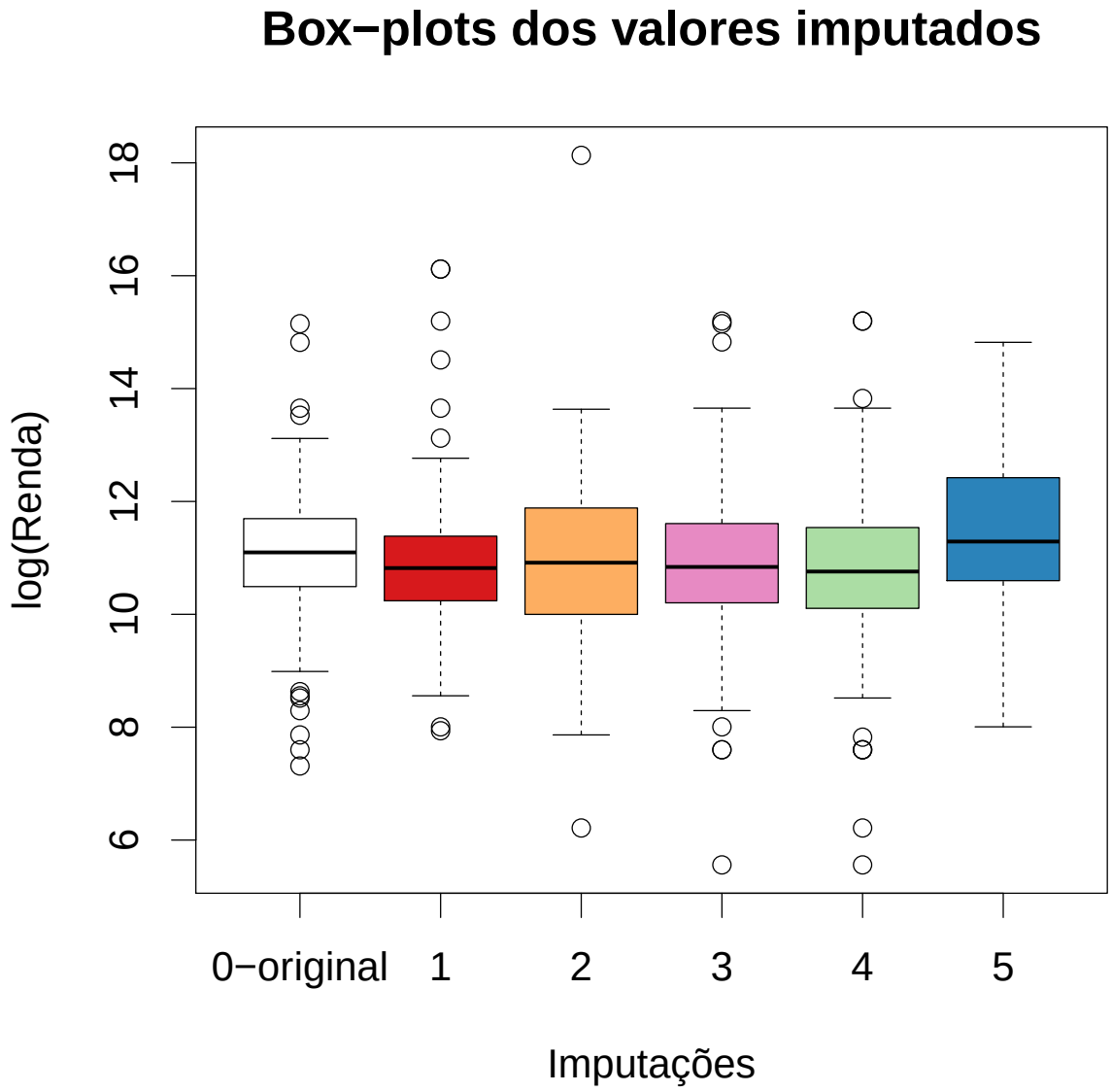


Por esses box-plots percebemos ainda o indício de pequenos desvios nas imputações comparada com o banco original. Portanto abaixo temos o QQ-plot no intuito de checar a adequação do ajuste:



Pelo QQ-plot acima percebemos para valores menores que o log(renda) de 10 e maiores que o log(renda) de 14, um deslocamento maior entre as imputações e a reta de adequação do modelo entre os quantis originais e o quantis imputados da log(renda).

