Prevendo Nota em Matemática no ENEM no estado de Roraima

Jadson Rodrigo

20/06/2021

Introdução

A análise de regressão é uma técnica utilizada para quantificar uma relação entre uma variável resposta e um conjunto de uma ou mais variáveis explicativas,sendo muito utilizada para ajustamento de curvas . Neste projeto,o foco será em estimar a nota em Matemática no Enem no estado de Roraima,de acordo com variáveis como Nota em Ciências Humanas,Sexo,Idade,Cor,etc .

Descrição do Banco de dados

Primeiramente,pode-se conhecer as variáveis do banco de dados "dados_218405 (1).csv" através das 6 primeiras linha do conjunto de dados .

```
## # A tibble: 6 x 16
##   UF    Idade Sexo  Cor   EstadoCivil NotaCN NotaCH NotaLC NotaMT NotaRED
##   <chr> <dbl> <chr> <chr> <chr>       <chr>  <chr>  <chr>  <chr>    <dbl>
## 1 RR       23 F     "PPI" Solteiro    496.4  547.8  551.9  314.3      560
## 2 RR       17 M     "PPI" Solteiro    424.4  566.3  490    328.4      340
## 3 RR       18 F     "PPI" Solteiro    464.2  536.5  444.1  341.9      600
## 4 RR       24 F     "PPI" Solteiro    571.9  593.5  538.8  583.6      640
## 5 RR       19 M     "PPI" Solteiro    475    565.1  454.5  351.9      560
## 6 RR       24 F     "N\x~ Solteiro    552.2  579.9  579.8  541.9      800
## # ... with 6 more variables: InstrPai <chr>, InstrMae <chr>, NPessoas <dbl>,
## #   Renda <chr>, EscolaEM <chr>, TurnoEM <chr>
```

Pode-se então dá um breve descritivo das variáveis do conjunto de dados :

UF=Estado de residência do aplicante,todas RR

Idade=Variável discreta relacionada a idade do aplicante

Sexo=Variável qualitativa relacionada ao sexo do aplicante,nos níveis Masculino e Feminino

Cor =Variável qualitativarelacionada a cor do aplicantes,nos níveis PPI e não PPI)

EstadoCivil = Variável qualitativa com os níveis Solteiro ou outro

NotaCN=Nota em Ciências da natureza

NotaCH = Nota em Ciências humanas

NotaLC = Nota em linguagens e códigos

NotaMT Nota em Matemática

NotaRED= Nota em redação

InstrPai=Nível de instrução do pai,nos níveis Fundamental,Médio ou Superior

InstrMae=Nível de instrução da mãe,nos níveis Fundamental,Médio ou Superior)

NPessoas=Número de pessoas que moram na casa do aplicante

Renda=Renda total da familia do aplicante,nos níveis 1. Ate 1 SM, 2. 1 a 3 SM ,3. 3 a 5 SM,4. 5 a 10 SM,5. 10 ou mais,onde SM=Salario minimo)

EscolaEM=Tipo de escola que o estudante estudou,nos níveis Publica/Privada,Somente Privada,Somente Publica

TurnoEM=Turno ensino médio,nos níveis Parte diurno/Parte noturno,somente diurno,somente noturno)

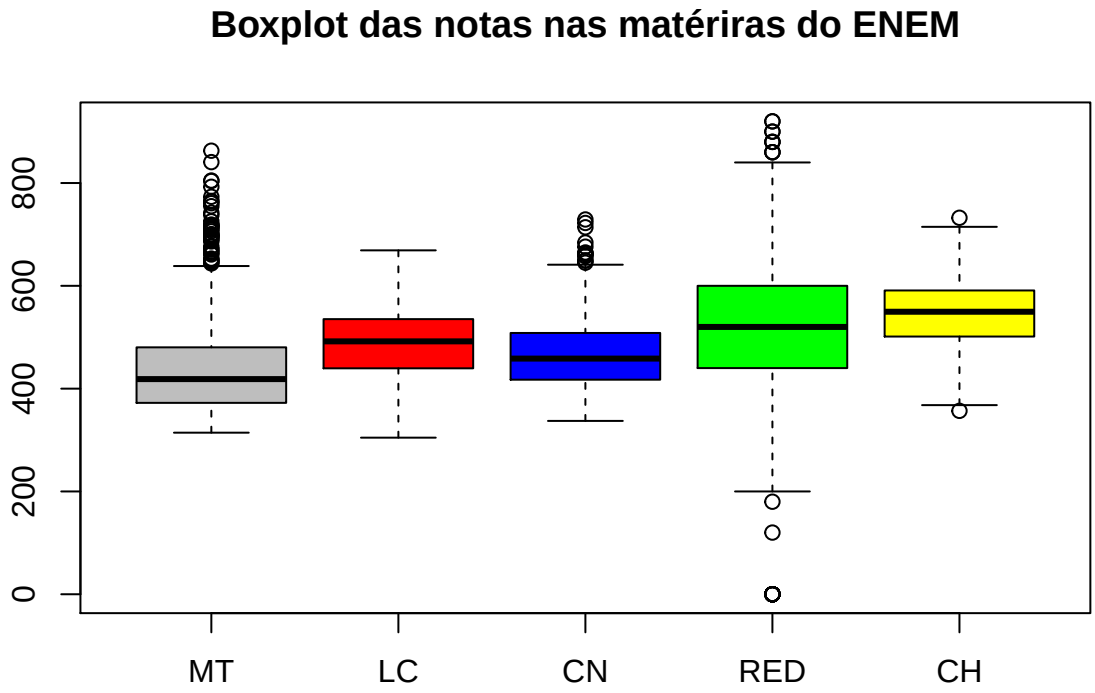
Após conhecer as variáveis é possivel conhecer as medidas sumárias das variáveis do banco de dados .

