Atividade Avaliativa

Análise de Regressão Linear

Ana Maria Alves da Silva

2025-04-05

- O conjunto de dados trata-se de 11 características clínicas utilizadas para a previsão de possíveis eventos relacionados a pressão arterial.

- A hipertensão arterial (HA) representa o principal fator de risco para desenvolvimento de doenças cardiovasculares e mortalidade em todo o mundo. É uma doença multifatorial, caracterizada e diagnosticada por níveis elevados e sustentados de pressão arterial (PA), possuindo, como critério clínico, em indivíduos maiores de 18 anos, níveis tensionais iguais ou maiores a 140 mmHg × 90 mmHg1.
- Há diversos fatores que podem ser responsáveis pelo desenvolvimento da doença. Entre eles podemos citar: idade, sexo, colesterol, açúcar no sangue em jejum, doença cardíaca, entre outros. O objetivo nesse trabalho é verificar dentre as variáveis disponíveis quais delas podem contribuir para a hipertensão arterial(RestingBP).
- Nesse estudo foram utilizados dados relacionados à variável dependente (pressão arterial) e às seguintes variáveis independentes: idade, sexo, colesterol sérico, açúcar no sangue em jejum, resultados de eletrocardiograma, doenças cardíacas , angina induzida por exercícios, frequência cardíaca, "OldPeak" e "ST" que é relacionada a inclinação do segmento ST do exercício de pico.

- Questões para a atividade avaliativa:

Item 1: (0.5 pts.) Realizar a Análise Descritiva dos Dados:

Para se familiarizar com os dados, faça uma boa caracterização de todas as variáveis quantitativas e qualitativas do arquivo, uma a uma, usando medidas resumo adequados a cada variável. Comente!

Solução:

Primeiramente, vamos carregar o conjunto de dados.

```
setwd <- "/Users/anamaria/especializacao/modulo_11/Atividade Avaliativa"
df <- read.csv("heart.csv", sep = ",")
print(dim(df))
```

```
## [1] 918  12
```

```
print(is.data.frame(df))
```

```
## [1] TRUE
```

Logo, o dataframe df possui 918 observações e 12 colunas. O código a seguir nos mostra os tipos de variávies presentes no dataframe.

```
str(df)
```

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : chr  "M" "F" "M" "F" ...
##  $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

Para verificar se há valores ausentes, podemos usar a função colSums e is.na no dataframe, conforme o código abaixo.

```
print(colSums(is.na(df)))
```

```
##            Age            Sex  ChestPainType      RestingBP    Cholesterol
##              0              0              0              0              0
##      FastingBS     RestingECG          MaxHR ExerciseAngina        Oldpeak
##              0              0              0              0              0
##       ST_Slope   HeartDisease
##              0              0
```

Ou ainda:

```
print(colnames(df)[colSums(is.na(df)) > 0])
```

```
## character(0)
```

Em ambos os casos, vemos que não há valores ausentes. Se houvesse valores ausentes teriamos que trata-los de alguma maneira. Para obtermos a estatísticas descritivas (média, mediana, variância, frequências, etc.) podemos usar:

```
print(summary(df))
```

```
##       Age            Sex            ChestPainType        RestingBP
##  Min.   :28.00   Length:918         Length:918         Min.   :  0.0
##  1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
##  Median :54.00   Mode  :character   Mode  :character   Median :130.0
##  Mean   :53.51                                         Mean   :132.4
##  3rd Qu.:60.00                                         3rd Qu.:140.0
##  Max.   :77.00                                         Max.   :200.0
##   Cholesterol      FastingBS       RestingECG           MaxHR
```

```
##  Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
##  1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
##  Median :223.0   Median :0.0000   Mode  :character   Median :138.0
##  Mean   :198.8   Mean   :0.2331                      Mean   :136.8
##  3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
##  Max.   :603.0   Max.   :1.0000                      Max.   :202.0
##  ExerciseAngina        Oldpeak          ST_Slope          HeartDisease
##  Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
##  Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
##                     Mean   : 0.8874                      Mean   :0.5534
##                     3rd Qu.: 1.5000                      3rd Qu.:1.0000
##                     Max.   : 6.2000                      Max.   :1.0000
```

Lembremos que as variáveis podem ser classificadas em dois grandes grupos:

- Qualitativas: Representam categorias e podem ser nominais, isso é sem ordem natural, ou podem ser ordinais, isto é com uma ordem definida.
- Quantitativas: Representam números e podem ser discretas, valores inteiros, ou podem ser contínuas, isso é com valores em números reais.

Além disso, as escalas de medida são:

- Nominal: Categorização sem ordem.
- Ordinal: Categorização com ordem definida.
- Intervalar: Diferenças entre valores fazem sentido, mas não há um zero absoluto.
- Razão: Diferenças e razões fazem sentido, e há um zero absoluto.

Abaixo, classificamos as variáveis do estudo conforme seu tipo e escala de medida:

1) Variáveis Quantitativas

1.1 Age

- **Tipo:** Quantitativa
- **Escala:** Razão
- **Justificativa:** Idade expressa quantidade de anos de vida. Zero implica ausência de idade (embora não seja observável em estudo de adultos), e faz sentido dizer que 60 anos é o dobro de 30 anos. Logo, há um zero absoluto e as razões são interpretáveis.

1.2 RestingBP

- **Tipo:** Quantitativa
- **Escala:** Razão
- **Justificativa:** Pressão arterial (mmHg) pode teoricamente partir de zero (ausência de pressão) e se mantém como uma grandeza física mensurável. As diferenças e proporções fazem sentido (por exemplo, 140 mmHg é 2 vezes 70 mmHg).

1.3 Cholesterol

- **Tipo:** Quantitativa
- **Escala:** Razão
- **Justificativa:** Valores de colesterol (mg/dL) também se baseiam em uma medida com zero absoluto (0 mg/dL = ausência de colesterol). Em geral, interpreta-se como uma quantidade mensurável, sendo assim razão.

1.4 MaxHR

- **Tipo:** Quantitativa
- **Escala:** Razão
- **Justificativa:** Batimentos cardíacos máximos por minuto (bpm) também se interpretam em relação a um zero absoluto (0 bpm = ausência de batimentos). 160 bpm, por exemplo, é o dobro de 80 bpm. Logo, escala de razão.

1.5 Oldpeak

- **Tipo:** Quantitativa
- **Escala:** Intervalo
- **Justificativa:** Oldpeak representa **desvio** (depressão ou elevação) do segmento ST em relação ao repouso, podendo ser **positivo ou negativo**. Isso indica que zero é um "ponto de referência" (nenhuma variação), mas não implica ausência de fenômeno. Assim, a interpretação é mais adequada como escala de intervalo (ex.: -2.0 vs. +2.0 são desvios distintos em direções opostas, mas não faz sentido dizer que +4.0 é o "dobro" de +2.0).

2) Variáveis Qualitativas

2.1 Sex

- **Tipo:** Qualitativa
- **Escala:** Nominal
- **Justificativa:** Valores "M" ou "F" (ou "0"/"1"), sem ordem intrínseca. É apenas uma categorização (masculino e feminino).

2.2 ChestPainType

- **Tipo:** Qualitativa
- **Escala:** Nominal (ou Ordinal, dependendo da codificação)
- **Justificativa:** Se for apenas um rótulo de tipos (ex.: "TA", "ATA", "NAP", "ASY"), costuma ser nominal, pois não há necessariamente uma hierarquia. Se houvesse uma clara progressão de severidade, poderíamos considerar ordinal.

2.3 FastingBS

- **Tipo:** Qualitativa (binária)
- **Escala:** Nominal
- **Justificativa:** Em geral é 0 ou 1 (ex.: "FastingBS>120 mg/dL" = 1). Não há ordem entre 0 e 1 que faça sentido em termos de magnitude; é apenas um indicador (sim ou não).

2.4 RestingECG

- **Tipo:** Qualitativa
- **Escala:** Nominal (ou Ordinal, conforme a codificação clínica)
- **Justificativa:** Se codificado como "Normal", "ST-T abnormality", "Left ventricle hypertrophy" etc., normalmente não há uma ordem estrita. Entretanto, alguns preferem considerá-la ordinal se existir progressão de gravidade. Geralmente, trata-se de níveis de anormalidade sem escala numérica explícita, então nominal costuma ser o mais frequente.

2.5 ExerciseAngina

- **Tipo:** Qualitativa (binária)
- **Escala:** Nominal
- **Justificativa:** Típico "Yes"/"No" (ou "Y"/"N"). Não há sentido em dizer que "Yes" é maior do que "No". É um estado (presença ou ausência de angina).

2.6 ST_Slope

- **Tipo:** Qualitativa (frequentemente)
- **Escala:** Ordinal, se os valores forem algo como "Up", "Flat", "Down"
- **Justificativa:** Em termos clínicos, "Up" < "Flat" < "Down" reflete gravidade crescente de anormalidade no segmento ST. Assim, há uma **ordem** plausível ("Up" é melhor que "Flat", e "Flat" é melhor que "Down"). Se não houver registro dessa hierarquia, trataríamos como nominal, mas é comum interpretá-la como ordinal.

2.7 HeartDisease

- **Tipo:** Qualitativa (binária)
- **Escala:** Nominal
- **Justificativa:** 0 ou 1 (ausência ou presença de doença). É um critério de classificação, sem noção de "maior" ou "menor".

Observações

- Em alguns estudos, certas variáveis categóricas podem ter **ordem implícita** (por exemplo, graus de dor: leve, moderada, intensa), configurando uma escala **ordinal**.
- Já as variáveis numéricas que podem assumir **zero absoluto** e permitir comparação de razões ("o dobro, o triplo") são da **escala de razão**.
- Variáveis em que se mede uma diferença em relação a um ponto de referência (podendo ter valores negativos) tendem a ser **intervalares**.
- Quando uma variável binária (0/1) se refere a "presença/ausência" (síndrome, condição, sintoma), a escala é nominal, pois não há valor maior ou menor, apenas categorias distintas.

Essa classificação ajuda a **escolher as análises**, por exemplo, testes paramétricos vs. não paramétricos, regressões lineares vs. logísticas etc. E a forma de descrever cada variável, i.e., estatísticas para variáveis quantitativas, frequências para variáveis qualitativas.

```
df$Sex <- factor(df$Sex,
                 levels = c("M", "F"),
                 labels = c("Masculino", "Feminino"))

df$ChestPainType <- factor(df$ChestPainType,
                           levels = c("TA", "ATA", "NAP", "ASY"),
                           labels = c("Tipica Angina", "Angina Atipica",
                                      "Dor Nao Anginosa", "Assintomatico"))

df$FastingBS <- factor(df$FastingBS,
                       levels = c(0, 1),
                       labels = c("<=120", ">120"))

df$RestingECG <- factor(df$RestingECG,
                        levels = c("Normal", "ST", "LVH"),
                        labels = c("Normal", "ST-T wave abnormality",
                                   "Left Ventricular Hypertrophy"))

df$ExerciseAngina <- factor(df$ExerciseAngina,
                            levels = c("N", "Y"),
                            labels = c("Nao", "Sim"))

df$ST_Slope <- factor(df$ST_Slope,
                      levels = c("Up", "Flat", "Down"),
                      ordered = TRUE)

df$HeartDisease <- factor(df$HeartDisease,
                          levels = c(0, 1),
                          labels = c("Nao", "Sim"))
```

Transformando as variáveis categóricas em fator

```r
df <- df %>%
  mutate(across(c(RestingBP, Cholesterol,), ~
ifelse(. == 0, mean(., na.rm = TRUE), .)))

# Confirmando o tratamento
summary(df)
```

Tratamento de valores ausentes

```r
skim(df)
```

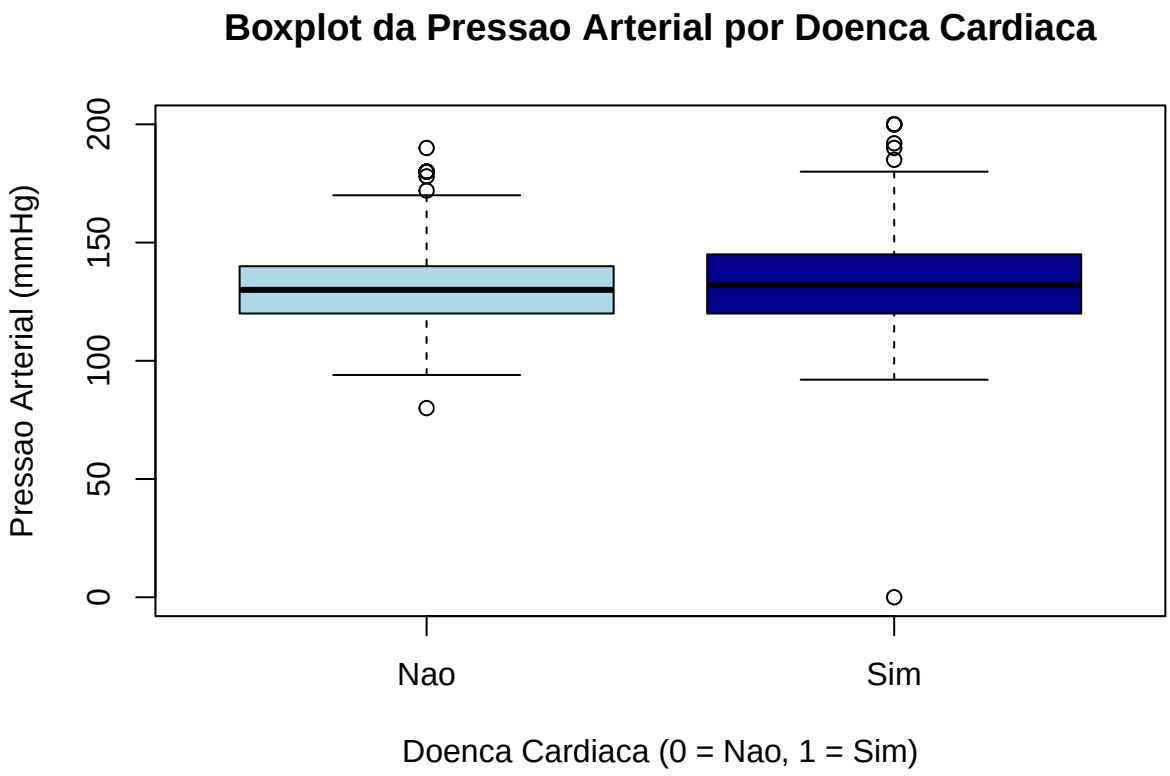
Análise exploratória mais completa

Item 2: (0.5 pts.) Faça uma análise descritiva gráfica dos dados usando boxplots e diagramas de dispersão. Comente!