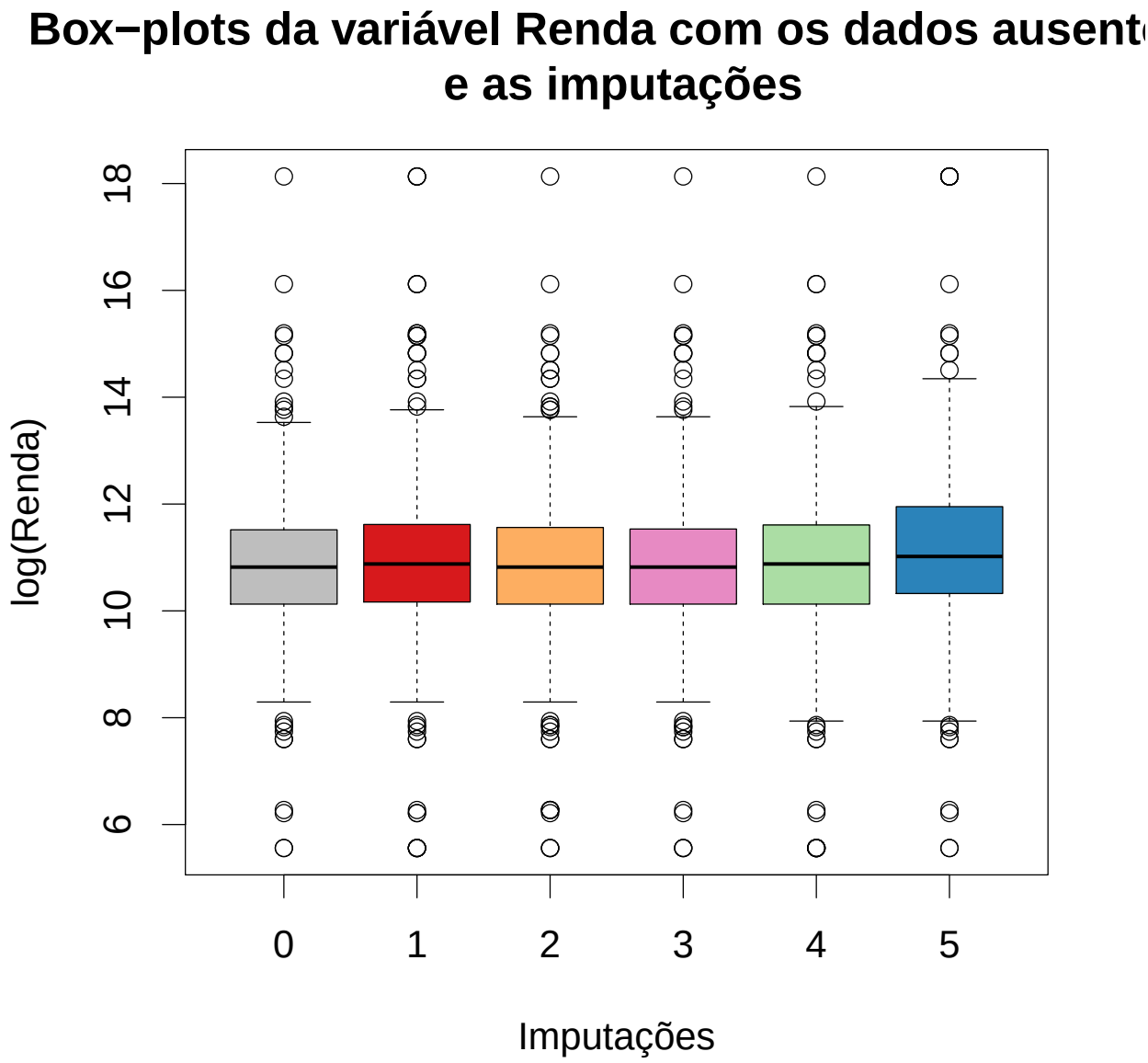
Como dito anteriormente, para gerar o mecanismo de perda aleatória conjuntamente com o gênero, criamos os valores ausente na variável renda no banco de dados original, das 500 observações presentes no banco de dados o mecanismo gerou 127 observações ausentes. Portanto analisamos somente essas observações originais codificadas para ausentes com os valores que foram imputados para elas. Assim temos abaixo os box-plots dessas 127 observações:



Tipo de Ensino

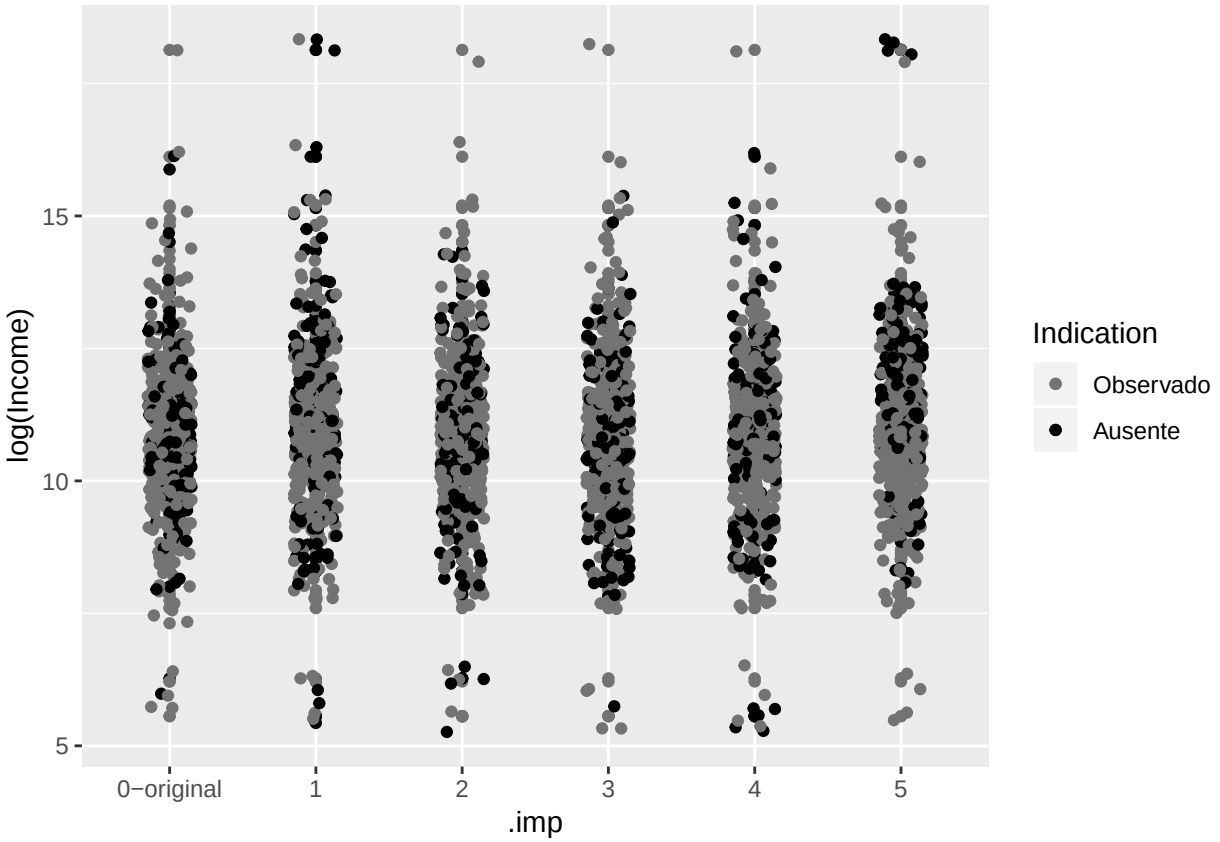
Para gerar os dados ausentes na renda, pelo caso de perda aleatória conjuntamente com o tipo de ensino, utilizamos uma distribuição bernoulli com probabilidade de sucesso de 0,05 para o ensino fundamental, 0,20 para o ensino médio e 0,40 para o ensino superior, e fixamos uma semente ao gerar os números aleatórios.