Análise descritiva das variáveis do conjunto de dados

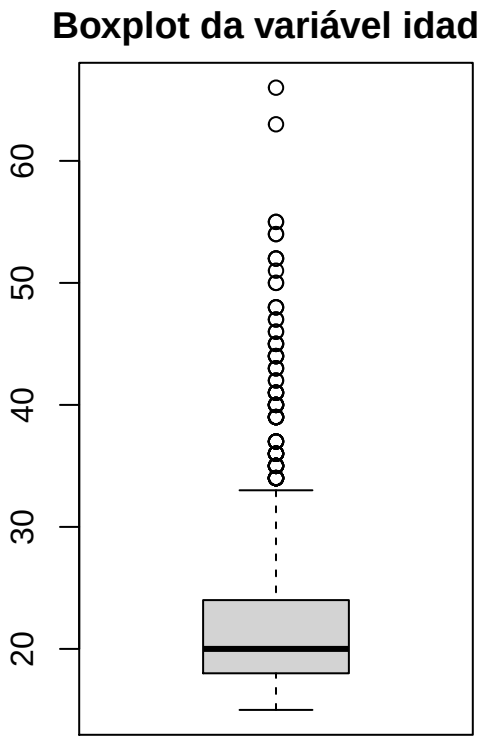
Após a descrição das variáveis , é possível estuda-las de uma forma mais profunda . Primeiramente , analisando as variáveis quantitativas .

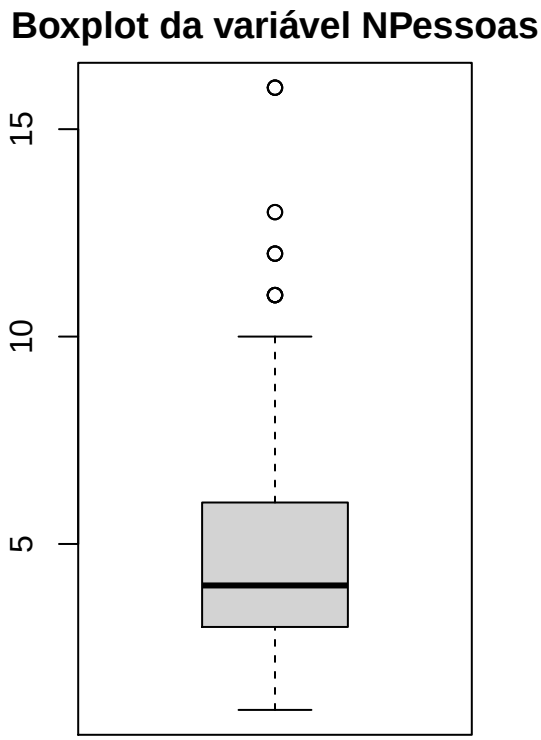
```
##      Idade           NotaMT          NotaLC          NotaCN     
##  Min.   :15.00   Min.   :314.3   Min.   :304.6   Min.   :337.2  
##  1st Qu.:18.00   1st Qu.:372.4   1st Qu.:439.6   1st Qu.:417.4  
##  Median :20.00   Median :418.6   Median :491.9   Median :458.4  
##  Mean   :22.64   Mean   :438.3   Mean   :486.7   Mean   :466.1  
##  3rd Qu.:24.00   3rd Qu.:480.3   3rd Qu.:535.2   3rd Qu.:508.3  
##  Max.   :66.00   Max.   :862.9   Max.   :669.0   Max.   :729.1  
##     NotaRED          NotaCH         NPessoas      
##  Min.   :  0.0   Min.   :356.8   Min.   : 1.000  
##  1st Qu.:440.0   1st Qu.:501.4   1st Qu.: 3.000  
##  Median :520.0   Median :549.6   Median : 4.000  
##  Mean   :519.7   Mean   :544.9   Mean   : 4.603  
##  3rd Qu.:600.0   3rd Qu.:591.0   3rd Qu.: 6.000  
##  Max.   :920.0   Max.   :732.4   Max.   :16.000
```

Ao se observar a tabela 1,pode-se verificar que a área que os aplicantes de Roraima vão melhor é Ciências humanas,com média 544.9 . E além disso,a área que os aplicantes vão pior é na variável resposta de estudo,Matemática,com média 438.3 . Por fim,pode-se verificar se existe outliers para as variáveis quantitativas por meio de boxplot,primeiramente construindo boxplot para as variáveis relacionadas a notas .