Solução:

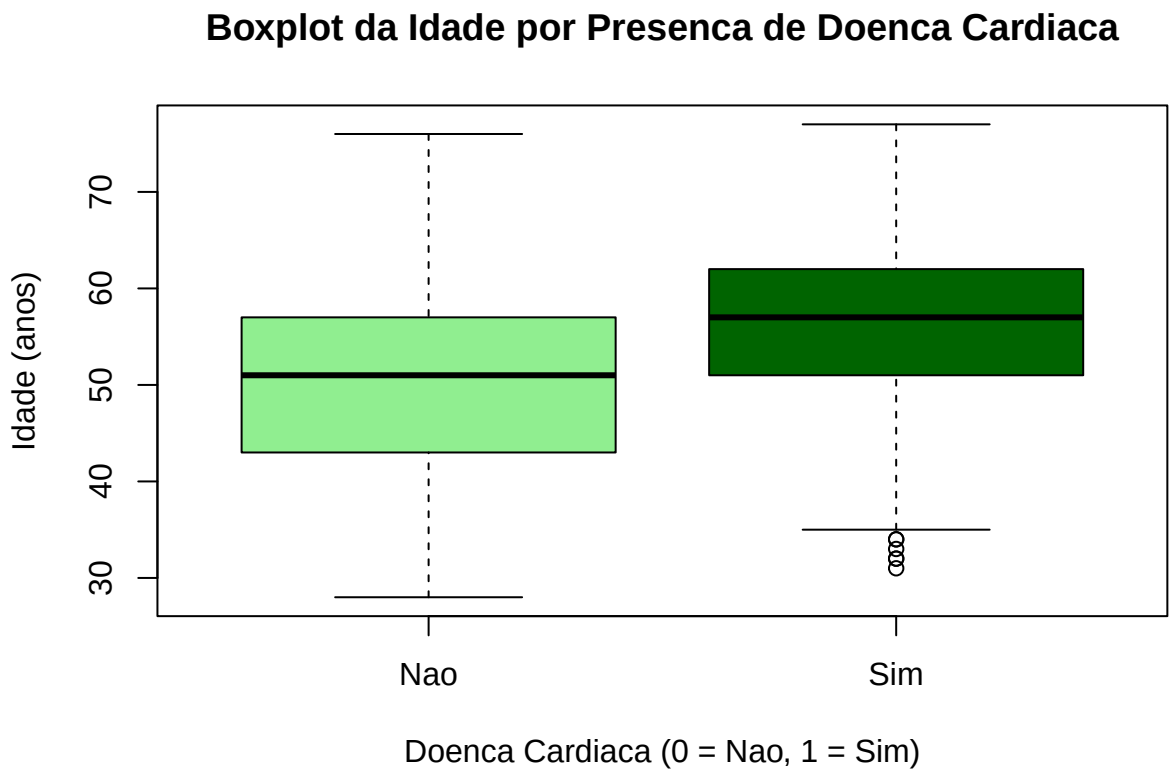
```r
boxplot(RestingBP ~ HeartDisease,
        data = df,
        main = "Boxplot da Pressao Arterial por Doenca Cardiaca",
        xlab = "Doenca Cardiaca (0 = Nao, 1 = Sim)",
        ylab = "Pressao Arterial (mmHg)",
        col = c("lightblue", "darkblue"),
        names = c("Nao", "Sim"))
```

1) Bloxplot



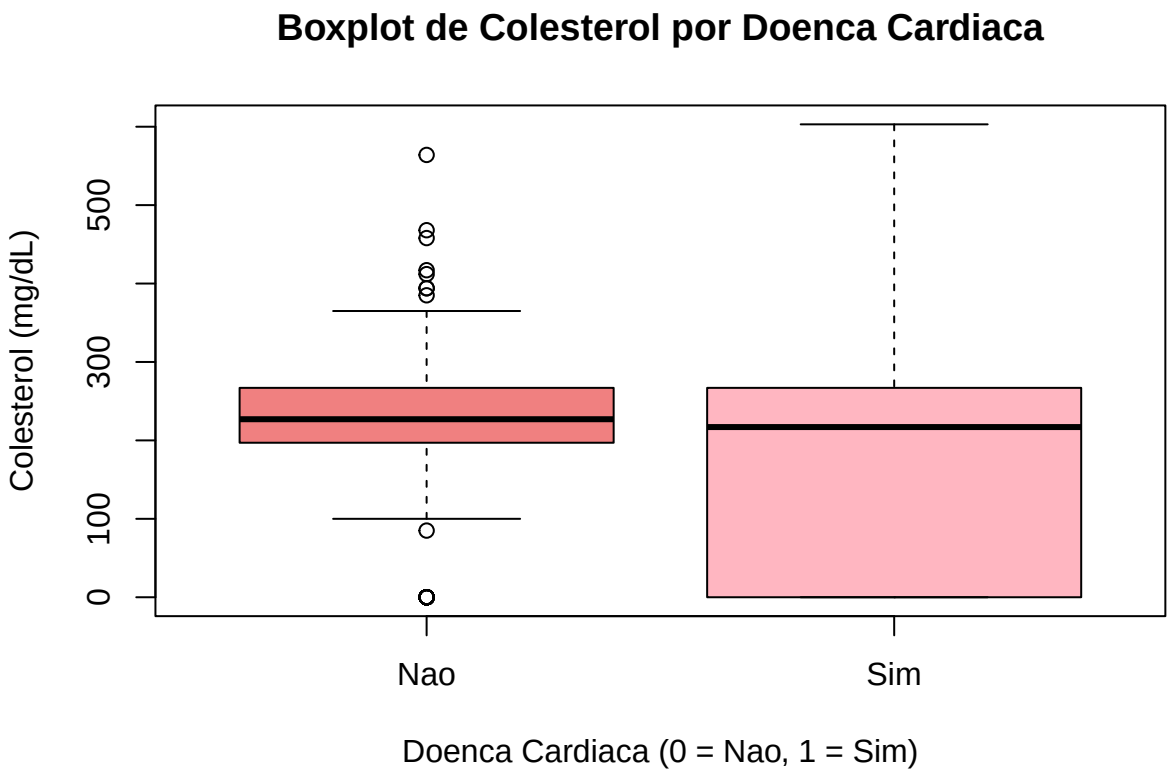
O boxplot indica que indivíduos com doença cardíaca tendem a apresentar valores medianos de pressão arterial ligeiramente mais elevados do que aqueles sem a doença. A amplitude interquartil é semelhante entre os grupos, porém o grupo com doença cardíaca apresenta mais valores extremos (outliers), sugerindo maior variabilidade. Essa diferença pode indicar que a pressão arterial elevada é mais comum em pacientes com problemas cardíacos.

```
boxplot(Age ~ HeartDisease,
        data = df,
        main = "Boxplot da Idade por Presenca de Doenca Cardiaca",
        xlab = "Doenca Cardiaca (0 = Nao, 1 = Sim)",
        ylab = "Idade (anos)",
        col = c("lightgreen", "darkgreen"),
        names = c("Nao", "Sim"))
```



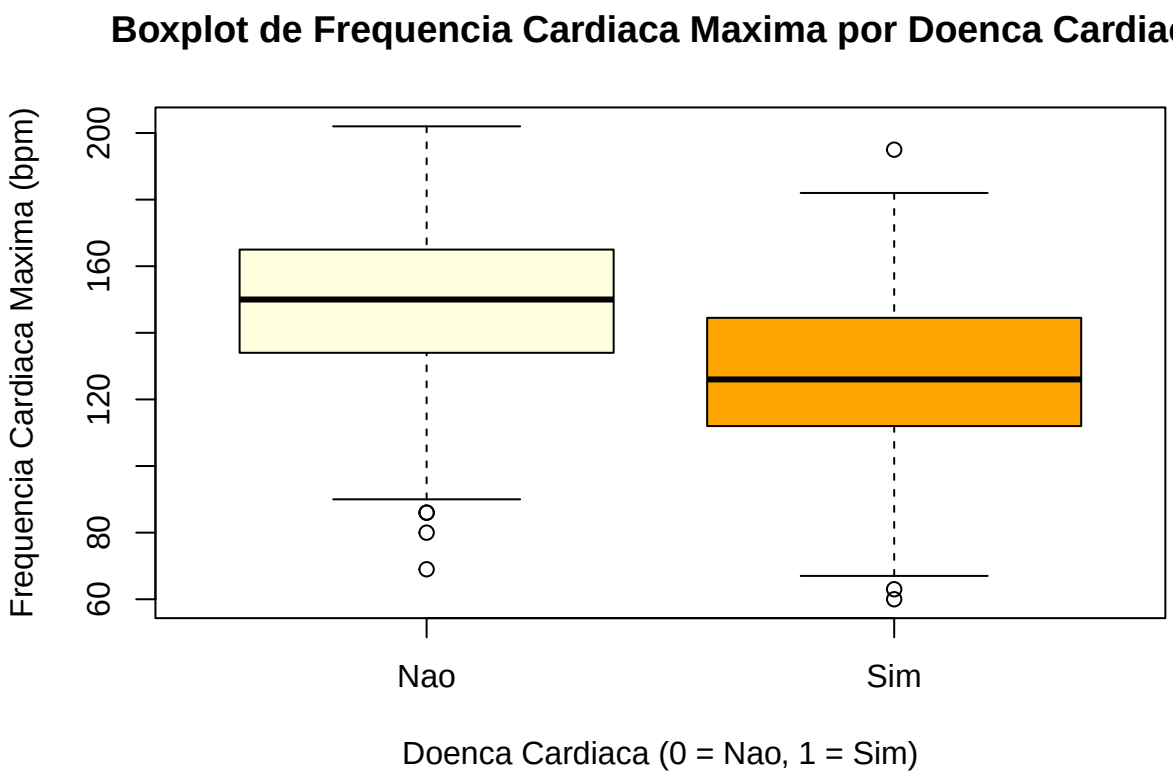
Observa-se que a mediana da idade dos indivíduos com doença cardíaca é superior à dos indivíduos sem a doença. O grupo com doença cardíaca também apresenta maior dispersão na idade, especialmente para idades mais avançadas. Isso sugere que a idade pode ser um fator de risco relevante, já que indivíduos mais velhos tendem a apresentar maior incidência de doença cardíaca.

```r
boxplot(Cholesterol ~ HeartDisease,
        data = df,
        main = "Boxplot de Colesterol por Doenca Cardiaca",
        xlab = "Doenca Cardiaca (0 = Nao, 1 = Sim)",
        ylab = "Colesterol (mg/dL)",
        col = c("lightcoral", "lightpink"),
        names = c("Nao", "Sim"))
```



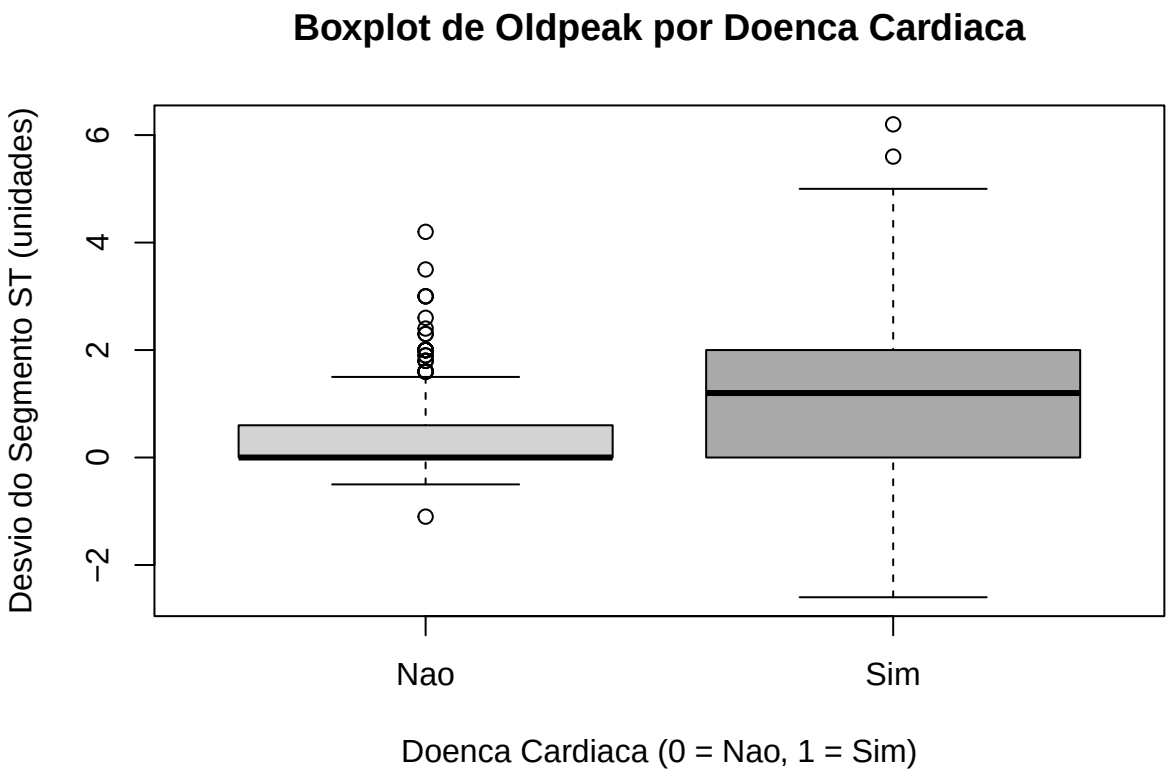
Os níveis de colesterol apresentam distribuições semelhantes entre os dois grupos. A mediana é levemente superior no grupo com doença cardíaca, mas a diferença é pequena. No entanto, ambos os grupos possuem diversos outliers, indicando que existem pacientes com colesterol muito elevado em ambas as categorias. Isso pode indicar que colesterol alto, isoladamente, não é um fator determinante da doença, mas ainda pode contribuir em conjunto com outras variáveis.

```r
boxplot(MaxHR ~ HeartDisease,
        data = df,
        main = "Boxplot de Frequencia Cardiaca Maxima por Doenca Cardiaca",
        xlab = "Doenca Cardiaca (0 = Nao, 1 = Sim)",
        ylab = "Frequencia Cardiaca Maxima (bpm)",
        col = c("lightyellow", "orange"),
        names = c("Nao", "Sim"))
```



Indivíduos com doença cardíaca tendem a atingir uma frequência cardíaca máxima mais baixa do que aqueles sem a condição. Isso é evidenciado por uma mediana inferior e uma menor dispersão nos valores mais altos. Essa observação pode estar associada a limitações cardíacas durante o esforço físico em pessoas com doença cardíaca.

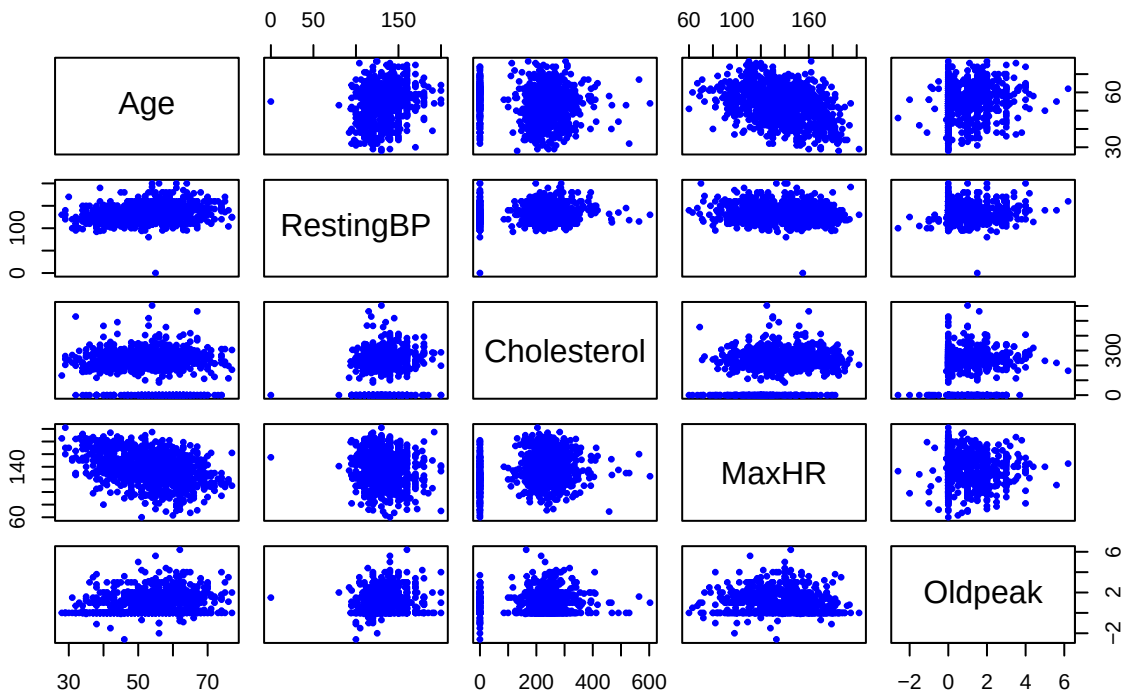
```
boxplot(Oldpeak ~ HeartDisease,
        data = df,
        main = "Boxplot de Oldpeak por Doenca Cardiaca",
        xlab = "Doenca Cardiaca (0 = Nao, 1 = Sim)",
        ylab = "Desvio do Segmento ST (unidades)",
        col = c("lightgray", "darkgray"),
        names = c("Nao", "Sim"))
```



O desvio do segmento ST (Oldpeak), que é uma medida obtida durante testes de esforço, é visivelmente maior em pacientes com doença cardíaca. A mediana desse grupo é consideravelmente mais elevada, e a distribuição é mais dispersa. Isso indica que o desvio do segmento ST é um bom indicativo da presença da doença, sendo um marcador importante para diagnóstico clínico.

```r
# Selecione apenas as variáveis numéricas de interesse:
vars_num <- c("Age", "RestingBP", "Cholesterol", "MaxHR", "Oldpeak")

# Gera a matriz de dispersão
pairs(df[ , vars_num],
      main = "Matriz de Dispersao das Variaveis Numericas",
      pch = 19,      # símbolo de plotagem
      cex = 0.5,     # tamanho dos pontos
      col = "blue")  # cor dos pontos
```

2) Gráficos de dispersão



Observando a matriz de dispersão acima, podemos extrair algumas impressões gerais sobre as relações e possíveis correlações entre as variáveis numéricas do conjunto:

1. Age vs. MaxHR

Nota-se uma tendência inversamente proporcional: à medida que a idade aumenta, a frequência cardíaca máxima tende a diminuir. Isso faz sentido fisiologicamente (máximo de batimentos por minuto costuma diminuir com o envelhecimento).

2. Age vs. RestingBP / Age vs. Cholesterol

Os pontos formam uma nuvem difusa, sem um alinhamento marcante que indique forte correlação linear. Pode haver leve tendência de aumento da pressão arterial com a idade, mas não é muito pronunciada pelos gráficos.

3. RestingBP vs. Cholesterol

Também exibe dispersão relativamente ampla. Não é possível visualizar um padrão claro de associação linear; há casos de colesterol elevado com pressão arterial tanto alta quanto moderada.

4. Oldpeak vs. as demais variáveis

A maioria dos pontos se concentra próxima ao valor 0 (sem grande depressão/elevação do ST), e surgem alguns outliers com valores altos ou negativos. Não se vê um padrão linear forte com Age, Cholesterol ou RestingBP, sugerindo pouca correlação direta.