Os box-plots abaixo estão indicando a distribuição entre os valores da renda com dados faltantes e as 5 imputações realizadas:



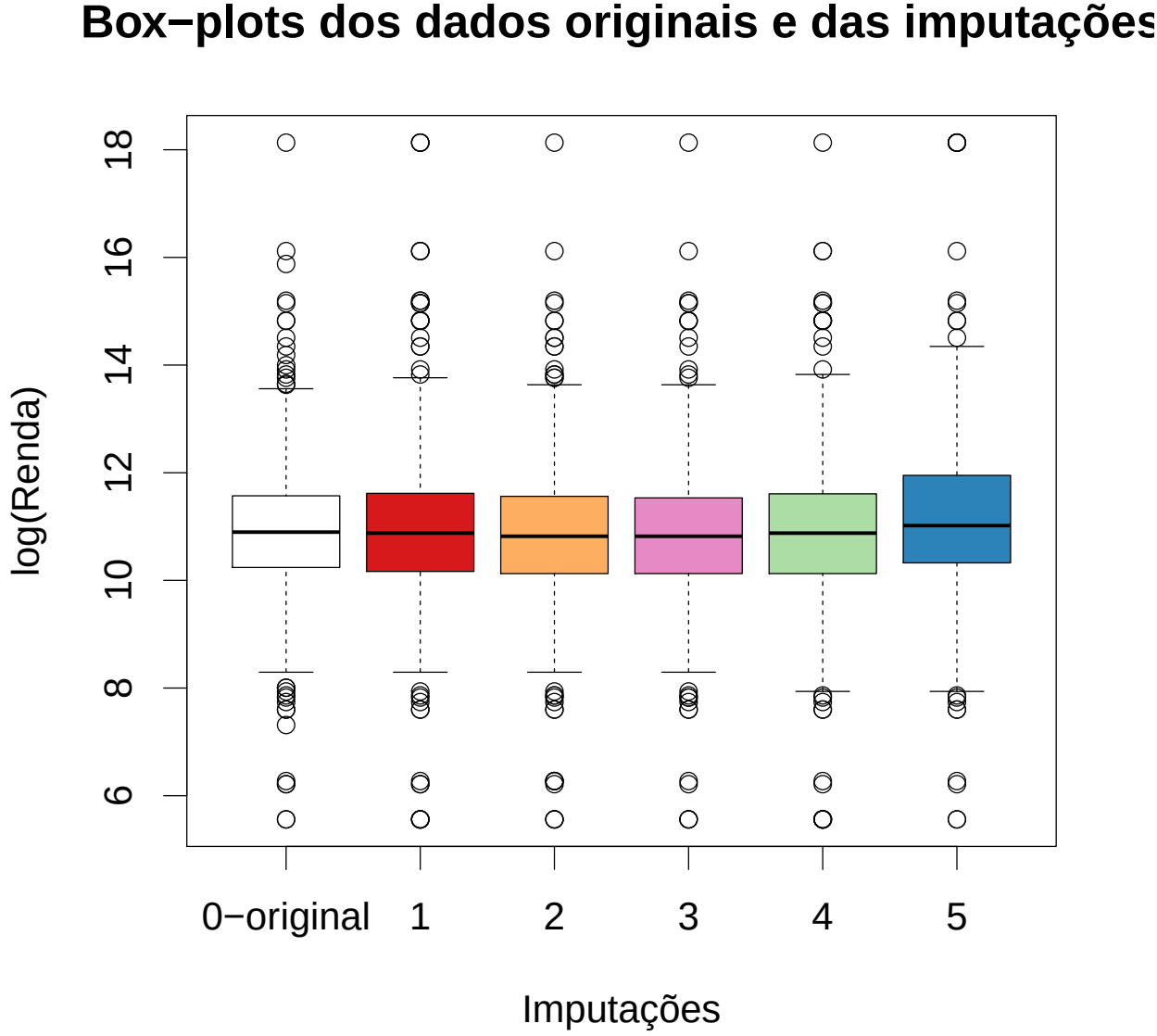
Pelos box-plots podemos perceber pequenas variações, a imputação nº 5 que apresentou maior divergência comparada com os outros box-plots.

Como temos o banco de dados com os valores originais da renda, o gráfico abaixo demonstra a distribuição dos pontos para os dados originais e os imputados da renda:



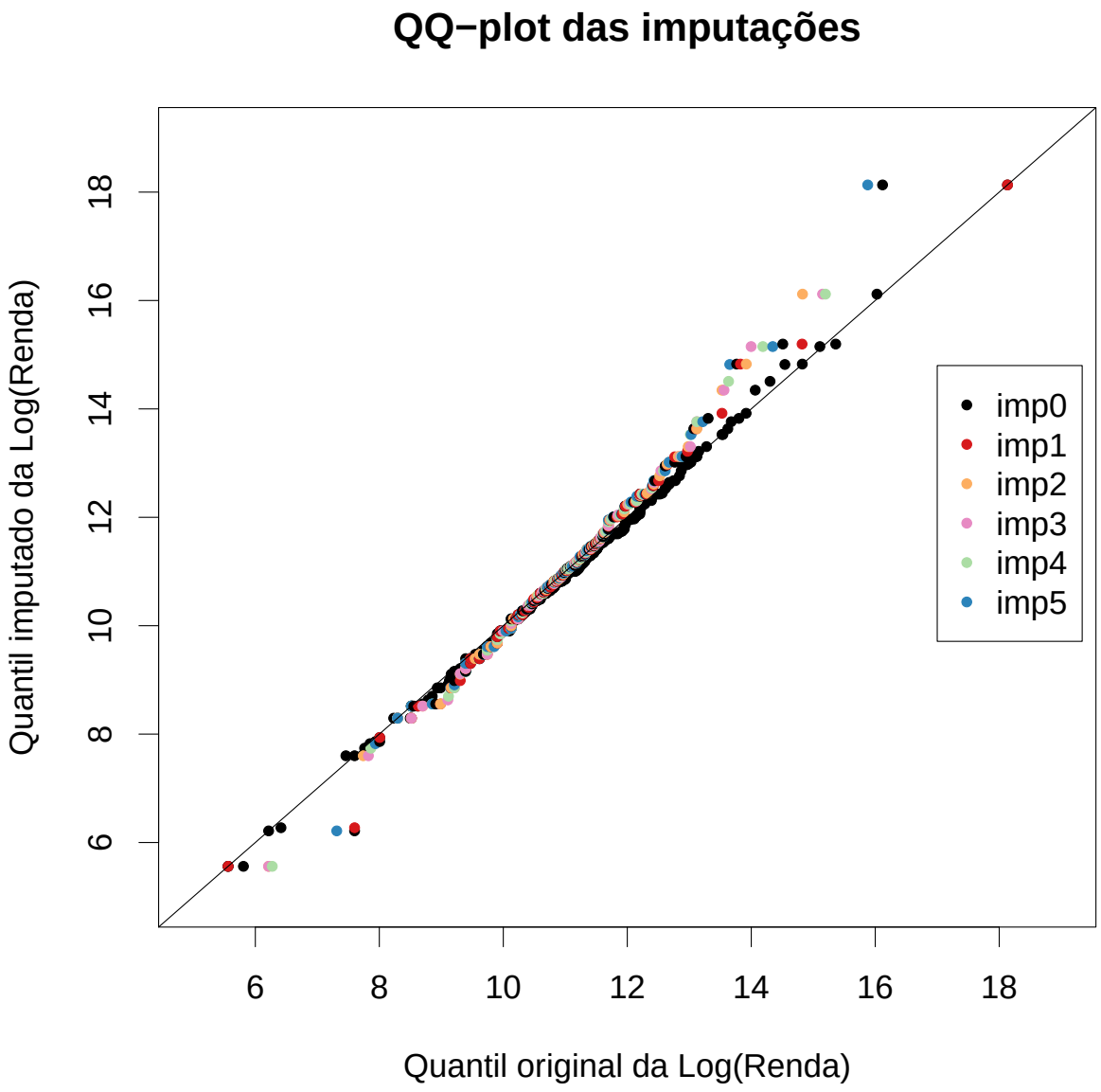
Pelo gráfico de dispersão percebemos que a imputação nº 5 não possui valores imputados para rendas menores, sendo que a original possui valores pequenos de renda, e as outras 4 imputações obtiveram valores imputados para rendas ao redor do valor da Log(renda) igual a 5. Percebemos também que somente as imputações de nº 1 e 5 possuem valores altos de renda imputados. É interessante observar o comportamento da imputação de nº 5, dado que não segue exatamente a distribuição das outras imputações, uma vez que não houve imputação de valores para Log(renda) ao redor de 5 e houveram imputações para Log(renda) ao redor de 15.

Para analisar as variações mencionadas anteriormente, temos abaixo os box-plots do banco de dados com os valores originais da renda e o banco de dados com os valores imputados da renda:



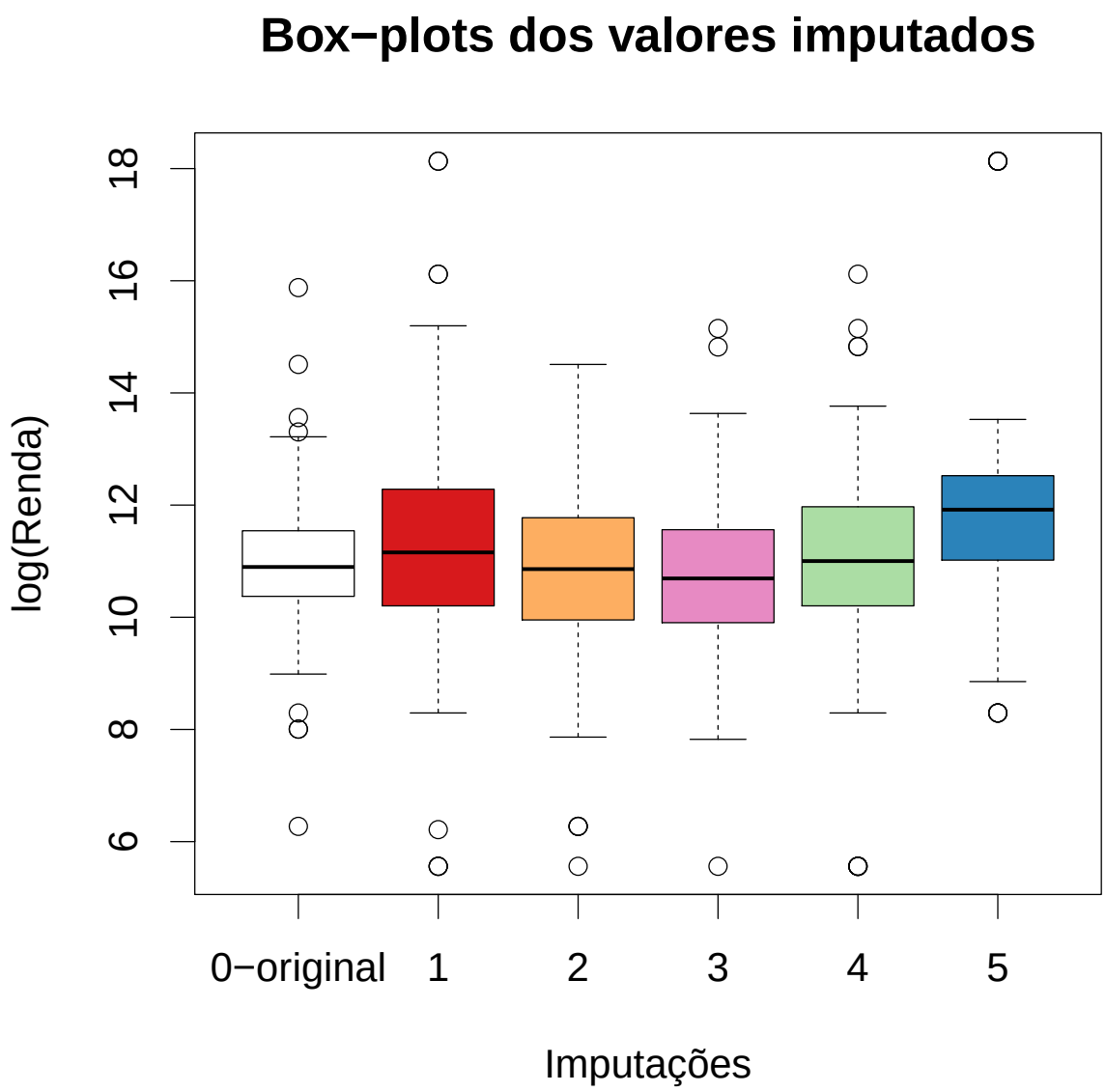
Pelos box-plots confirmamos a variação maior encontrada na imputação nº 5, quando está é comparada com os dados originais da renda.