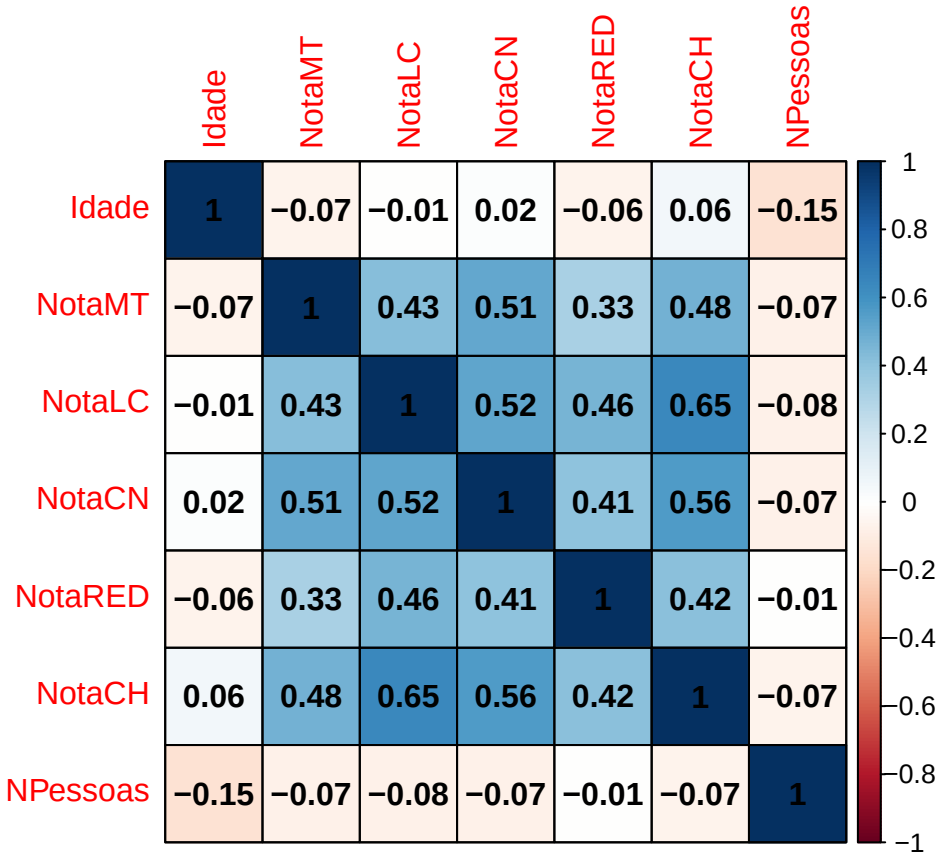
É possível construir agora boxplot para as variáveis relacionada a número de pessoas que moram na casa do aplicante,e também de idade .





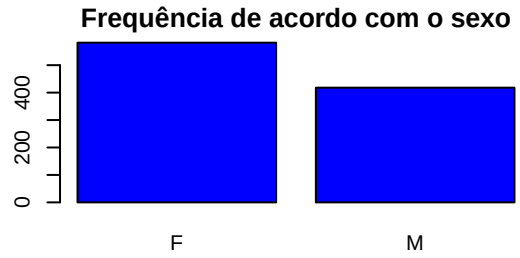
Uma possível análise que pode ser feita em relação aos boxplots é que é possível observar algumas presenças de outliers nas variáveis ,principalmente na variável NotaMT e NotaRED,porém a decisão aqui é não retirá-las,pois segundo a coluna de máximos e mínimos da Tabela 1 não há valores que fogem a regra do ENEM,isto é,as notas estão no intervalo 0-1000,e também um aplicante de 15 a 66 anos podem realizar ENEM,e é "possível" morar 16 pessoas em um casa . E retirar esses valores poderiam comprometer a predição de novas observações com esses valores .

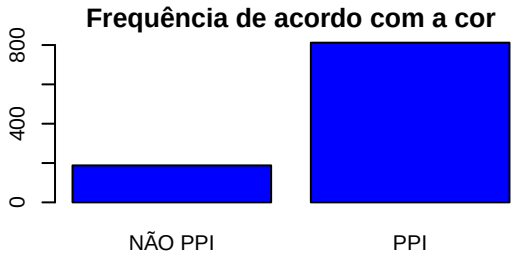
Por fim,pode-se verificar como a variável NotaMT ,Nota em Matemática se relaciona com as outras variáveis quantitativas do conjunto de dados ,através de uma matriz de correlação .Onde uma relação de duas variáveis próxima de 1 entre duas variáveis indica uma alta relação,ou seja ,quando uma aumenta a outra também aumenta . Enquanto que um coeficiente de correlação de -1 indica o oposto,ou seja ,se uma aumenta a outra diminui . Além disso, uma correlação entre duas variáveis pode ocasionar o problema de multicolinearidade .