Com MaxHR também não há grande alinhamento.

5. MaxHR vs. RestingBP / MaxHR vs. Cholesterol

Observa-se grande espalhamento dos pontos, sem relação linear evidente. Podem existir tendências mais sutis ou não lineares, mas, se houver, não são claras no gráfico de dispersão.

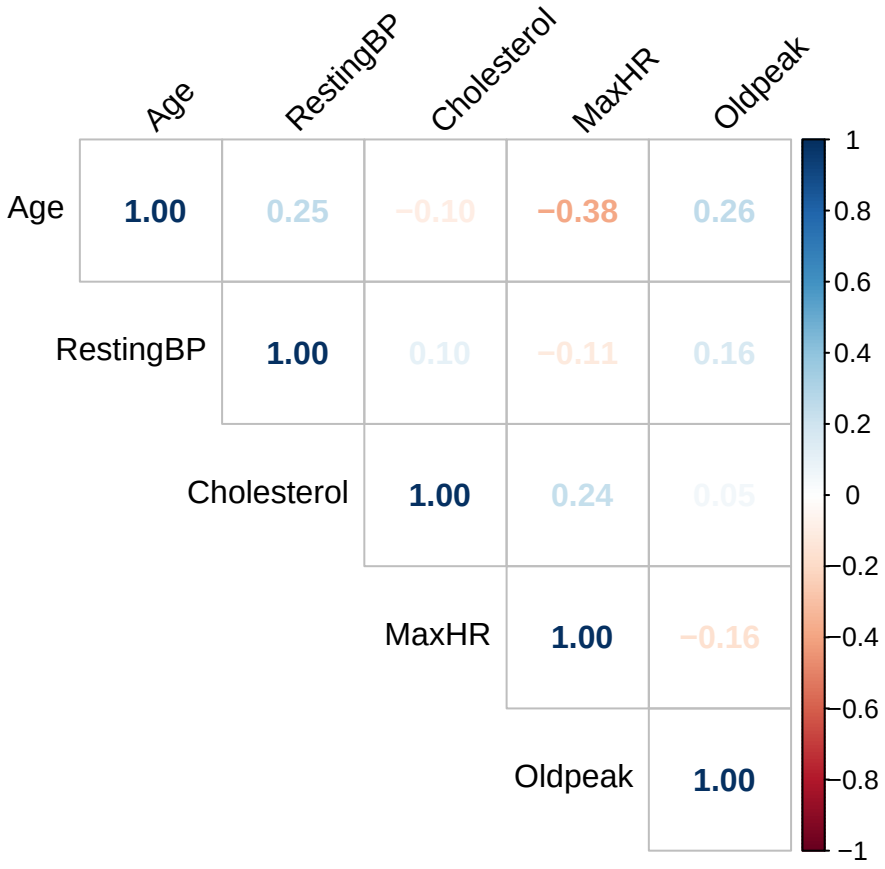
6. Dispersão geral

A maioria dos pares apresenta nuvens de pontos sem alinhamento marcante, sugerindo associações fracas ou inexistentes em termos de linearidade. É claro que, para confirmar, seria ideal calcular coeficientes de correlação e eventualmente testar relações não lineares.

Item 3: (0.5 pts.) Calcule correlações lineares de Pearson entre as variáveis contínuas (faça um gráfico de correlações), comente.

Solução:

```r
df_num <- df[, vars_num]

corr_mat <- cor(df_num, method = "pearson", use = "complete.obs")

library(corrplot)

corrplot(corr_mat, method = "number",
         type = "upper",
         tl.col = "black",
         tl.srt = 45)
```



Na matriz de correlação acima vemos que:

1. Age vs. MaxHR é a correlação de maior magnitude – ainda assim, moderada e negativa. Isso significa que pessoas mais velhas tendem a ter frequência cardíaca máxima menor.
2. Age vs. RestingBP e Age vs. Oldpeak mostram correlações positivas porém fracas. Ou seja, a pressão arterial de repouso e o desvio ST podem aumentar levemente com a idade, mas sem grande intensidade.
3. As demais correlações ficam entre -0,16 e +0,24, indicando relações lineares muito fracas, isso é, próximas de zero. Por exemplo, RestingBP e Cholesterol pouco se correlacionam, assim como Cholesterol e MaxHR .

Nenhum par apresenta correlação forte acima de 0,7 ou abaixo de -0,7, sugerindo que não há multicolinearidade severa entre essas variáveis numéricas e que as associações lineares são em geral modestas.

Item 4: (0.5 pt.) Proponha um modelo de regressão normal linear com todas as variáveis explicativas.Comente!

Solução:

A ideia é considerar RestingBP como a variável resposta (dependente) e todas as outras variáveis como explanatórias.

```r
fit_full <- lm(RestingBP ~ Age + Sex + ChestPainType + Cholesterol +
                          FastingBS + RestingECG + MaxHR + ExerciseAngina +
                          Oldpeak + ST_Slope + HeartDisease,
               data = df)

summary(fit_full)
```

```
##
## Call:
## lm(formula = RestingBP ~ Age + Sex + ChestPainType + Cholesterol +
##     FastingBS + RestingECG + MaxHR + ExerciseAngina + Oldpeak +
##     ST_Slope + HeartDisease, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.321  -10.815   -2.189    9.212   66.917
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            106.767518   7.002473  15.247  < 2e-16
## Age                                      0.415338   0.071593   5.801  9.1e-09
## SexFeminino                              0.490402   1.525490   0.321 0.747927
## ChestPainTypeAngina Atipica             -2.705286   3.012559  -0.898 0.369424
## ChestPainTypeDor Nao Anginosa           -4.504395   2.903954  -1.551 0.121223
## ChestPainTypeAssintomatico              -5.124268   2.834624  -1.808 0.070980
## Cholesterol                              0.022128   0.005933   3.730 0.000204
## FastingBS>120                            2.420005   1.488275   1.626 0.104289
## RestingECGST-T wave abnormality          2.893179   1.557652   1.857 0.063580
## RestingECGLeft Ventricular Hypertrophy   0.669568   1.566360   0.427 0.669141
## MaxHR                                   -0.020745   0.028502  -0.728 0.466907
## ExerciseAnginaSim                        3.667165   1.504643   2.437 0.014992
## Oldpeak                                  1.649325   0.678880   2.429 0.015315
## ST_Slope.L                              -4.422694   1.963249  -2.253 0.024515
## ST_Slope.Q                              -1.650014   1.287877  -1.281 0.200456
## HeartDiseaseSim                          0.608255   1.794102   0.339 0.734666
##
## (Intercept)                            ***
## Age                                    ***
## SexFeminino
## ChestPainTypeAngina Atipica
## ChestPainTypeDor Nao Anginosa
## ChestPainTypeAssintomatico             .
## Cholesterol                            ***
## FastingBS>120
## RestingECGST-T wave abnormality        .
## RestingECGLeft Ventricular Hypertrophy
## MaxHR
## ExerciseAnginaSim                      *
## Oldpeak                                *
## ST_Slope.L                             *
## ST_Slope.Q
## HeartDiseaseSim
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.6 on 902 degrees of freedom
## Multiple R-squared:  0.1107, Adjusted R-squared:  0.09591
## F-statistic: 7.486 on 15 and 902 DF,  p-value: 7.288e-16
```

Note que Age, Cholesterol, ExerciseAngina, Oldpeak e o efeito linear de ST_Slope têm associação estatisticamente significativa com a pressão arterial de repouso, embora os coeficientes sejam em geral modestos. O modelo, apesar de estatisticamente significativo, explica apenas cerca de 10–11% da variação em RestingBP, sugerindo que muitos outros fatores (genéticos, ambiente, medicações, entre outros) influenciam essa medida. Em termos de aplicação, se o objetivo for prever RestingBP, esse modelo não é muito robusto (dada a baixa R^2). Se for para verificar que algumas variáveis (Age, Cholesterol, etc.) exercem algum efeito, ainda que modesto, então o modelo cumpre esse papel.

Item 5: (0.8 pts.) Faça uma análise de Multicolinearidade. Comente!

Solução:

Como ja calculamos a matriz de correlação anteriormente, vamos calcular Fatores de inflação de variância (VFI)

```r
# 3. Cálculo dos VIFs
library(car)
vif_values <- vif(fit_full)
print(vif_values)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Age             1.349445  1        1.161656
## Sex             1.144602  1        1.069861
## ChestPainType   1.648388  3        1.086867
## Cholesterol     1.246085  1        1.116282
## FastingBS       1.172988  1        1.083045
## RestingECG      1.171618  2        1.040391
## MaxHR           1.558250  1        1.248299
## ExerciseAngina 1.614984  1        1.270820
## Oldpeak         1.551376  1        1.245543
## ST_Slope        2.099851  2        1.203780
## HeartDisease    2.356579  1        1.535115
```

Note que embora exista alguma correlação entre os preditores, os GVIFs indicam que nenhuma variável está excessivamente inflando a variância dos coeficientes. Portanto, a multicolinearidade não parece ser um grande obstáculo para este conjunto de dados. Agora, vamos calcular o Índice de Condição, que são autovalores da matriz de design (X) para obter o condition index, que também indica a presença de multicolinearidade (índices muito altos sugerem problemas).

```r
X <- model.matrix(fit_full)
eigen_values <- eigen(t(X) %*% X)$values
condition_index <- sqrt(max(eigen_values) / eigen_values)
print(condition_index)
```

```
##  [1]    1.000000    4.423263   17.734421  235.658285  389.711165  495.041000
##  [7]  593.315020  613.822647  645.458529  680.941220  729.447089  792.009453
## [13]  903.311357  940.347021 1852.262236 3382.410460
```

No vetores acima, há vários índices bastante altos (na casa de centenas ou milhares), o que sinaliza forte a severa multicolinearidade no modelo. Em outras palavras, há combinações de variáveis que são quase linearmente dependentes, deixando o modelo sensível com coeficientes instáveis, erros-padrão inflado). Agora, vamos calcular a Regressão Ridge.

```
library(MASS)
ridge_model <- lm.ridge(RestingBP ~ Age + Cholesterol + MaxHR + Oldpeak,
                        data = df,
                        lambda = seq(0, 0.2, 0.001))

dim_coef <- dim(coef(ridge_model))
print(dim_coef)
```

```
## [1] 201   5
```

```
print(length(ridge_model$lambda))
```

```
## [1] 201
```

```
df_coef <- as.data.frame(coef(ridge_model))
df_coef$lambda <- ridge_model$lambda

# Reorganiza as colunas para que lambda seja a primeira
df_coef <- df_coef[, c("lambda", setdiff(names(df_coef), "lambda"))]

# Visualize as primeiras linhas para conferir
head(df_coef)
```

```
##       lambda       V1       Age Cholesterol       MaxHR  Oldpeak
## 0.000  0.000 106.8636 0.4454818  0.02156174 -0.02943850 1.618059
## 0.001  0.001 106.8636 0.4454813  0.02156171 -0.02943852 1.618058
## 0.002  0.002 106.8636 0.4454808  0.02156169 -0.02943854 1.618058
## 0.003  0.003 106.8637 0.4454803  0.02156166 -0.02943856 1.618057
## 0.004  0.004 106.8637 0.4454798  0.02156164 -0.02943857 1.618057
## 0.005  0.005 106.8638 0.4454793  0.02156161 -0.02943859 1.618056
```

```
# Cria o gráfico interativo com plotly:
p <- plot_ly(df_coef, x = ~lambda)
for(col in names(df_coef)[-1]) {
  p <- p %>% add_lines(y = as.formula(paste0("~", col)), name = col)
}
p <- p %>% layout(title = "Trajetória dos Coeficientes na Regressão Ridge",
                  xaxis = list(title = "Lambda"),
                  yaxis = list(title = "Coeficientes"))
```

Age (coeficiente positivo): Indica que, a cada aumento de 1 ano de idade, a pressão arterial estimada (RestingBP) tende a subir em torno de “0.44 mmHg”, segurando as demais variáveis constantes. Cholesterol (coeficiente positivo, ~0.02): Sugere que cada aumento de 1 mg/dL de colesterol aumenta ligeiramente a pressão arterial. MaxHR (coeficiente negativo, ~-0.02): Cada batimento cardíaco máximo adicional está associado a uma redução pequena na pressão arterial estimada. Oldpeak (coeficiente positivo, ~1.61): Para cada 1 unidade de desvio ST, a pressão arterial de repouso aumenta, em média, cerca de 1.6 mmHg.

Com os resultados acima e com a matriz de correlação calculada anteriormente, vemos que embora haja alguma correlação entre as variáveis explicativas, os resultados dos testes (VIF/GVIF, matrizes de correlação e índices de condição) sugerem que a multicolinearidade está em níveis aceitáveis. Isso significa que os coeficientes estimados podem ser interpretados com uma confiança razoável, sem que haja uma inflacão excessiva na variância dos estimadores.

Item 6: (0.8 pts.) Aplique o método do StepWise para verificar qual o melhor ajuste. Comente!

Solução:

```
step_model <- step(fit_full, direction = "both")
```

```
## Start:  AIC=5281.73
## RestingBP ~ Age + Sex + ChestPainType + Cholesterol + FastingBS +
##     RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
##     HeartDisease
##
##                  Df Sum of Sq    RSS    AIC
## - Sex             1      32.0 279559 5279.8
## - HeartDisease    1      35.6 279562 5279.8
## - ChestPainType   3    1323.0 280850 5280.1
## - MaxHR           1     164.2 279691 5280.3
## - RestingECG      2    1069.4 280596 5281.2
## <none>                        279527 5281.7
## - FastingBS       1     819.4 280346 5282.4
## - ST_Slope        2    1628.4 281155 5283.1
## - Oldpeak         1    1829.1 281356 5285.7
## - ExerciseAngina  1    1840.8 281368 5285.8
## - Cholesterol     1    4311.2 283838 5293.8
## - Age             1   10429.9 289957 5313.4
##
## Step:  AIC=5279.83
## RestingBP ~ Age + ChestPainType + Cholesterol + FastingBS + RestingECG +
##     MaxHR + ExerciseAngina + Oldpeak + ST_Slope + HeartDisease
##
##                  Df Sum of Sq    RSS    AIC
## - HeartDisease    1      25.0 279584 5277.9
## - ChestPainType   3    1320.9 280880 5278.2
## - MaxHR           1     155.5 279714 5278.3
## - RestingECG      2    1066.5 280625 5279.3
## <none>                        279559 5279.8
## - FastingBS       1     812.7 280372 5280.5
## - ST_Slope        2    1629.1 281188 5281.2
## + Sex             1      32.0 279527 5281.7
## - ExerciseAngina  1    1818.9 281378 5283.8
## - Oldpeak         1    1828.5 281387 5283.8
## - Cholesterol     1    4449.5 284008 5292.3
## - Age             1   10543.4 290102 5311.8
##
## Step:  AIC=5277.91
```

```
## RestingBP ~ Age + ChestPainType + Cholesterol + FastingBS + RestingECG +
##     MaxHR + ExerciseAngina + Oldpeak + ST_Slope
##
##                   Df Sum of Sq    RSS    AIC
## - ChestPainType    3    1297.1 280881 5276.2
## - MaxHR            1     162.7 279747 5276.4
## - RestingECG       2    1066.8 280651 5277.4
## <none>                         279584 5277.9
## - FastingBS        1     884.9 280469 5278.8
## - ST_Slope         2    1612.4 281196 5279.2
## + HeartDisease     1      25.0 279559 5279.8
## + Sex              1      21.4 279562 5279.8
## - Oldpeak          1    1921.8 281506 5282.2
## - ExerciseAngina   1    1956.5 281540 5282.3
## - Cholesterol      1    4473.5 284057 5290.5
## - Age              1   10616.0 290200 5310.1
##
## Step:  AIC=5276.16
## RestingBP ~ Age + Cholesterol + FastingBS + RestingECG + MaxHR +
##     ExerciseAngina + Oldpeak + ST_Slope
##
##                   Df Sum of Sq    RSS    AIC
## - MaxHR            1      67.4 280948 5274.4
## - RestingECG       2    1097.6 281979 5275.7
## <none>                         280881 5276.2
## - FastingBS        1     844.7 281726 5276.9
## - ST_Slope         2    1651.5 282532 5277.5
## + ChestPainType    3    1297.1 279584 5277.9
## + Sex              1      31.1 280850 5278.1
## + HeartDisease     1       1.2 280880 5278.2
## - ExerciseAngina   1    1435.9 282317 5278.8
## - Oldpeak          1    1892.5 282774 5280.3
## - Cholesterol      1    4890.3 285771 5290.0
## - Age              1   10847.3 291728 5308.9
##
## Step:  AIC=5274.38
## RestingBP ~ Age + Cholesterol + FastingBS + RestingECG + ExerciseAngina +
##     Oldpeak + ST_Slope
##
##                   Df Sum of Sq    RSS    AIC
## - RestingECG       2    1126.4 282075 5274.1
## <none>                         280948 5274.4
## - FastingBS        1     835.0 281783 5275.1
## - ST_Slope         2    1618.2 282567 5275.7
## + MaxHR            1      67.4 280881 5276.2
## + Sex              1      23.5 280925 5276.3
## + HeartDisease     1       0.0 280948 5276.4
## + ChestPainType    3    1201.7 279747 5276.4
## - ExerciseAngina   1    1670.6 282619 5277.8
## - Oldpeak          1    1843.6 282792 5278.4
## - Cholesterol      1    4837.6 285786 5288.1
## - Age              1   12705.6 293654 5313.0
##
## Step:  AIC=5274.06
```

```
## RestingBP ~ Age + Cholesterol + FastingBS + ExerciseAngina +
##     Oldpeak + ST_Slope
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         282075 5274.1
## + RestingECG       2    1126.4 280948 5274.4
## - FastingBS        1    1005.3 283080 5275.3
## - ST_Slope         2    1657.1 283732 5275.4
## + MaxHR            1      96.1 281979 5275.7
## + Sex              1      19.4 282055 5276.0
## + HeartDisease     1       0.0 282075 5276.1
## + ChestPainType    3    1220.8 280854 5276.1
## - ExerciseAngina   1    1875.4 283950 5278.1
## - Oldpeak          1    1904.1 283979 5278.2
## - Cholesterol      1    4698.5 286773 5287.2
## - Age              1   14227.2 296302 5317.2
```

```
# 3. Exibir o resumo do modelo selecionado
summary(step_model)
```

```
##
## Call:
## lm(formula = RestingBP ~ Age + Cholesterol + FastingBS + ExerciseAngina +
##     Oldpeak + ST_Slope, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.698  -11.024   -1.287    9.145   69.464
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       99.247504   3.826993  25.934  < 2e-16 ***
## Age                0.446863   0.065959   6.775 2.23e-11 ***
## Cholesterol        0.021674   0.005567   3.893 0.000106 ***
## FastingBS>120      2.625373   1.457820   1.801 0.072051 .
## ExerciseAnginaSim  3.398077   1.381482   2.460 0.014089 *
## Oldpeak            1.645129   0.663770   2.478 0.013375 *
## ST_Slope.L        -4.416714   1.930551  -2.288 0.022378 *
## ST_Slope.Q        -1.720661   1.205351  -1.428 0.153774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.61 on 910 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.09569
## F-statistic: 14.86 on 7 and 910 DF,  p-value: < 2.2e-16
```