Abaixo temos o QQ-plot:



O intuito principal do QQ-plot acima é checar a adequação da distribuição. Por ele percebemos deslocamentos do ajuste da reta para valores de Log(renda) menores que 8 e acima de 13, indicando que as variações encontradas estão interferindo na adequação do modelo.

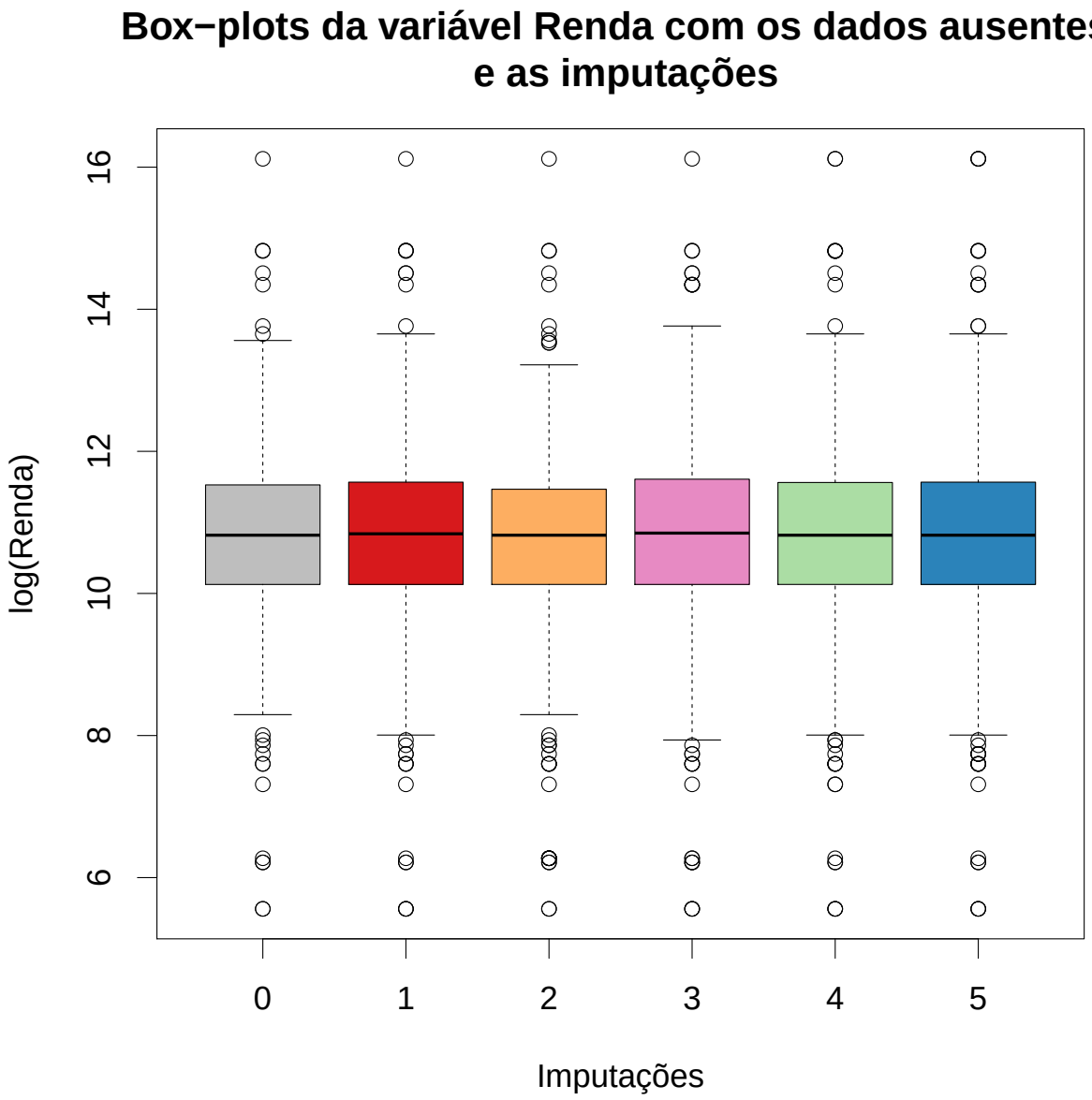
Como dito anteriormente, para gerar o mecanismo de perda aleatória conjuntamente com o tipo de ensino, criamos os valores ausente na variável renda no banco de dados original, das 500 observações presentes no banco de dados o mecanismo gerou 137 observações ausentes. Portanto analisamos somente essas observações originais codificadas para ausentes com os valores que foram imputados para elas. Assim temos abaixo os box-plots dessas 137 observações:



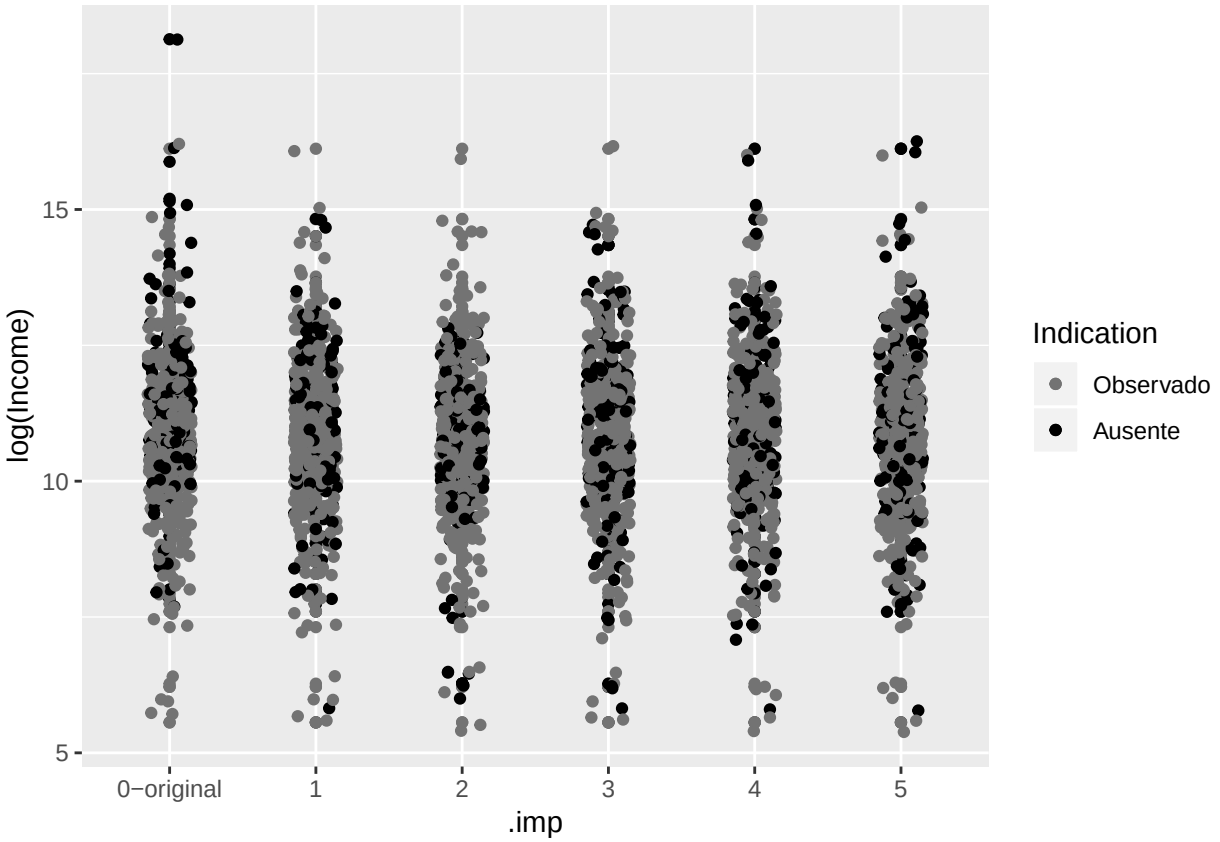
Perda Não Aleatória

O mecanismo de perda não aleatória consiste no motivo pelo qual os dados estão ausentes está diretamente relacionado aos valores não observados da variável de interesse. Para gerar os dados ausentes desse mecanismo avaliamos a perda da variável Renda através da função logit ..... (colocar fórmula aqui), em que utilizamos os valores do mínimo e do máximo da renda anual, observados no banco de dados original, para solucionar o sistema de equações e encontrar os betas (colocar fórmula dos betas aqui). Após encontrar as probabilidades através da função aplicamos nos dados da renda para gerar os valores ausentes.

Os box-plots da renda com valores ausentes e as 5 imputações realizadas podem ser vistos abaixo:

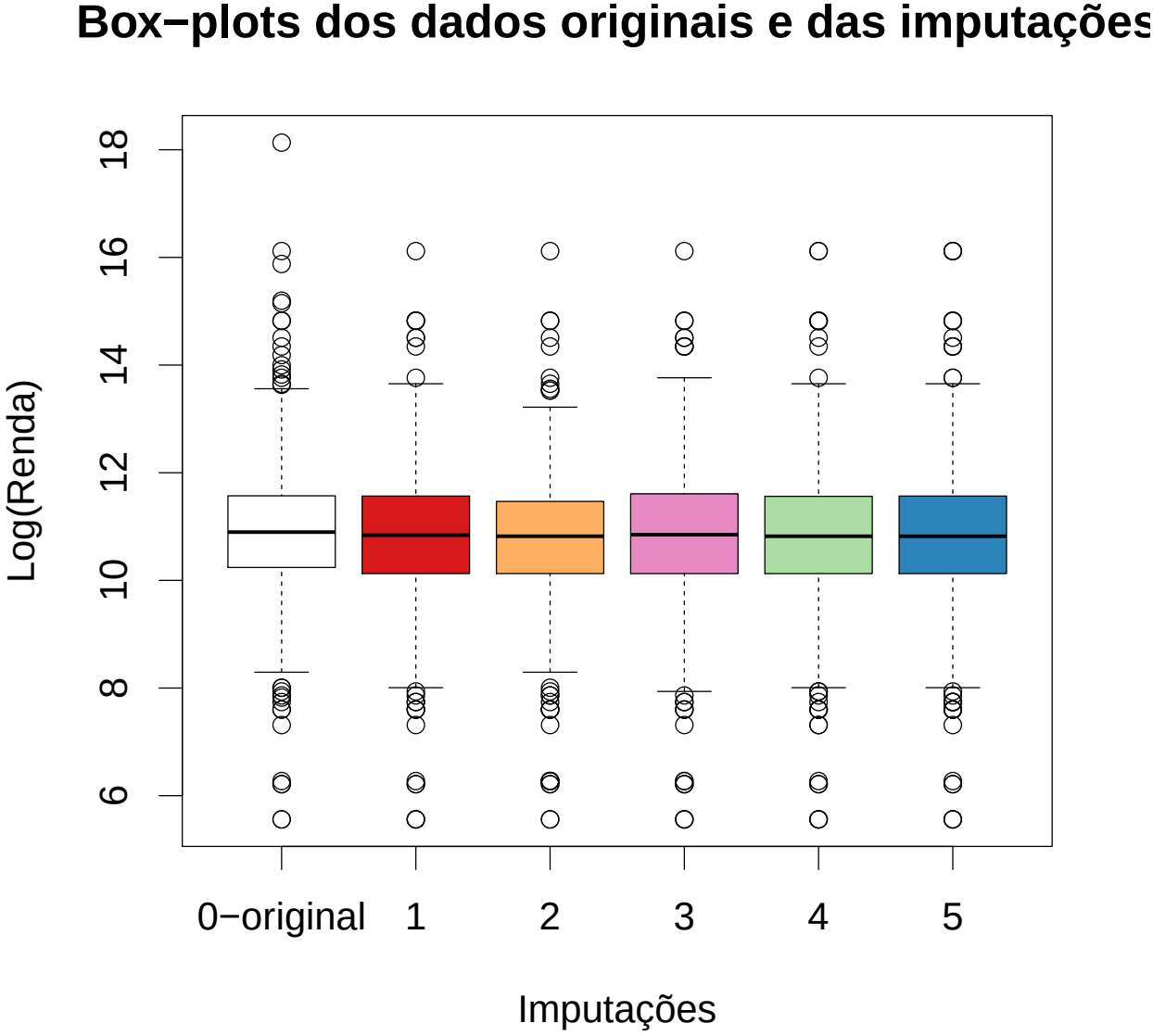


Nos box-plots há indícios de variações nas imputações. Portanto, temos abaixo o gráfico de dispersão para os valores originais da renda, que foram codificados como ausentes, e os valores imputados da renda:



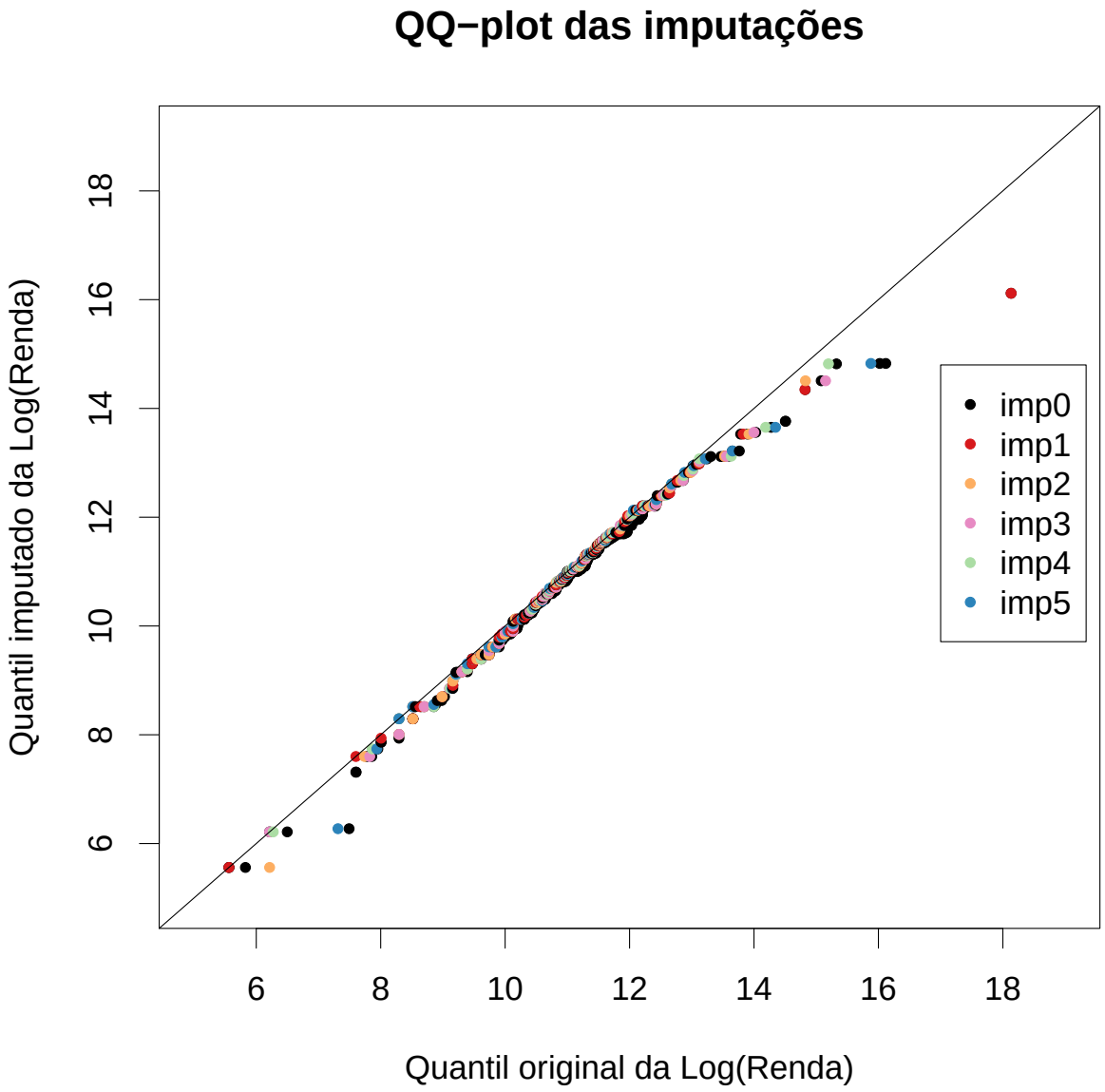
Pelo gráfico de dispersão acima comprovamos alguns indícios de variações, temos que os dados originais não posseum codificação de dados ausentes para log(renda) ao redor de 5, e as imputações obtiveram valores imputados próximos desse valor, vemos também que a imputação nº 2 a partir de aproximadamente log(renda) igual a 13 não há mais imputações, sendo que nos dados originais percebemos codificação de variável ausente acima desse valor de renda.