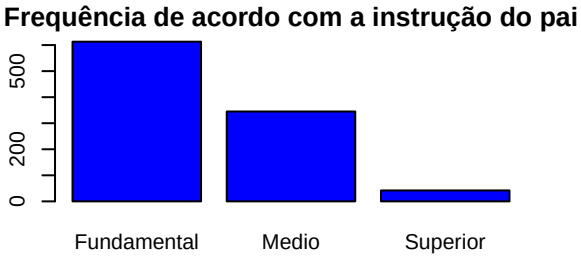
Ao se analisar a matriz,pode-se verificar que a variável NotaCN é a que mais está correlacionada com a variável NotaMT, o que faz de certa forma sentido,pois são duas matérias de exatas,e quando uma pessoa vai bem em Matemática ela normalmente vai bem em Natureza,devido a isso NotaCN provavelmente entrará no modelo .Além disso,NotaCH,NotaLC,NotaRED,tem uma correlação considerável com a variável NotaMT,o que é um indício que elas entrarão no modelo de regressão . E as variáveis idade e NPessoas têm uma correlação quase nula com a variável NotaMT ,com grau de correlação de -0.07 . A respeito da multicolinearidade,as variáveis apresentam entre si uma correlação máxima de 0.65,que ocorre as variáveis NotaLC e NotaCH,o que é bom para o modelo,pois evita problemas de multicolinearidade,tais como o modelo não ser único,efeitos na soma de quadrados,etc .

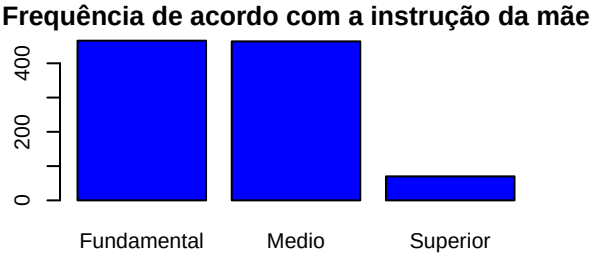
Por fim, pode-se realizar um estudo para as variáveis qualitativas do conjunto de dados,observando como os níveis estão distribuidos em cada uma das variáveis .