O método Stepwise, aplicado com base no critério AIC, foi usado para reduzir o modelo completo e identificar um conjunto de preditores que, juntos, fornecem o melhor compromisso entre ajuste e simplicidade. O processo eliminou variáveis que não contribuíam significativamente para explicar a variação em RestingBP. O modelo final resultante inclui as seguintes variáveis:

- Age
- Cholesterol

- FastingBS
- ExerciseAngina
- Oldpeak
- ST_Slope

Item 7: (0.8 pts.) No modelo escolhido pelo StepWise faça a interpretação dos parâmetros.

Solução:

No modelo final escolhido pelo stepwise temos que:

- Age: Para cada aumento de 1 ano na idade, mantendo as demais variáveis constantes, a pressão arterial em repouso aumenta, em média, cerca de 0.45 mmHg. Esse efeito é altamente significativo, indicando que a idade exerce um efeito positivo sobre a RestingBP.
- Cholesterol: Para cada aumento de 1 mg/dL no colesterol sérico, a RestingBP aumenta em média 0.022 mmHg, mantendo as outras variáveis constantes. Embora o efeito seja pequeno por unidade, mudanças maiores em colesterol podem ter impacto acumulado, e o efeito é estatisticamente significativo.
- FastingBS: Este parâmetro indica o efeito da categoria de FastingBS. Pacientes com FastingBS $>$ 120 mg/dL têm, em média, uma RestingBP 2.63 mmHg maior do que aqueles com FastingBS $<=$ 120 mg/dL, mantendo os demais fatores constantes. O p-valor é marginal, sugerindo efeito de tendência.
- ExerciseAnginaSim: Indica que pacientes com angina induzida por exercício (sim) têm, em média, uma pressão arterial de repouso 3.40 mmHg maior do que os pacientes sem angina induzida, com os demais fatores controlados. Esse efeito é estatisticamente significativo.
- Oldpeak: Cada aumento de 1 unidade em Oldpeak (o desvio do segmento ST) está associado a um acréscimo de aproximadamente 1.65 mmHg em RestingBP, mantendo as demais variáveis constantes. Esse efeito é estatisticamente significativo.
- ST_Slope.L: O contraste linear da variável ordinal ST_Slope (que pode representar, por exemplo, uma ordem de "Up", "Flat" e "Down") apresenta um coeficiente negativo de -4.42. Isso significa que, conforme avançamos na ordem linear (ou seja, passando de uma categoria de melhor para pior, de acordo com a codificação dos contrastes), há uma redução de 4.42 mmHg na RestingBP, em média. Esse efeito é estatisticamente significativo.
- ST_Slope.Q: O contraste quadrático para ST_Slope tem coeficiente -1.72, mas com p-valor de aproximadamente 0.154, indicando que o efeito não é estatisticamente significativo. Assim, o componente quadrático não contribui de forma relevante para o modelo.

Item 8: (0.8 pts.) Faça uma anova entre os 2 modelos, com todas as variáveis e aquele com as variáveis dependentes escolhidas pelo StepWise. Comente!

Solução

```
anova_result <- anova(step_model, fit_full)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: RestingBP ~ Age + Cholesterol + FastingBS + ExerciseAngina +
##     Oldpeak + ST_Slope
## Model 2: RestingBP ~ Age + Sex + ChestPainType + Cholesterol + FastingBS +
##     RestingECG + MaxHR + ExerciseAngina + Oldpeak + ST_Slope +
##     HeartDisease
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    910 282075
## 2    902 279527  8    2547.9 1.0277 0.4131
```

A hipótese nula nesse teste é de que o modelo reduzido (step_model) se ajusta tão bem quanto o modelo completo (fit_full) para explicar RestingBP. Como o p-value é alto, não rejeitamos a hipótese nula, indicando que o modelo completo não é estatisticamente superior ao modelo reduzido.

Isso significa que há ausência de ganho significativo de ajuste, isso é, as variáveis excluídas não contribuem de forma significativa para reduzir o erro do modelo. Dado que o modelo reduzido não perde poder explicativo de forma estatisticamente relevante e ainda apresenta menor AIC, ele é mais parcimonioso. Logo, é preferível ao modelo completo.

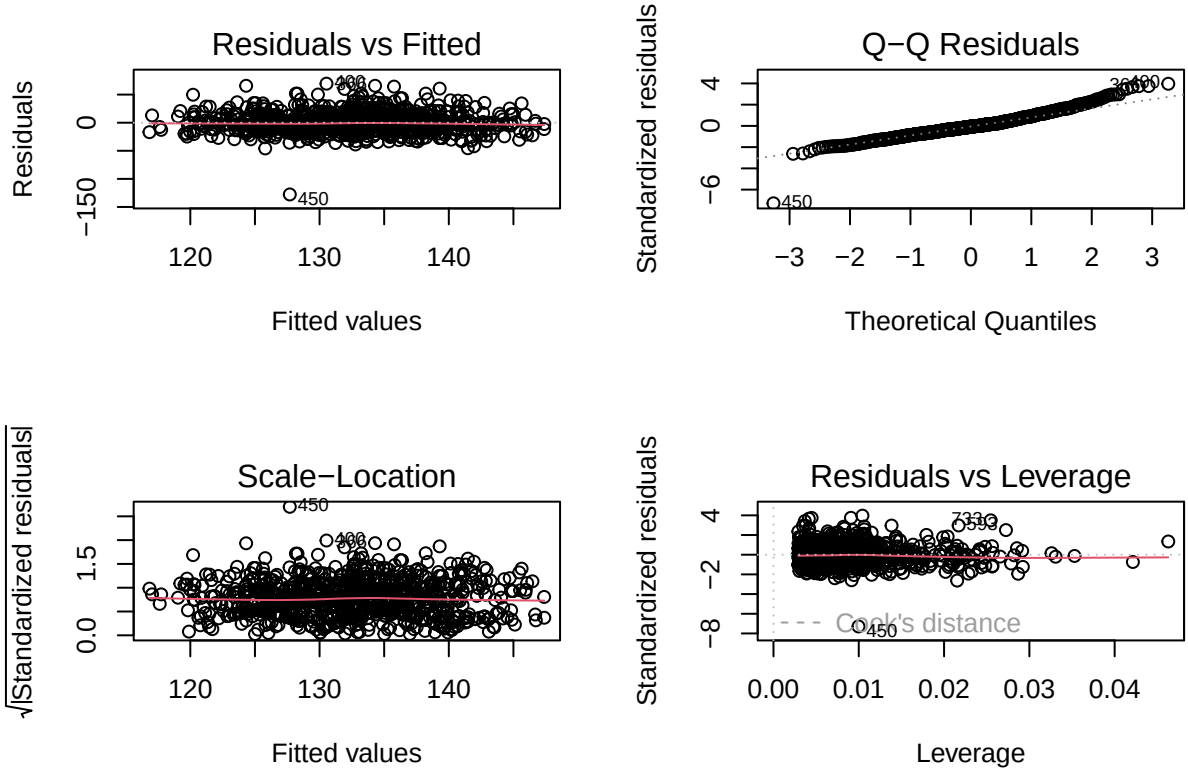
O modelo final com Age, Cholesterol, FastingBS, ExerciseAngina, Oldpeak e ST_Slope explica a variação em RestingBP tão bem quanto o modelo maior, sem sobrecarregar com variáveis que não agregam melhora significativa.

Item 9: (0.8 pts.) Para o modelo escolhido em 5) aplicar as análises de resíduos verificando as condições de normalidade e heterocedasticidade (Usar os programas postados na segunda aula).Comente todas elas!

Solução:

Acredito que a questão se refere ao modelo escolhido no item 6, após aplicar o stepwise.

```r
par(mfrow = c(2, 2))
plot(step_model)
```



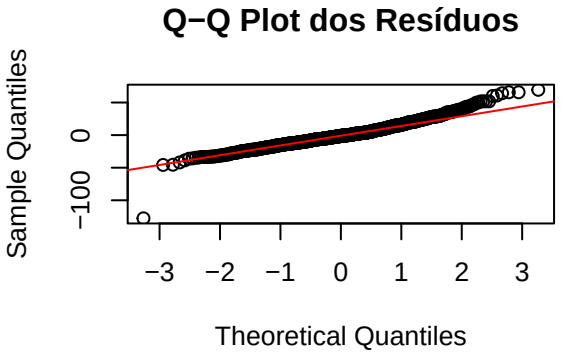

```r
# 2. Análise de Normalidade dos Resíduos

## 2.1 Gráfico Q-Q (normal quantile-quantile plot)
qqnorm(residuals(step_model), main = "Q-Q Plot dos Resíduos")
qqline(residuals(step_model), col = "red")

## 2.2 Teste de Shapiro-Wilk
normality_test <- shapiro.test(residuals(step_model))
print(normality_test)
```

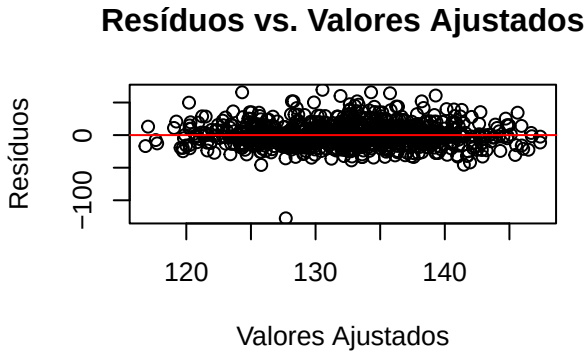
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step_model)
## W = 0.96012, p-value = 4.051e-15
```

```r
# 3. Análise de Homocedasticidade

## 3.1 Gráfico de Resíduos vs. Valores Ajustados
plot(fitted(step_model), residuals(step_model),
     xlab = "Valores Ajustados",
     ylab = "Resíduos",
     main = "Resíduos vs. Valores Ajustados")
abline(h = 0, col = "red")
```

```r
## 3.2 Teste de Breusch-Pagan
library(lmtest)
bp_test <- bptest(step_model)
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  step_model
## BP = 8.2432, df = 7, p-value = 0.3116
```





1. Normalidade de Resíduos:

- Q-Q Plot: Observa-se que os pontos formam uma curva, desviando-se da linha reta nos extremos (caudas). Isso sugere que os resíduos não seguem exatamente uma distribuição normal.
- Teste de Shapiro-Wilk possui um p-valor extremamente baixo indica que rejeitamos a hipótese nula de normalidade dos resíduos. Em outras palavras, os resíduos não são normais segundo esse teste.

2. Homoscedasticidade (Variância Constante)

- Resíduos vs. Valores Ajustados: Não se nota um padrão de "funil" ou crescimento/dispersão sistemática; a distribuição dos pontos é relativamente homogênea em torno de zero.

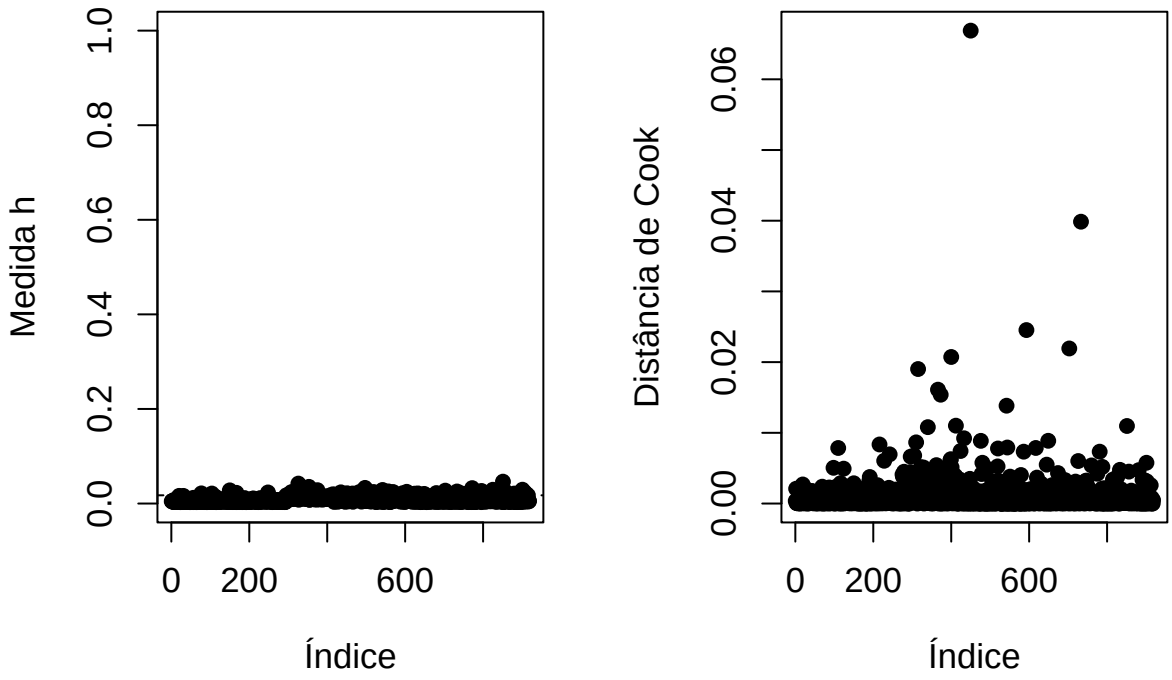
- Teste de Breusch-Pagan: Como o p-valor é maior que 0,05, não há evidência de heterocedasticidade. Logo, não rejeitamos a hipótese de variância constante dos resíduos.

Logo, temos que o *Pressuposto de Normalidade* foi Violado. O modelo linear clássico (OLS) supõe resíduos normais, mas aqui esse pressuposto não foi atendido. Dependendo do tamanho da amostra, o Teorema Central do Limite pode amenizar esse problema para inferência.

Pressuposto de Homocedasticidade foi atendido. Não há indicação de variância não constante ao longo dos valores ajustados.

Em síntese, o modelo não apresenta problemas de heterocedasticidade, mas não satisfaz a normalidade dos resíduos, podendo demandar ajustes ou métodos alternativos para a análise mais rigorosa.

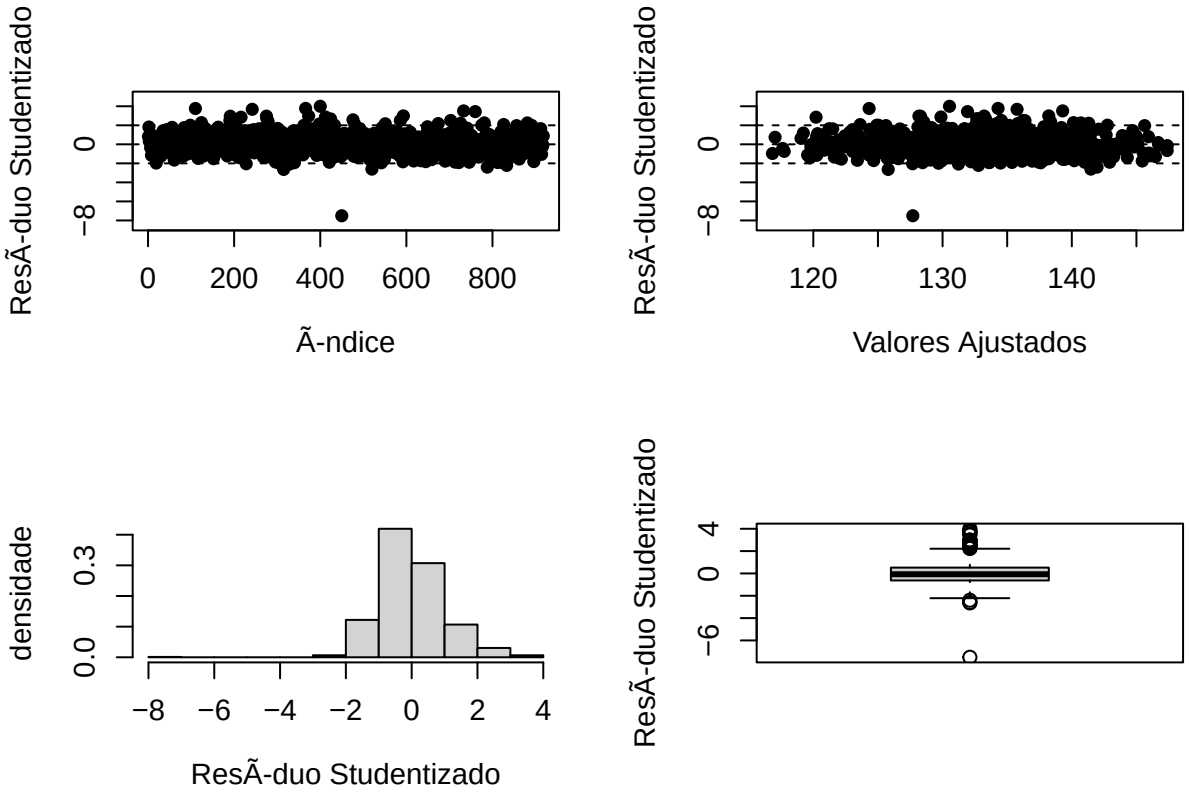
```
source("anainflu_norm.R")