Temos abaixo os box-plots para os valores originais e os imputados da renda:



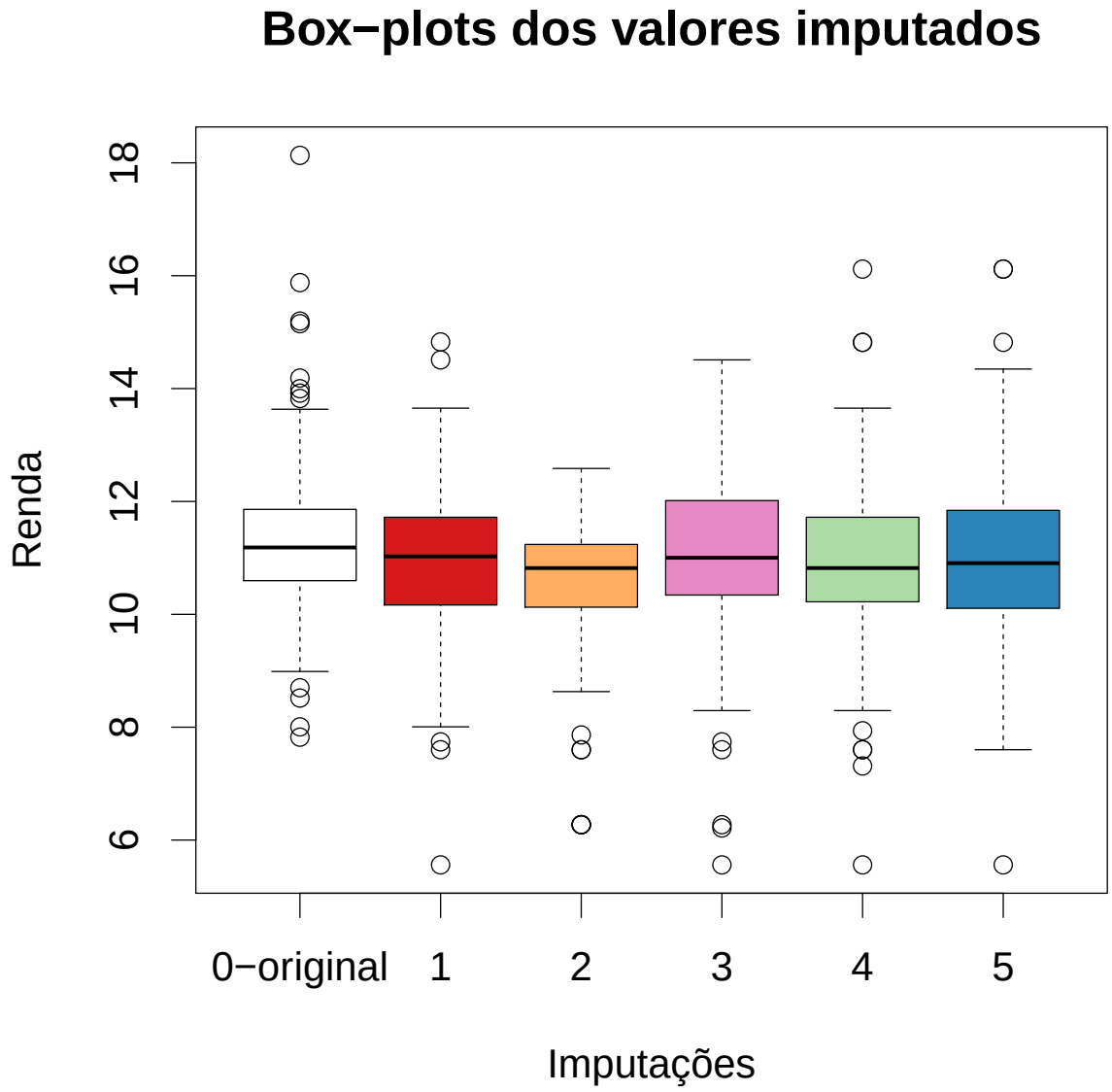
Para os box-plots acima percebemos distribuições divergentes entre os dados originais e os dados imputados, e também existe a confirmação que a imputação nº 2 está divergente entre os dados originais e as outras imputações.

Para a adequação do modelo temos o QQ-plot abaixo:



O QQ-plot indica que a adequação do modelo, entre os dados de renda com valores ausentes e os dados de renda imputados, está demonstrando alguns deslocamentos do ajuste da reta.

O mecanismo de perda não aleatória aplicado na variável renda do banco de dados original gerou 144 observações ausentes. Portanto analisamos abaixo somente essas observações, tanto as que foram retiradas do banco original, e depois codificadas como ausentes, tanto as 144 observações que foram imputadas. Assim temos abaixo os box-plots dessas 144 observações:



CONCLUSÃO

REFERÊNCIAS BIBLIOGRÁFICAS

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. John Wiley & Sons, New York.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data (Monographs on Statistics and Applied Probability). Chapman & Hall.

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.