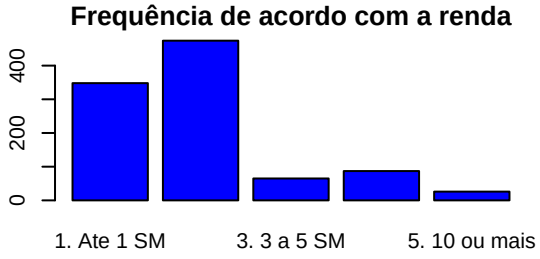

Frequência de acordo com o sexo
0
200
400
F
M

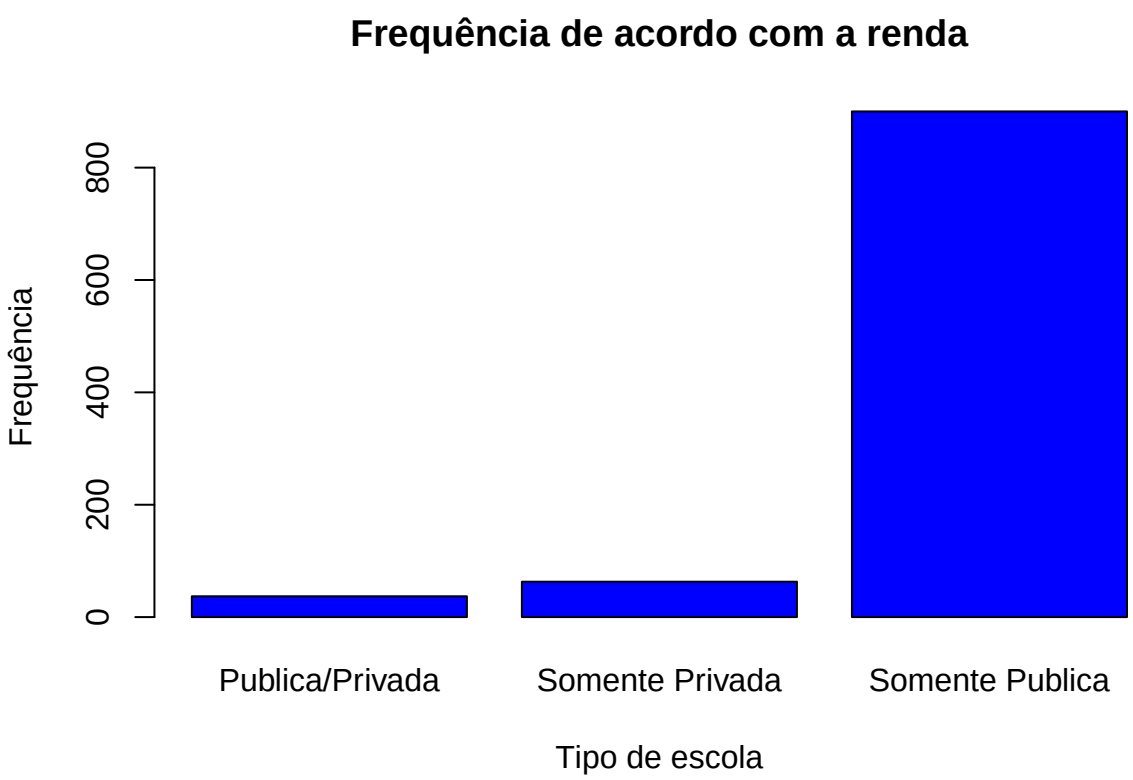

Frequência de acordo com a cor
0
400
800
NÃO PPI
PPI

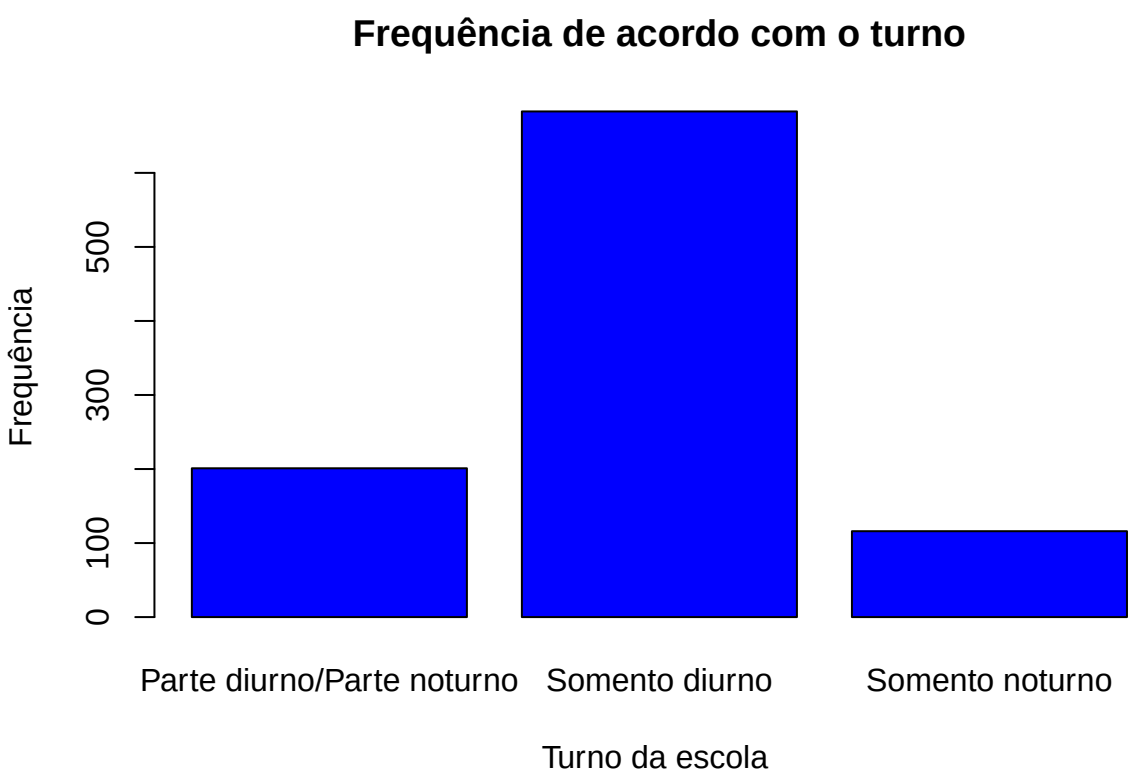

Frequência de acordo com o estado civil
0
400
800
Outro
Solteiro


Frequência de acordo com a instrução do pai
0
200
500
Fundamental
Medio
Superior


Frequência de acordo com a instrução da mãe
0
200
400
Fundamental
Medio
Superior


Frequência de acordo com a renda
0
200
400
1. Ate 1 SM
3. 3 a 5 SM
5. 10 ou mais


Frequência de acordo com a renda
Frequência
0
200
400
600
800
Publica/Privada
Somente Privada
Somente Publica
Tipo de escola



Dos gráficos , é possivel retirar alguns insights . A maioria dos aplicantes são de escola pública,e o período em que há mais estudantes é somente diurno ,a maior parte dos aplicantes são solteiro . Outra análise a ser feita é referente a educação ao nível de instrução dos país e das mãe , onde há a predominância do nível Fundamental,principalmente no que se refere aos pais .

Modelagem

Para a construção dos modelos para predidizir a nota em Matemática tem-se que criar variáveis dummys para as variáveis qualitativas ,e definir um nível de referência para elas . O nível de refência para a variável sexo é "Masculino",para a cor é "NãoPPI",para o estado cívil é "Outro",para instrução mãe e pai o nível de referência é "Fundamental" ,para a renda é "1. Ate 1 SM",para o tipo de escola é "Publica/Privada",e para a variável referente ao turno é "Parte diurno/Parte noturno" .

Há 5 métodos para selecionar variáveis para ajustar um modelo de Regressão ,que são AIC,PRESSp, BIC,Cp de mallow,R2 ajustado . Então,pode-se selecionar o modelo para cada critério,e verificar qual modelo se ajusta melhor utilizando o critério do MSPR(erro quadrático médio),onde quando um modelo apresenta MSPR menor dos outros, é um indicio de que aquele modelo é melhor do que os demais .