# 1. Para identificar observações influentes e ver medidas associadas
anainflu_norm(step_model)
```



Em relação a Medida h (Leverage), temos que há observações muito acima da média são consideradas de alta alavancagem. No gráfico, quase todos os pontos estão em valores baixos, sugerindo que não há muitas ou quase nenhuma observações com alavancagem extrema.

Pontos com distância de Cook muito alta são altamente influentes. Neste caso, a maioria das distâncias está concentrada perto de 0, com algumas um pouco maiores, mas ainda distantes dos limiares usuais de alerta. Isso indica que nenhum ponto se destaca como fortemente influente para distorcer o ajuste.

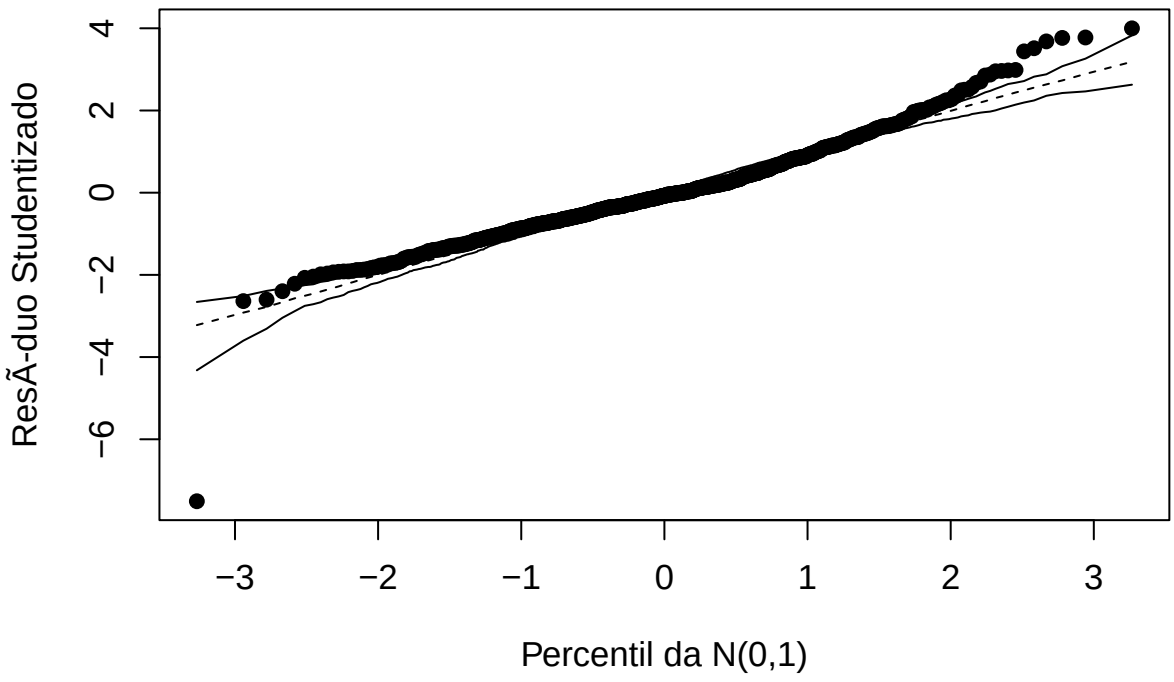
```
source("Diag2.norm.R")
diag2norm(step_model)
```



- Resíduo Studentizado vs. Índice

Mostra como os resíduos estão distribuídos ao longo das observações (índice). A maioria dos pontos se mantém na faixa de -2 a 2, porém há um ou outro ponto ligeiramente fora de ± 3. Esses pontos podem ser considerados outliers em termos de resíduo e merecem investigação. A ausência de um padrão crescente ou decrescente indica que não há grandes problemas de correlação serial ou mudança sistemática ao longo das observações.

```
source("Envel_norm.R")

envelnorm(step_model)
```



Pelo gráfico acima vemos que a maior parte dos pontos está relativamente próxima da reta, sugerindo que os resíduos, no centro da distribuição, não se afastam tanto da normal.

Entretanto, há pontos nas caudas que desviam bastante (particularmente na cauda esquerda em torno de –6 e na direita acima de +3), indicando outliers ou alguma assimetria nas caudas.

Portanto, apesar de os resíduos não seguirem perfeitamente a distribuição normal (violando o pressuposto de normalidade), a condição de homocedasticidade está satisfeita. Isso significa que, embora o modelo apresente resíduos com distribuição não normal, a variância dos erros é constante. Dependendo do objetivo da análise e do tamanho da amostra, essa violação da normalidade pode ser menos preocupante para a predição, mas pode afetar a validade dos testes estatísticos. Em situações onde a normalidade é crucial, pode ser considerado aplicar transformações ou métodos robustos para melhorar o ajuste.

Item 10: (0.8 pts.) Caso algum pressuposto do modelo de regressão normal linear falhe, tente fazer alguma transformação (vistas em aula) com o objetivo de atingir os pressupostos do modelo. Comente!

Solução:

Como podemos ver no item anterior, o pressuposto de normalidade não foi antendido.

1. Transformação Logaritmica:

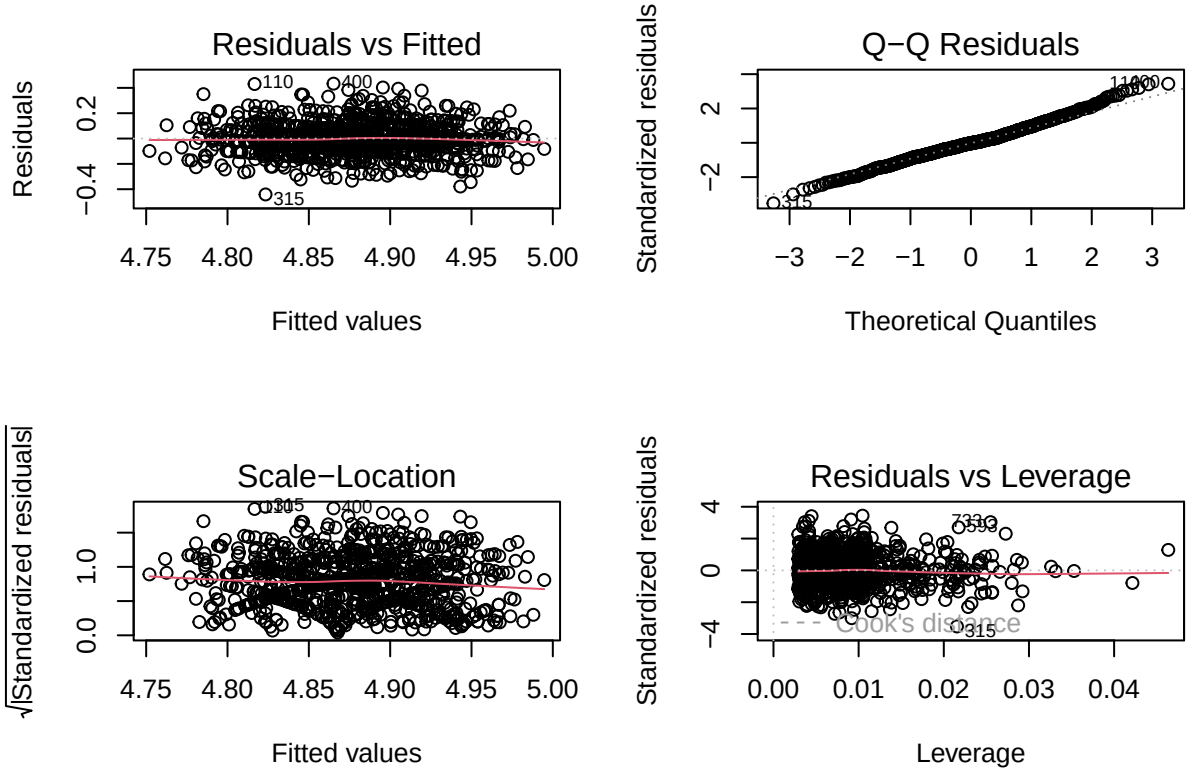
Uma estratégia bastante comum para resolver a violação do pressuposto de normalidade é transformar a variável resposta. No caso, como os resíduos do modelo original (com RestingBP) não seguem uma

distribuição normal, podemos aplicar, por exemplo, uma transformação logarítmica à variável resposta. Essa abordagem foi discutida em aula e pode ser aplicada diretamente ao modelo.

```r
# Ajuste do modelo com a transformação logarítmica da variável resposta
df_clean <- df[df$RestingBP > 0, ]

step_model_log <- lm(log(RestingBP) ~ Age + Cholesterol + FastingBS
                     + ExerciseAngina + Oldpeak + ST_Slope, data = df_clean)

# Exibir o resumo do novo modelo
summary(step_model_log)
```

```
##
## Call:
## lm(formula = log(RestingBP) ~ Age + Cholesterol + FastingBS +
##     ExerciseAngina + Oldpeak + ST_Slope, data = df_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44135 -0.07972 -0.00377  0.07407  0.43313
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.627e+00  2.752e-02 168.126  < 2e-16 ***
## Age                3.422e-03  4.742e-04   7.217 1.12e-12 ***
## Cholesterol        1.508e-04  4.012e-05   3.759 0.000181 ***
## FastingBS>120      1.465e-02  1.049e-02   1.397 0.162878
## ExerciseAnginaSim  2.092e-02  9.945e-03   2.103 0.035715 *
## Oldpeak            1.382e-02  4.774e-03   2.894 0.003896 **
## ST_Slope.L        -3.774e-02  1.388e-02  -2.719 0.006664 **
## ST_Slope.Q        -1.767e-02  8.672e-03  -2.037 0.041903 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1266 on 909 degrees of freedom
## Multiple R-squared:  0.1106, Adjusted R-squared:  0.1037
## F-statistic: 16.14 on 7 and 909 DF,  p-value: < 2.2e-16
```

```r
# Análise dos resíduos do modelo transformado

## 1. Gráficos de diagnóstico básicos
par(mfrow = c(2, 2))
plot(step_model_log)
```



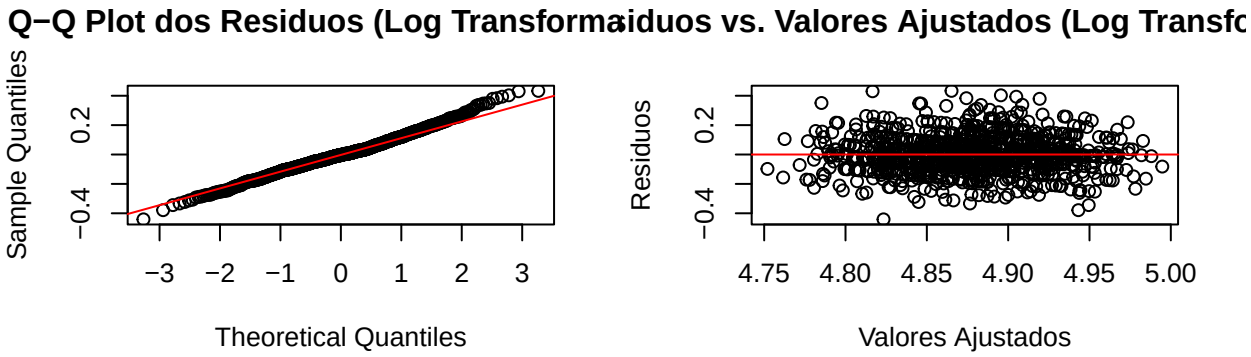

```
## 2. Gráfico Q-Q dos resíduos
qqnorm(residuals(step_model_log), main = "Q-Q Plot dos Residuos (Log Transformado)")
qqline(residuals(step_model_log), col = "red")

## 3. Teste de Shapiro-Wilk para normalidade
normality_test_log <- shapiro.test(residuals(step_model_log))
print(normality_test_log)
```

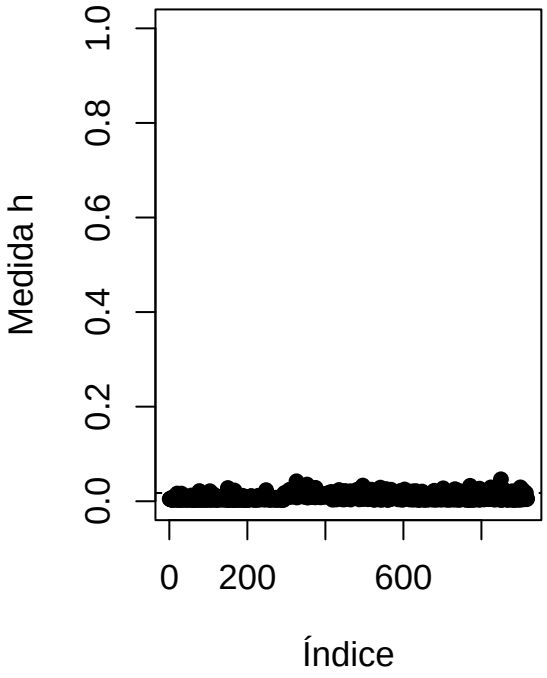
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step_model_log)
## W = 0.99363, p-value = 0.0006021
```

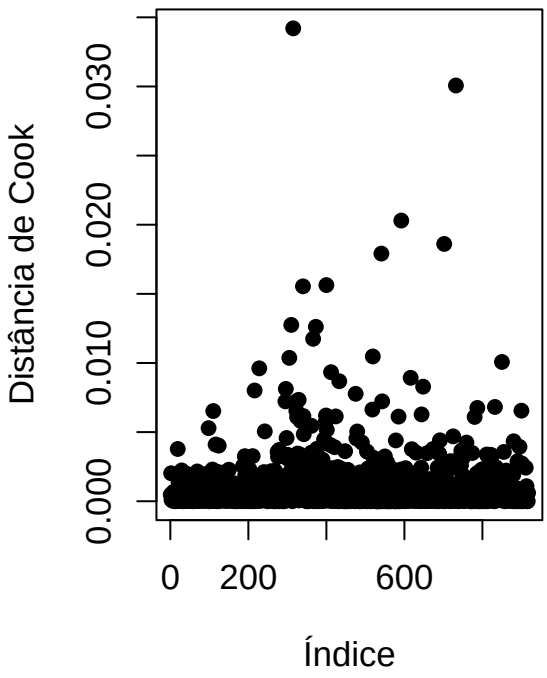
```
## 4. Gráfico de Resíduos vs. Valores Ajustados (para verificar homocedasticidade)
plot(fitted(step_model_log), residuals(step_model_log),
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     main = "Residuos vs. Valores Ajustados (Log Transformado)")
abline(h = 0, col = "red")

## 5. Teste de Breusch-Pagan para homocedasticidade
library(lmtest)
bp_test_log <- bptest(step_model_log)
print(bp_test_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  step_model_log
## BP = 9.5648, df = 7, p-value = 0.2146
```