Modelo 1 = Segundo o critério do AIC e PRESSp :

```
##
## Call:
## lm(formula = NotaMT ~ Idade + Cor + NotaCN + NotaCH + NotaLC +
##     Renda_5.10oumais + EscolaEM_SomentePrivada, data = dadostreino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.841  -48.948   -6.082   41.350  271.594
```

```
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               74.59435   28.69912   2.599 0.009544 **
## Idade                     -1.54498    0.39055  -3.956 8.41e-05 ***
## Cor                      -17.61197    7.17106  -2.456 0.014295 *
## NotaCN                     0.38949    0.05227   7.452 2.76e-13 ***
## NotaCH                     0.29051    0.05775   5.031 6.24e-07 ***
## NotaLC                     0.14148    0.05316   2.661 0.007961 **
## Renda_5.10oumais          64.36887   17.69366   3.638 0.000295 ***
## EscolaEM_SomentePrivada   38.83486   12.69559   3.059 0.002307 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.42 on 692 degrees of freedom
## Multiple R-squared:  0.3931, Adjusted R-squared:  0.3869
## F-statistic: 64.03 on 7 and 692 DF,  p-value: < 2.2e-16
```

Modelo 2 = Segundo o critério do BIC :

```
##
## Call:
## lm(formula = NotaMT ~ Idade + NotaCN + NotaCH + NotaLC + Renda_5.10oumais +
##     EscolaEM_SomentePublica, data = dadostreino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -213.492  -47.867   -6.772   42.397  249.484
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               89.98389   32.28666   2.787 0.005465 **
## Idade                     -1.53739    0.39189  -3.923 9.61e-05 ***
## NotaCN                     0.39103    0.05283   7.401 3.92e-13 ***
## NotaCH                     0.29146    0.05796   5.029 6.30e-07 ***
## NotaLC                     0.14918    0.05325   2.802 0.005227 **
## Renda_5.10oumais          63.39071   18.13816   3.495 0.000504 ***
## EscolaEM_SomentePublica -35.57292   11.23051  -3.168 0.001605 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.68 on 693 degrees of freedom
## Multiple R-squared:  0.3878, Adjusted R-squared:  0.3825
## F-statistic: 73.17 on 6 and 693 DF,  p-value: < 2.2e-16
```

Modelo 3 = Segundo o critério do Cp de mallow e R ajustado :

```
##
## Call:
## lm(formula = NotaMT ~ Idade + Cor + NotaCN + NotaCH + NotaLC +
##     Renda_5.10oumais + EscolaEM_SomentePublica, data = dadostreino)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -221.00  -49.19   -5.97   41.03  240.37
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              111.94782   33.43880   3.348 0.000859 ***
## Idade                     -1.54911    0.39056  -3.966 8.06e-05 ***
## Cor                      -17.31125    7.17717  -2.412 0.016125 *
## NotaCN                     0.37924    0.05288   7.172 1.90e-12 ***
## NotaCH                     0.28999    0.05776   5.020 6.57e-07 ***
## NotaLC                     0.14355    0.05311   2.703 0.007048 **
## Renda_5.10oumais          59.84183   18.13522   3.300 0.001017 **
## EscolaEM_SomentePublica -33.98821   11.21094  -3.032 0.002523 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.42 on 692 degrees of freedom
## Multiple R-squared:  0.3929, Adjusted R-squared:  0.3868
## F-statistic: 63.99 on 7 and 692 DF,  p-value: < 2.2e-16
```