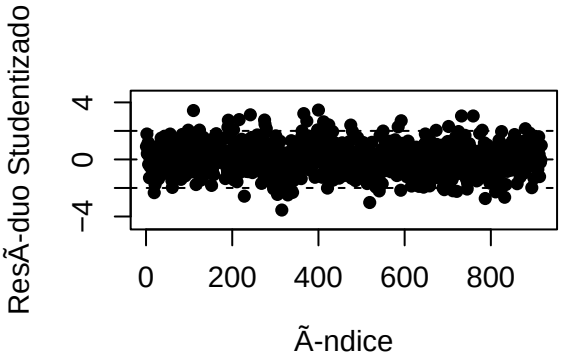
```
anainflu_norm(step_model_log)
```

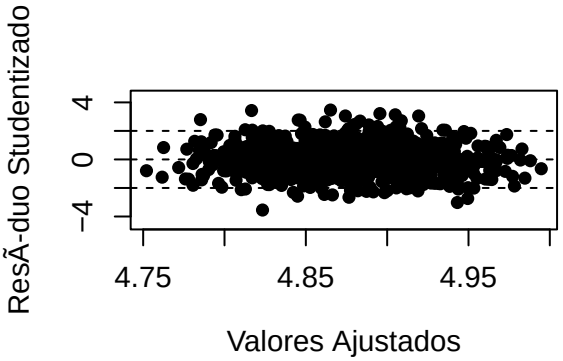


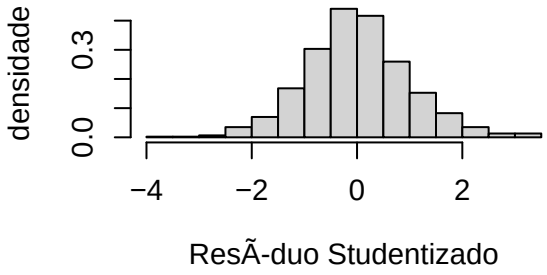


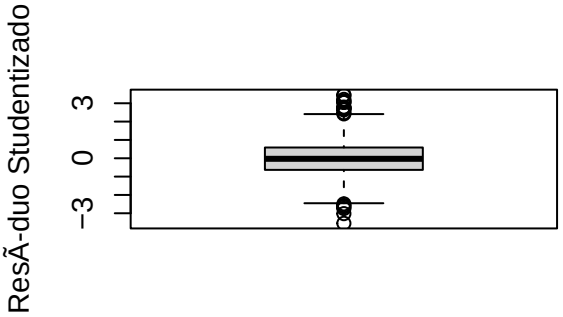



```
diag2norm(step_model_log)
```

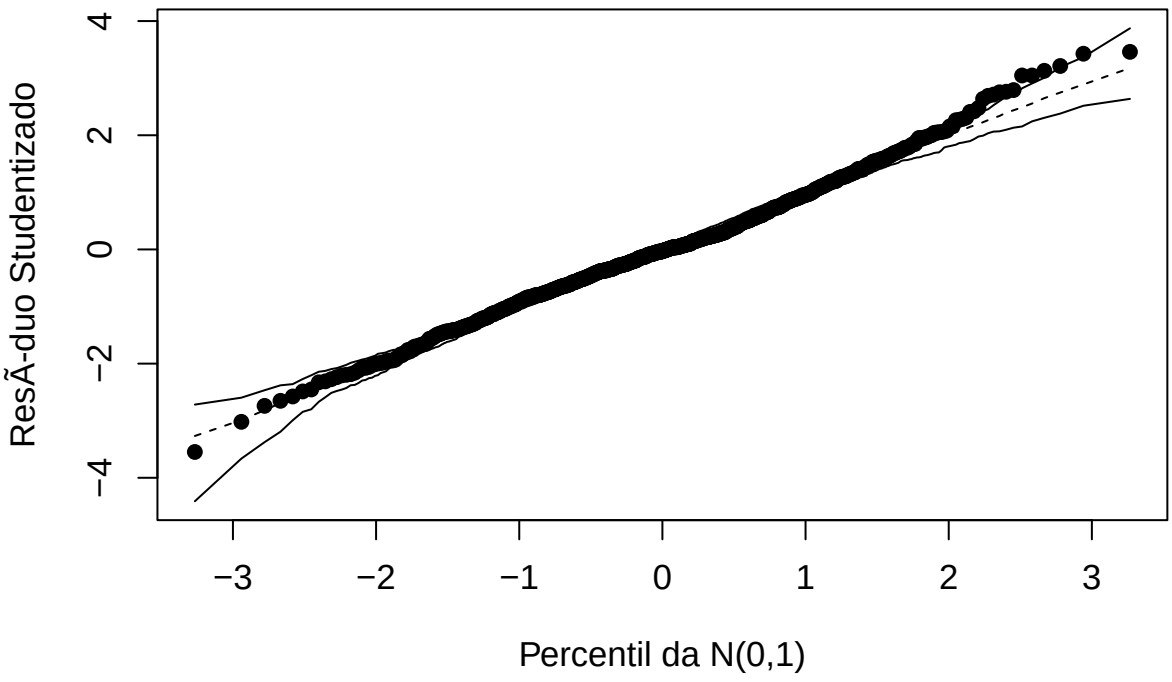









```
envelnorm(step_model_log)
```



Note que a transformação logarítmica melhorou a distribuição dos resíduos (em comparação ao modelo original), mas ainda não atingiu uma normalidade "ideal". O modelo transformado segue estatisticamente significativo e mantém homocedasticidade, com interpretação dos coeficientes em termos de variação percentual na pressão arterial de repouso.

2. Centralização das Variáveis:

Outra alternativa é utilizar a centralização de variáveis, isto é, subtraindo a média de cada variável. Assim, os coeficientes terão uma interpretação onde o intercepto corresponde à média da RestingBP para um paciente "médio" (ou seja, com as variáveis centrais igual a zero). Em nosso caso, considerando o modelo selecionado pelo Stepwise:

```
# Vamos supor que df_clean é o seu data frame já filtrado para RestingBP > 0
df_center <- df_clean

# Centralizando as variáveis numéricas
df_center$Age_c <- df_center$Age - mean(df_center$Age, na.rm = TRUE)
df_center$Cholesterol_c <- df_center$Cholesterol - mean(df_center$Cholesterol,
                                                       na.rm = TRUE)
df_center$Oldpeak_c <- df_center$Oldpeak - mean(df_center$Oldpeak,
                                                na.rm = TRUE)

# Podemos conferir as médias (elas devem ser próximas de zero)
mean(df_center$Age_c, na.rm = TRUE)          # deve ser 0
```

```
## [1] 1.67811e-15
```

```
mean(df_center$Cholesterol_c, na.rm = TRUE)   # deve ser 0
```

```
## [1] 1.126297e-14
```

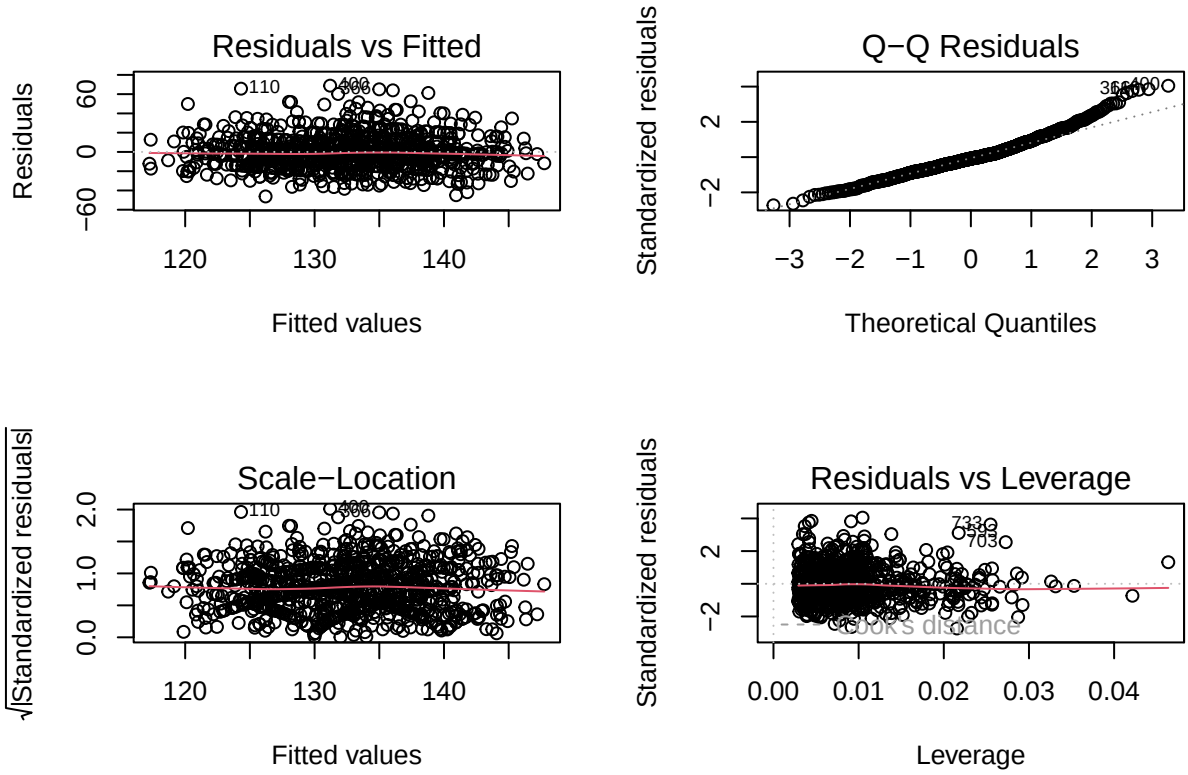
```
mean(df_center$Oldpeak_c, na.rm = TRUE)       # deve ser 0
```

```
## [1] -4.80791e-17
```

```
centered_model <- lm(RestingBP ~ Age_c + Cholesterol_c + FastingBS +
                     ExerciseAngina + Oldpeak_c + ST_Slope,
                     data = df_center)

summary(centered_model)
```

```
##
## Call:
## lm(formula = RestingBP ~ Age_c + Cholesterol_c + FastingBS +
##     ExerciseAngina + Oldpeak_c + ST_Slope, data = df_center)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.213 -10.972  -1.492   9.001  68.798
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       129.298395   1.089349 118.693  < 2e-16 ***
## Age_c               0.447611   0.064040   6.990 5.33e-12 ***
## Cholesterol_c       0.018865   0.005418   3.482 0.000521 ***
## FastingBS>120       2.220209   1.416424   1.567 0.117352
## ExerciseAnginaSim   2.883050   1.343031   2.147 0.032083 *
## Oldpeak_c           1.791338   0.644747   2.778 0.005576 **
## ST_Slope.L         -4.445880   1.874374  -2.372 0.017903 *
## ST_Slope.Q         -2.049709   1.171094  -1.750 0.080412 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.09 on 909 degrees of freedom
## Multiple R-squared:  0.105,  Adjusted R-squared:  0.09814
## F-statistic: 15.24 on 7 and 909 DF,  p-value: < 2.2e-16
```

```
# Análise dos resíduos do modelo transformado

## 1. Gráficos de diagnóstico básicos
par(mfrow = c(2, 2))
plot(centered_model)
```

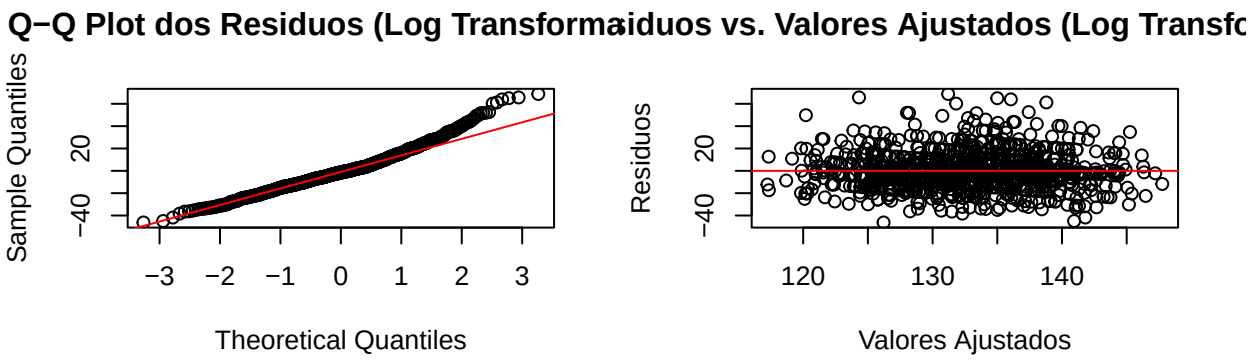



```
## 2. Gráfico Q-Q dos resíduos
qqnorm(residuals(centered_model), main = "Q-Q Plot dos Residuos (Log Transformado)")
qqline(residuals(centered_model), col = "red")

## 3. Teste de Shapiro-Wilk para normalidade
normality_test_center <- shapiro.test(residuals(centered_model))
print(normality_test_center)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(centered_model)
## W = 0.97463, p-value = 1.493e-11
```

```
## 4. Gráfico de Resíduos vs. Valores Ajustados (para verificar homocedasticidade)
plot(fitted(centered_model), residuals(centered_model),
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     main = "Residuos vs. Valores Ajustados (Log Transformado)")
abline(h = 0, col = "red")

## 5. Teste de Breusch-Pagan para homocedasticidade
library(lmtest)
bp_test_center <- bptest(centered_model)
print(bp_test_center)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  centered_model
## BP = 8.9535, df = 7, p-value = 0.256
```



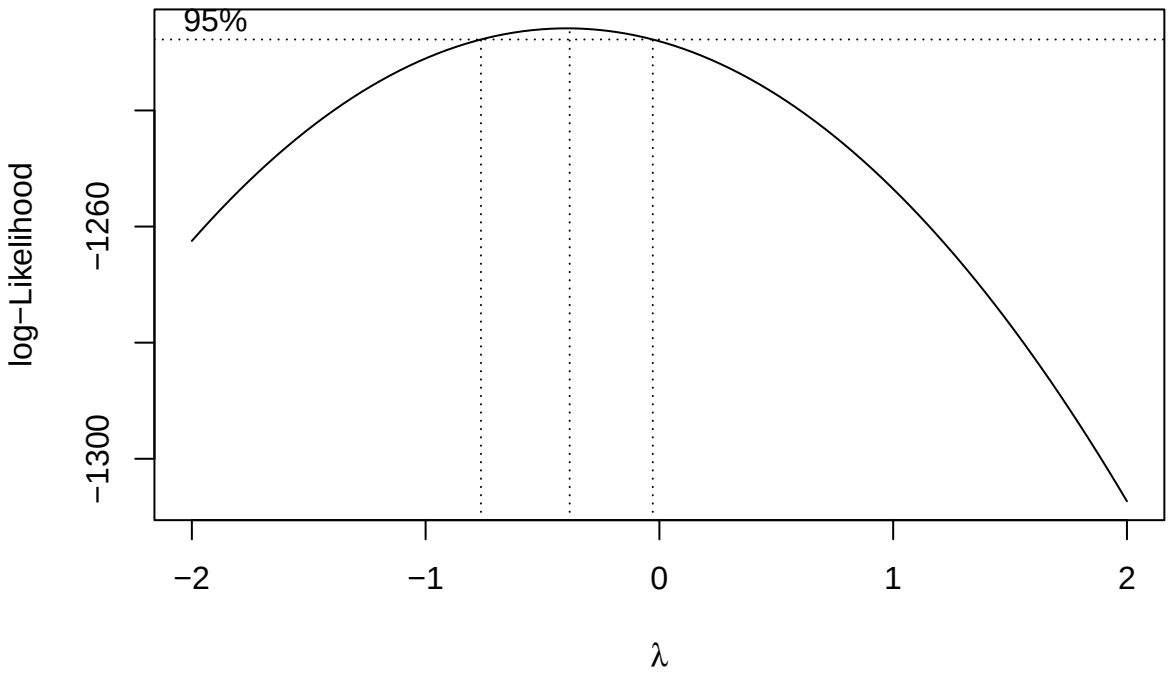
A centralização cumpre seu papel de melhorar a interpretação do intercepto e possivelmente reduzir colinearidades entre variáveis, mas não resolve a violação da normalidade.

A homocedasticidade permanece atendida, o que é positivo para a confiabilidade das estimativas de mínimos quadrados.

Item 11: (0.8 pts.) Caso nenhuma transformação funcione, tente utilizar Box-Cox. Comente!