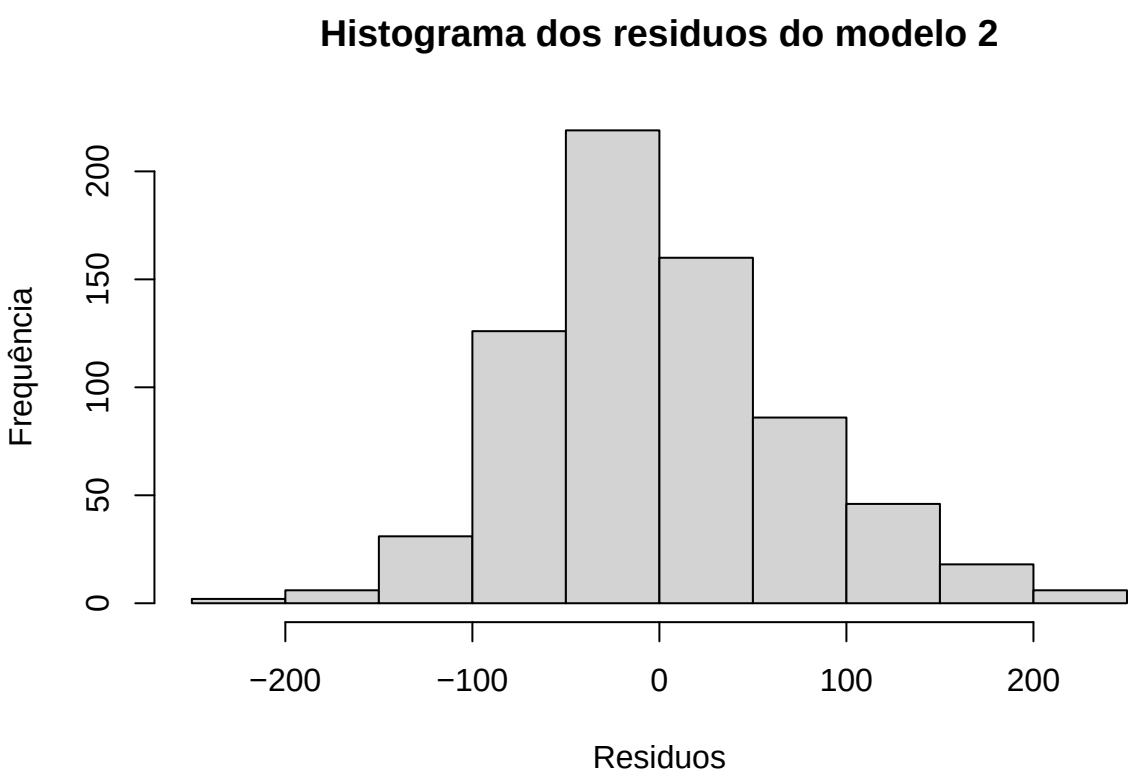
Por fim,pode-se calcular o MSPR com os dados de teste,e verificar qual dos modelos apresenta o menor MSPR .

```
##    Modelo     MSPR
## 1 Modelo1 5923.386
## 2 Modelo2 5906.820
## 3 Modelo3 5914.324
```

Ao se observar o MSPR pode-se observar que o melhor modelo é o segundo modelo ,que apresenta um MSPR de 5906.82 .

Pode-se verificar então se o modelo possui os pressuposto da regressão linear,que é homogeneidade e normalidade dos residuos .

Para a normalidade pode-se construir um histograma dos residuos .



Do gráfico pode-se suspeitar que o residuos não seguem uma distribuição normal,pois o gráfico não possui formato de sino . Porém,para se confirmar tem-se que realizar um teste ,sendo que o teste escolhido é o Teste de Shapiro ,em que as hipóteses são :

H0= Os residuos seguem uma distribuição normal

H1= Os residuos não seguem uma distribuição normal

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo2$residuals
## W = 0.98383, p-value = 5.446e-07
```

A um nível de significância de 5%,pode-se observar que os residuos não seguem uma distribuição normal,logo o modelo fere um dos pressupostos da regressão linear,que é a linearidade . Portanto,pode-se tentar alguma transformação . Uma transformação possível é transformação a nota em Matemática como :

$$NotaMT* = 1/NotaMT$$

Logo,ajustando o modelo com a transformação,tem-se que o modelo ajustado é dado por :

```
##
## Call:
## lm(formula = 1/NotaMT ~ Idade + NotaCN + NotaCH + NotaLC + Renda_5.10oumais +
```

```
##     EscolaEM_SomentePublica, data = dadostreino)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.047e-03 -2.526e-04 -2.410e-06  2.298e-04  1.090e-03
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.869e-03  1.558e-04  24.827  < 2e-16 ***
## Idade                     8.510e-06  1.891e-06   4.499 7.99e-06 ***
## NotaCN                   -1.468e-06  2.550e-07  -5.756 1.29e-08 ***
## NotaCH                   -1.354e-06  2.797e-07  -4.840 1.60e-06 ***
## NotaLC                   -7.872e-07  2.570e-07  -3.063  0.00228 **
## Renda_5.10oumais         -2.142e-04  8.754e-05  -2.447  0.01466 *
## EscolaEM_SomentePublica   1.289e-04  5.420e-05   2.378  0.01767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0003508 on 693 degrees of freedom
## Multiple R-squared:  0.326,  Adjusted R-squared:  0.3202
## F-statistic: 55.87 on 6 and 693 DF,  p-value: < 2.2e-16
```

Por fim,pode-se realizar um teste de shapiro nesse modelo transformado ,tem-se então :

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo2.1$residuals
## W = 0.99679, p-value = 0.1769
```

Pode-se observar que o p_valor do teste é 0.1769, logo a um nível de significância de 5%, não se rejeita a hipótese de que os residuos seguem uma distribuição normal .

Pode-se então realizar um teste de homogeneidade ,para verificar homocedasticidade , em que as hipóteses são :