Solução:

```r
# Ajuste o modelo completo com as variáveis selecionadas (usando df_clean)
modelo_temp <- lm(RestingBP ~ Age + Cholesterol + FastingBS + ExerciseAngina
                  + Oldpeak + ST_Slope, data = df_clean)

bc <- boxcox(modelo_temp, plotit = TRUE)
```




```r
# Encontre o lambda que maximiza o log-verossímil:
lambda_opt <- bc$x[which.max(bc$y)]
cat("Lambda ótimo:", lambda_opt, "\n")
```

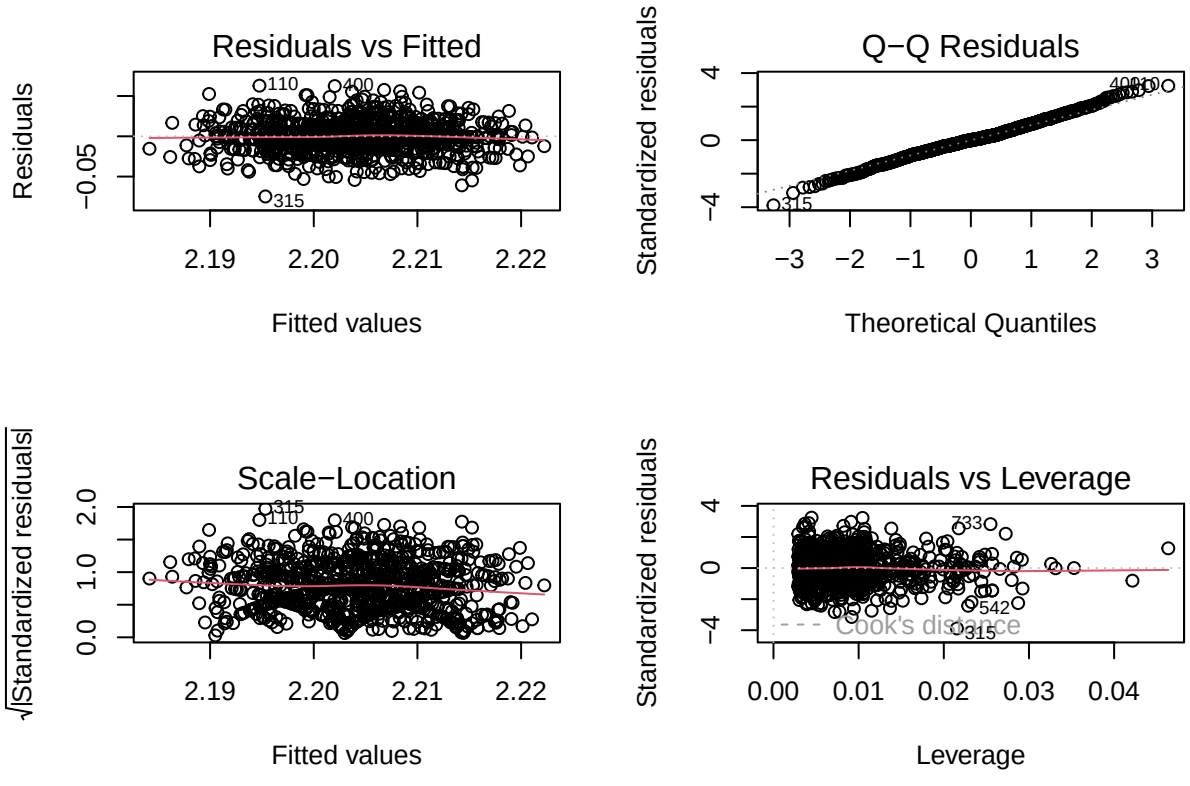
```
## Lambda ótimo: -0.3838384
```

```r
if(abs(lambda_opt) < 0.001) {
  df_clean$Trans_RestingBP <- log(df_clean$RestingBP)
} else {
  df_clean$Trans_RestingBP <- (df_clean$RestingBP^lambda_opt - 1) / lambda_opt
}

# Ajuste o modelo utilizando a variável resposta transformada
bc_model <- lm(Trans_RestingBP ~ Age + Cholesterol + FastingBS + ExerciseAngina
               + Oldpeak + ST_Slope, data = df_clean)
summary(bc_model)
```

```
##
## Call:
## lm(formula = Trans_RestingBP ~ Age + Cholesterol + FastingBS +
##     ExerciseAngina + Oldpeak + ST_Slope, data = df_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.074660 -0.011964 -0.000169  0.011616  0.062823
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.165e+00  4.222e-03 512.875  < 2e-16 ***
## Age                 5.292e-04  7.274e-05   7.276 7.45e-13 ***
## Cholesterol         2.374e-05  6.154e-06   3.858 0.000122 ***
## FastingBS>120       2.131e-03  1.609e-03   1.325 0.185653
## ExerciseAnginaSim   3.167e-03  1.525e-03   2.076 0.038179 *
## Oldpeak             2.140e-03  7.323e-04   2.923 0.003554 **
## ST_Slope.L         -6.053e-03  2.129e-03  -2.843 0.004565 **
## ST_Slope.Q         -2.853e-03  1.330e-03  -2.145 0.032243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01942 on 909 degrees of freedom
## Multiple R-squared:  0.1121, Adjusted R-squared:  0.1052
## F-statistic: 16.39 on 7 and 909 DF,  p-value: < 2.2e-16
```

```r
# Análise dos resíduos do modelo transformado:
par(mfrow = c(2, 2))
plot(bc_model)
```




```r
qqnorm(residuals(bc_model), main = "Q-Q Plot dos Residuos (Box-Cox)")
qqline(residuals(bc_model), col = "red")
```

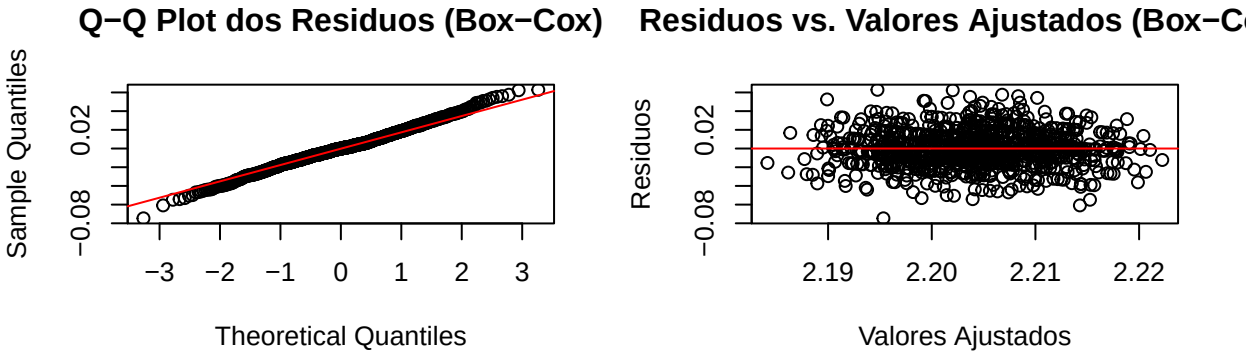
```r
# Teste de Shapiro-Wilk para normalidade dos resíduos
normality_test_bc <- shapiro.test(residuals(bc_model))
print(normality_test_bc)
```

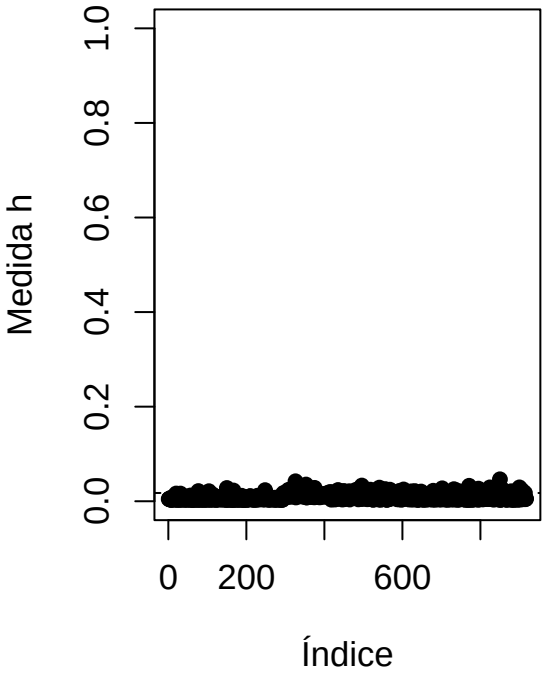
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bc_model)
## W = 0.99572, p-value = 0.01211
```

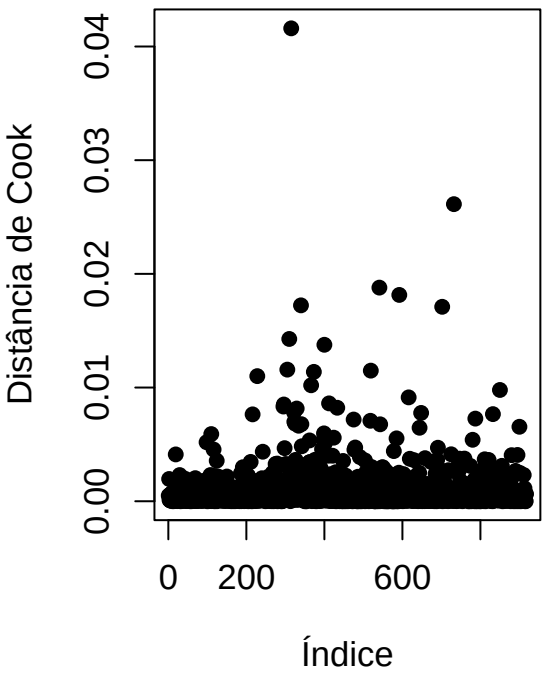
```r
# Gráfico de Resíduos vs. Valores Ajustados para verificar homocedasticidade
plot(fitted(bc_model), residuals(bc_model),
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     main = "Residuos vs. Valores Ajustados (Box-Cox)")
abline(h = 0, col = "red")

# Teste de Breusch-Pagan para heterocedasticidade
library(lmtest)
bp_test_bc <- bptest(bc_model)
print(bp_test_bc)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bc_model
## BP = 10.696, df = 7, p-value = 0.1524
```

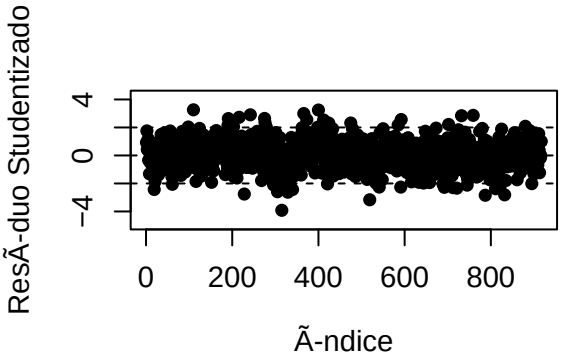
```r
anainflu_norm(bc_model)
```

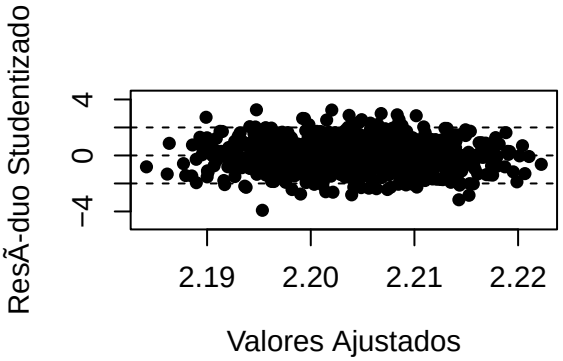

Q−Q Plot dos Residuos (Box−Cox)
Sample Quantiles
−0.08
0.02
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Residuos vs. Valores Ajustados (Box−C
Residuos
−0.08
0.02
2.19
2.20
2.21
2.22
Valores Ajustados

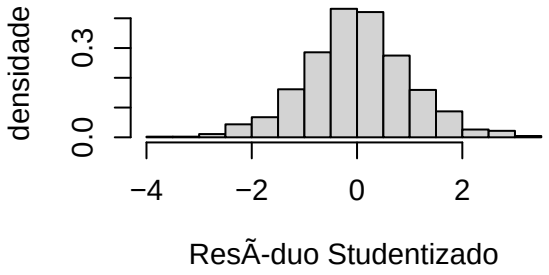


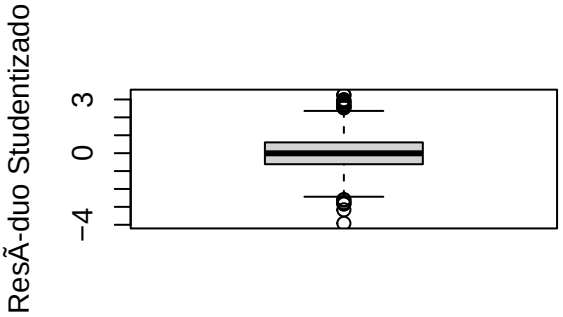



```
diag2norm(bc_model)
```

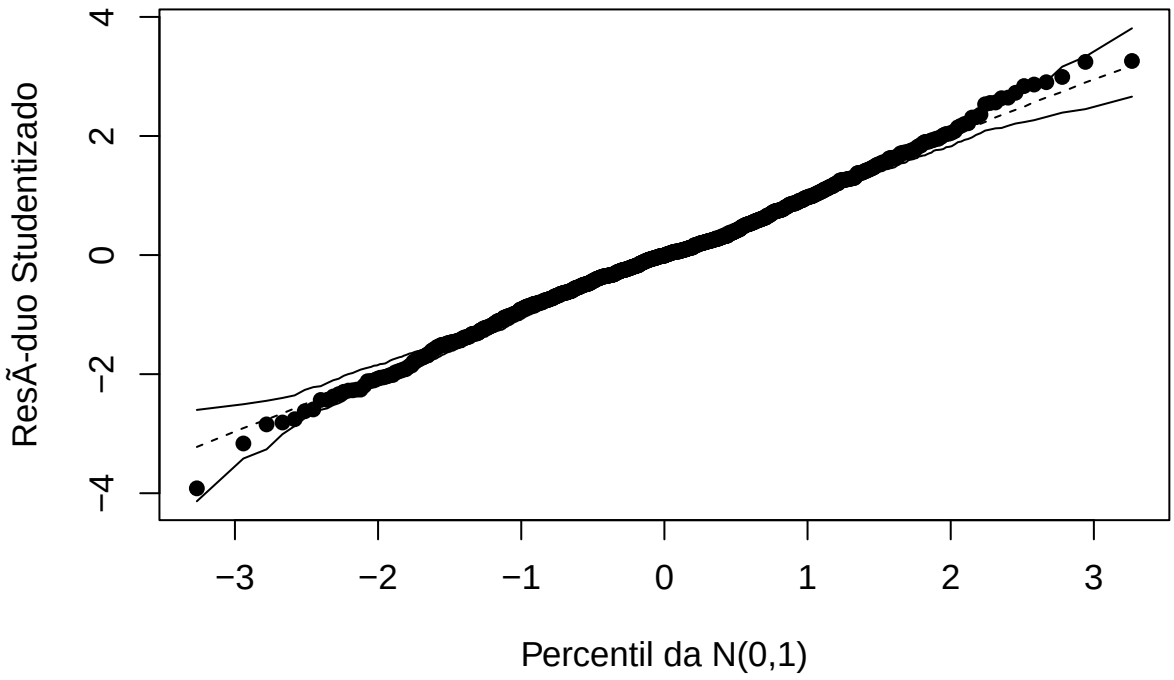









```
envelnorm(bc_model)
```



Com base nos resultados acima, podemos concluir que a aplicação da transformação Box-Cox foi eficaz para aproximar os resíduos de uma distribuição normal, melhorando os pressupostos do modelo de regressão linear. Note que o teste de Shapiro-Wilk apresentou W = 0.99572 e um p-valor de 0.01211. Isso indica que, sob o nível de significância convencional de 0.05, a hipótese nula de normalidade dos resíduos é rejeitada – ou seja, os resíduos não seguem perfeitamente uma distribuição normal. Contudo, o valor de W está muito próximo de 1, o que sugere que os resíduos estão bastante próximos de uma distribuição normal. Em amostras grandes, mesmo pequenos desvios podem resultar em p-valores significativos, mas podem não comprometer as inferências do modelo. Assim, embora tecnicamente haja uma violação do pressuposto de normalidade, a transformação Box-Cox melhorou consideravelmente a distribuição dos resíduos e, na prática, o modelo pode ser considerado aceitável para análises inferenciais.

Item 12: (0.8 pts.) Eliminar individualmente os pontos mais discrepantes (se existirem), e verificar se houve mudança inferencial. Comente!

Solução:

Primeiramente vamos verificar se há outliers.