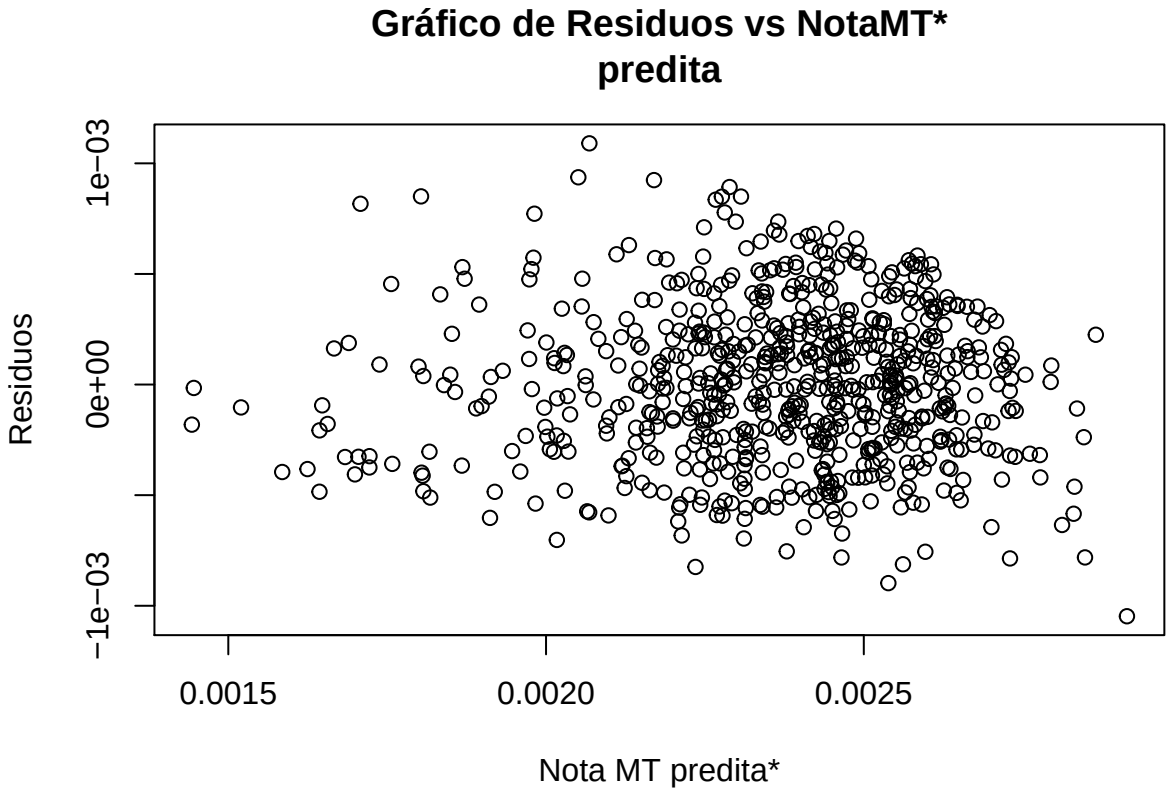
```
H0= A  variância  do  modelo  é  constante
H1 = A  variância  do modelo  não  é constante
```

Realizando o teste de Breusch-Pagan test tem-se :

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2.1
## BP = 7.3615, df = 6, p-value = 0.2887
```

Pode-se observar que o p_valor do teste de homogeneidade é de 0.2887 , logo a um nível de significância de 5% ,não se rejeita a hipótese de que o modelo apresenta variância constante ,logo o modelo apresenta homocedasticidade .

É possível então construir um gráfico para verificar como os residuos estão de acordo com a nota predita pelo modelo .



O gráfico de homogeneidade diz que a variância está constante,pois os resíduos se encontram ao redor de 0, e a variabilidade é constante,pois não se tem uma grande variação de acordo com o valor de NotaMT* predita. Logo,pode-se concluir que a transformação na variável NotaMT deu certo ,e o modelo ajustado está correto .

Conclusão

Portanto ,do summary do modelo pode-se retirar algumas conclusões :

1) Os coeficientes do modelo são :

```
##                          modelo2.1$coefficients
## (Intercept)                        3.868813e-03
## Idade                              8.510147e-06
## NotaCN                            -1.467701e-06
## NotaCH                            -1.353848e-06
## NotaLC                            -7.871967e-07
## Renda_5.10oumais                  -2.142031e-04
## EscolaEM_SomentePublica            1.289079e-04
```

2) O modelo consegue explicar 32.6% da variabilidade dos dados

3) As variáveis importantes no modelo são a idade, NotaCN(Ciências da Natureza) , NotaCH(Ciências humanas),NotaLC(Linguagens e Códigos) , Renda_5.10oumais , EscolaEM_SomentePublica. Enquanto que variáveis como a NotaRED(Nota Redação), Sexo,EstadoCivil,CorAsvariáveis foram descartadas .

4) As variáveis NotaCN,NotaCH,NotaLC,Renda_5.10oumais tem uma relação negativa com NotaMT* ,o que pode indicar que quanto maior o valor dessas variáveis maior será a nota em Matemática,pois

quanto menor o valor de NotaMT* maior NotaMT (nota em Matemática),pois NotaMT* e NotaMT são inversamente proporcionais. Enquanto que as variáveis Idade e EscolaEM_SomentePublica tem uma relação positiva,o que indica que quanto maior a idade e quando é somente escola pública tende-se a ter um pior desempenho em Matemática .

5) Outro fator que pode ser relevante para a análise é em relação a condição social do estudante,pois o modelo mostra que pessoas de escola pública(normalmente com menor poder aquisitivo) tendem a tirar uma nota pior em Matemática,enquanto que quem possui Renda_5.10oumais tende a ir melhor .