```
outliers <- unique(c(
  which(abs(rstudent(bc_model)) > 3),
  which(cooks.distance(bc_model) > 0.05),
  which(hatvalues(bc_model) > (2 * (length(coef(bc_model))) / nrow(df_clean)))
))

print(outliers)
```

```
##  [1] 110 315 400 519  77 104 150 167 248 304 308 310 323 324 325 326 330 331 338
## [20] 340 341 342 353 375 414 419 433 435 441 450 462 464 477 481 483 492 496 497
## [39] 500 504 508 518 521 534 537 541 543 547 556 559 568 570 578 592 598 603 616
## [58] 624 632 648 679 702 732 733 734 739 748 750 771 782 795 806 814 824 850 858
## [77] 879 900 904 907 908 914
```

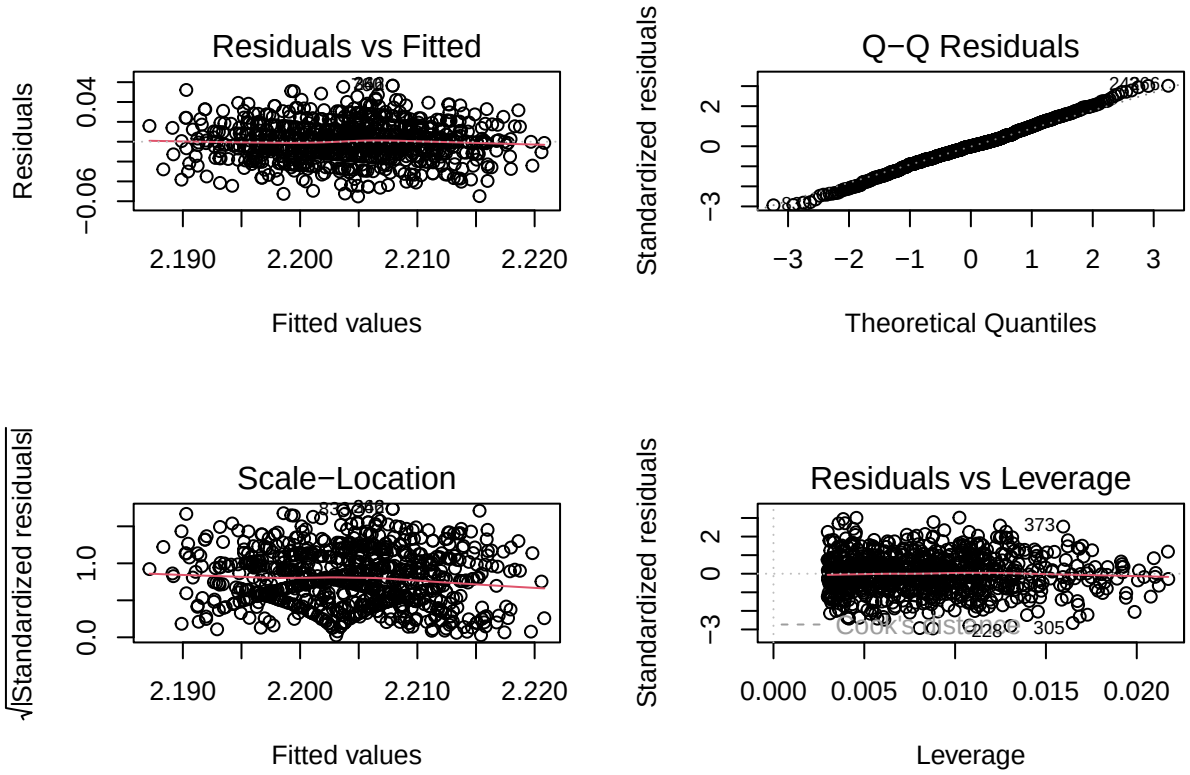
Agora, vamos remover os outliers e ver como fica a performance do modelo.

```
df_sem_outliers <- df_clean[-outliers, ]
bc_model_sem_outliers <- lm(Trans_RestingBP ~ Age + Cholesterol + FastingBS +
                            ExerciseAngina + Oldpeak + ST_Slope,
                            data = df_sem_outliers)
summary(bc_model_sem_outliers)
```

```
##
## Call:
## lm(formula = Trans_RestingBP ~ Age + Cholesterol + FastingBS +
##     ExerciseAngina + Oldpeak + ST_Slope, data = df_sem_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05518 -0.01174 -0.00007  0.01108  0.05648
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.170e+00  4.178e-03 519.288  < 2e-16 ***
## Age                5.194e-04  7.412e-05   7.007 5.03e-12 ***
## Cholesterol        2.002e-05  6.552e-06   3.055  0.00232 **
## FastingBS>120      2.427e-03  1.679e-03   1.445  0.14870
## ExerciseAnginaSim  3.944e-03  1.593e-03   2.476  0.01348 *
## Oldpeak            1.243e-03  8.477e-04   1.466  0.14290
## ST_Slope.L        -3.620e-04  1.130e-03  -0.320  0.74872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01882 on 828 degrees of freedom
## Multiple R-squared:  0.1032, Adjusted R-squared:  0.09666
## F-statistic: 15.87 on 6 and 828 DF,  p-value: < 2.2e-16
```

A remoção dos outliers não alterou drasticamente os coeficientes das variáveis principais, mas reduziu a variância residual e corrigiu possíveis inferências enviesadas. O modelo está mais estável e a inferência agora é mais confiável, com menor influência de pontos extremos. Vamos realizar a análise de resíduos.

```
# Análise dos resíduos do modelo transformado:
par(mfrow = c(2, 2))
plot(bc_model_sem_outliers)
```



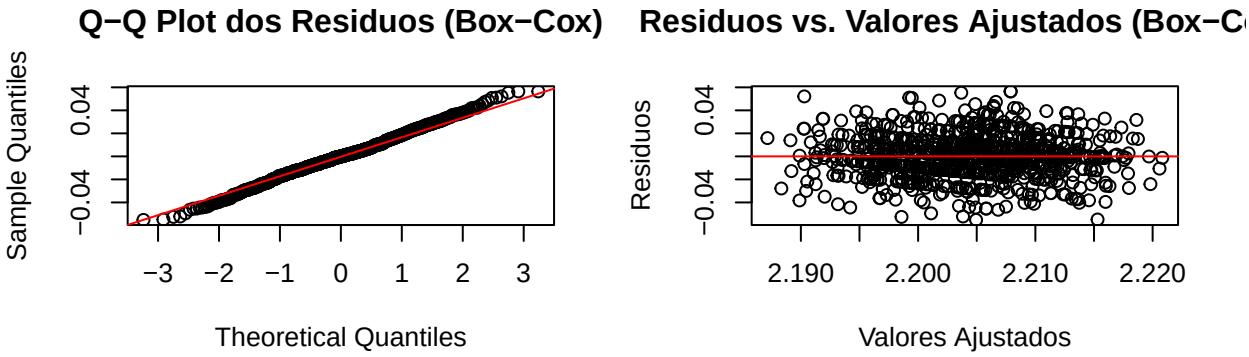

```r
qqnorm(residuals(bc_model_sem_outliers), main = "Q-Q Plot dos Residuos (Box-Cox)")
qqline(residuals(bc_model_sem_outliers), col = "red")

# Teste de Shapiro-Wilk para normalidade dos resíduos
normality_test_bc <- shapiro.test(residuals(bc_model_sem_outliers))
print(normality_test_bc)
```

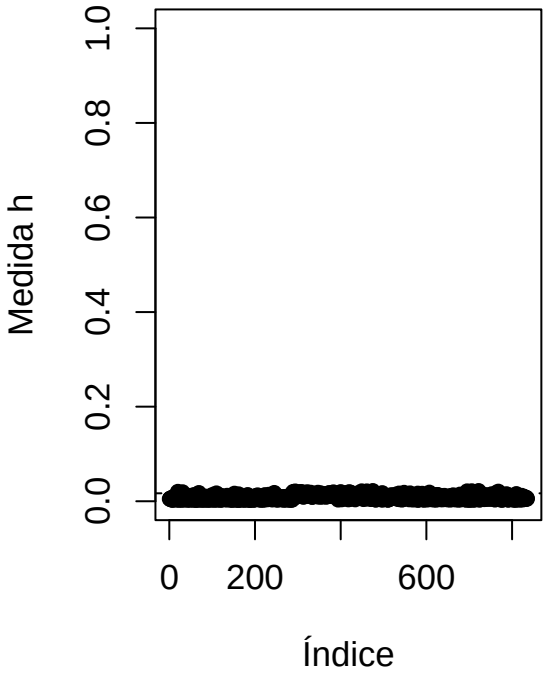
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bc_model_sem_outliers)
## W = 0.99689, p-value = 0.1054
```

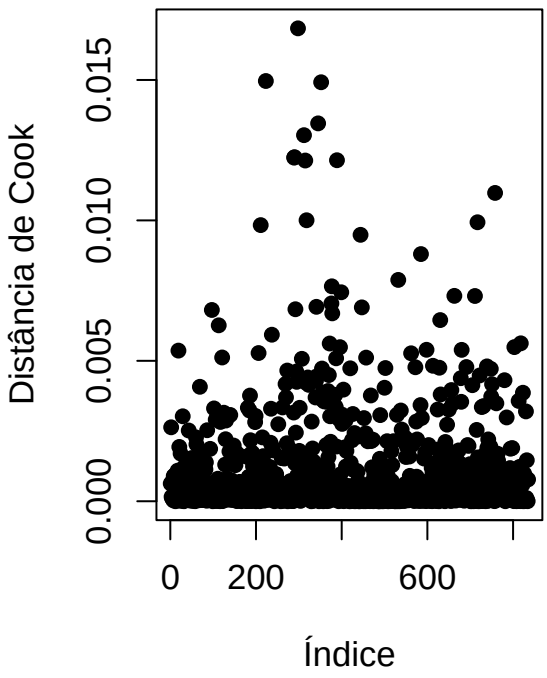
```r
# Gráfico de Resíduos vs. Valores Ajustados para verificar homocedasticidade
plot(fitted(bc_model_sem_outliers), residuals(bc_model_sem_outliers),
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     main = "Residuos vs. Valores Ajustados (Box-Cox)")
abline(h = 0, col = "red")

# Teste de Breusch-Pagan para heterocedasticidade
library(lmtest)
bp_test_bc <- bptest(bc_model_sem_outliers)
print(bp_test_bc)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bc_model_sem_outliers
## BP = 6.791, df = 6, p-value = 0.3406
```

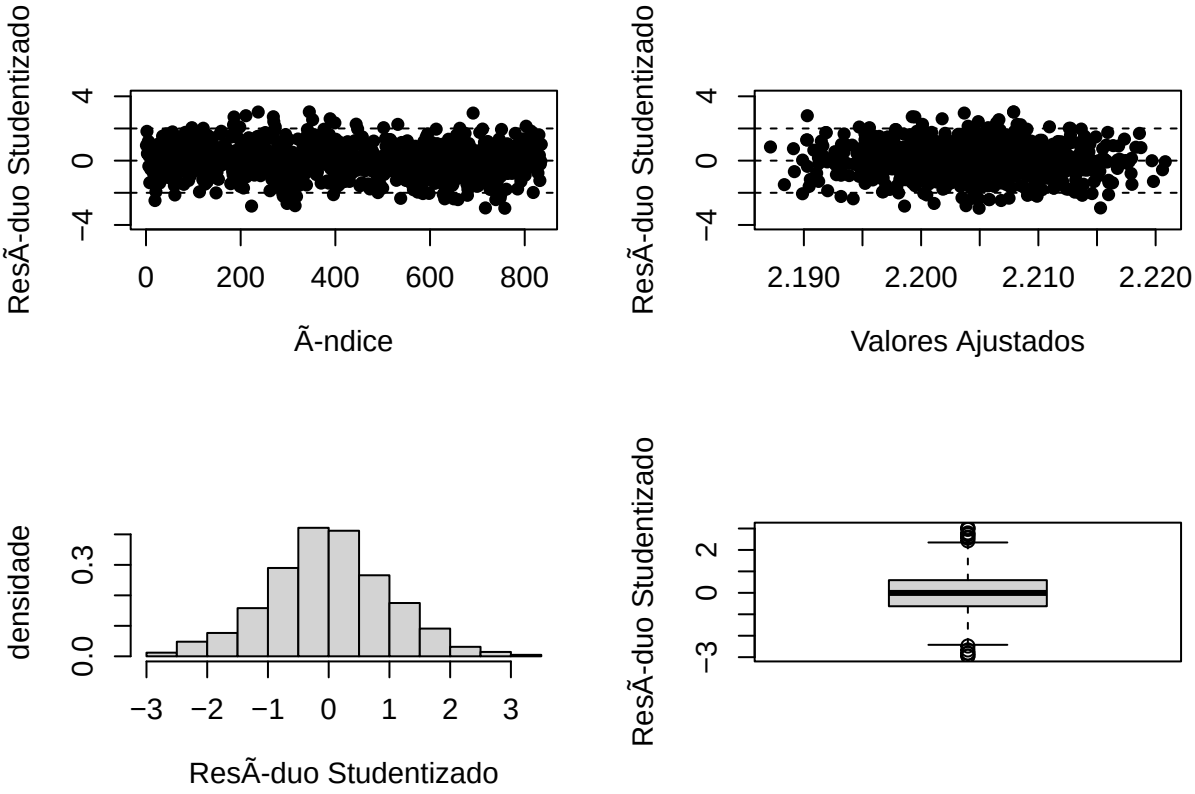
```
anainflu_norm(bc_model_sem_outliers)
```



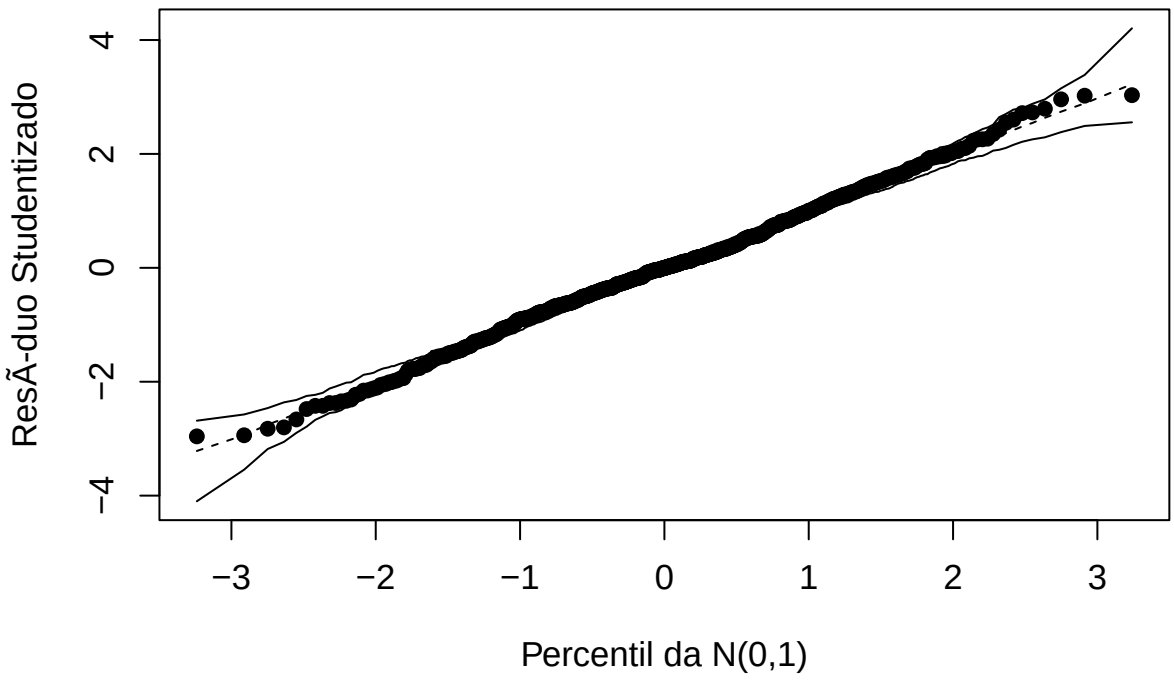





```
diag2norm(bc_model_sem_outliers)
```




```
envelnorm(bc_model_sem_outliers)
```



Note que agora, o pressuposto de normalidade dos resíduos é atendida uma vez que com base nas evidências dos gráficos e do teste de Shapiro-Wilk, podemos afirmar que o pressuposto de normalidade dos resíduos foi atendido ao nível de significância de 10% após a transformação Box-Cox e a remoção de outliers. Além de satisfazer a homocedasticidade e manter boa distribuição dos resíduos. Portanto, é um modelo estatisticamente válido para inferência.

Item 13: (0.8 pts.) Tente atingir a normalidade ao nível de significância de 10%. Comente!

Solução:

Esse item foi resolvido no item anterios, quando usamos box-cox sem outliers.

Item 14: (0.8 pts.) Elaborar a conclusão.

Este estudo teve como objetivo avaliar os fatores associados à pressão arterial de repouso (RestingBP), a partir de um conjunto de variáveis clínicas e demográficas.

O modelo final ajustado, com transformação Box-Cox e remoção de outliers, atende aos pressupostos da regressão linear ao nível de significância de 10%. Embora o poder explicativo seja limitado, o modelo permite identificar fatores com associação estatística significativa com a pressão arterial de repouso, sendo uma ferramenta válida para inferência e apoio à tomada de decisão clínica.