DataScience - R Notebook IMC

Code

Autor: Thiago Reichert

Dicas de Uso:

Atalhos para uso no R: Clicar em *Run* atalho *Ctrl+Shift+Enter* para executar cada comando.

Clicar em *Insert Chunk* ou atalho *Ctrl+Alt+I* para adicionar um linha de comando.

Para abrir o manual de uma função deve-se usar o ? seguido do nome da função, por exemplo ?dim, ?length

---

INFORMAÇÕES E COLETA DE DADOS

Dataset: Foi utilizado o dataset “500-person-gender-height-weight-bodymassindex” baixado em https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex (https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex) .

Os dados do dataset possuem as informações corporais de 500 pessoas.

Há 4 varíáveis:

- Gender (Sexo) : Male (Masculino) , Female (Feminino)
- Height (Altura) : Number (cm)
- Weight (Peso) : Number (kg)
- Index : 0 - Extremely Weak (muito magro)

  1 - Weak (magro)

  2 - Normal (normal)

  3 - Overweight (acima do peso)

  4 - Obesity (obeso)

  5 - Extreme Obesity (muito obeso)

Para realizar a coleta dos dados, a variável “imc” recebe os valores do arquivo imc.csv
OBS: Salvar o arquivo imc.csv no diretório de Worskpace do R.

Hide

```
imc <- read.csv("imc.csv")
```

EXPLORAÇÃO DOS DADOS

Estrutura do dataset:

Antes de iniciar a exploração dos dados, é interessante saber qual tipo de dados do dataset.

Hide

```
str(imc)
```

```
'data.frame':   500 obs. of  4 variables:
 $ gender: Factor w/ 2 levels "Female","Male": 2 2 1 1 2 2 2 2 2 1 ...
 $ height: int  174 189 185 195 149 189 147 154 174 169 ...
 $ weight: int  96 87 110 104 61 104 92 111 90 103 ...
 $ index : int  4 2 4 3 3 3 5 5 3 4 ...
```

Com o comando str, é possível identificar que o dataset possui 500 registros e 4 variáveis, sendo uma varíavel gender com dois valores (female e male) (dados categóricos), height do tipo int, weight do tipo int (dados intervalares) e index do tipo int (dados ordinais).

Resumo das Estatísticas:

Com a função “length()” é possível verificar o tamanho de um objeto, de uma varíavel, de uma amostra de valores de uma varíavel etc.
Com a função “dim()” é possível verificar a dimensão de um objeto ou um dataset.

Quantidade de variáveis/colunas do dataset:

Hide

```
length(imc)
```

```
[1] 4
```

Quantidade de dados (utilizando a coluna “gender” por exemplo):

Hide

```
dim(imc)
```

```
[1] 500   4
```

Quantidade de homens e mulheres respectivamente:

Hide

```
length(imc$gender[imc$gender=='Male'])
```

```
[1] 245
```

Hide

```
length(imc$gender[imc$gender=='Female'])
```

```
[1] 255
```

Para visualizar os primeiros elementos do dataset pode ser usado a função “head()”:

Hide

```
head(imc)
```

	gender <fctr>	height <int>	weight <int>	index <int>
1	Male	174	96	4
2	Male	189	87	2
3	Female	185	110	4
4	Female	195	104	3
5	Male	149	61	3
6	Male	189	104	3

6 rows

Esta é maneira de identificar como pode ser iniciado a exploração dos dados.

Valores Máximos

Com a função “max()” é possível identificar os valores máximos dos pesos e das alturas:

Valor máximo da altura de todas as pessoas:

Hide

```
max(imc$height)
```

```
[1] 199
```

Valor máximo do peso de todas as pessoas:

Hide

```
max(imc$weight)
```

```
[1] 160
```

Com a função “min()” é possível identificar os valores mínimos dos pesos e das alturas:

Valor mínimo da altura de todas as pessoas:

Hide

```
min(imc$height)
```

```
[1] 140
```

Valor mínimo do peso de todas as pessoas:

Hide

```
min(imc$weight)
```

```
[1] 50
```

Saber os valores máximos e mínimos são interessantes para identificar o intervalo de cada uma das variáveis, neste caso, peso e altura. Agora sabe-se que a altura possui um intervalo de 140cm até 199cm e o peso de 50kg a 160kg. O índice não há necessidade, pois sabe-se que o intervalo é entre 1 e 5 conforme especificação do dataset.

Médias

Com a função “mean()” é possível identificar a média dos pesos, alturas e índices:

Média das alturas de todas as pessoas:

Hide

```
mean(imc$height)
```

```
[1] 169.944
```

Média de peso de todas as pessoas:

Hide

```
mean(imc$weight)
```

```
[1] 106
```

Média do índice de obesidade:

Hide

```
mean(imc$index)
```

```
[1] 3.748
```

Apenas utilizando a média é possível identificar alguns dados interessantes para a análise. O conjunto de dados apresenta uma população com média de altura 169 cm e média de peso 106 kg, acima do peso ideal e próxima da obesidade, conforme a média de índices de 3,748.

Medianas

Com a função “median()” é possível identificar a mediana dos pesos, alturas e índices:

Mediana das alturas de todas as pessoas:

Hide

```
median(imc$height)
```

```
[1] 170.5
```

Mediana de peso de todas as pessoas:

Hide

```
median(imc$weight)
```

```
[1] 106
```

Mediana do índice de obesidade:

Hide

```
median(imc$index)
```

```
[1] 4
```

As médias e medianas possuem valores muitos próximos. Com a mediana é possível identificar que a maioria do conjunto está com o índice corporal de obesidade (4).

_____ Quartis

Com a função “quantile()” é possível identificar os quartis dos pesos, alturas e índices:

Quartis das alturas de todas as pessoas:

Hide

```
quantile(imc$height)
```

```
   0%   25%   50%   75%  100% 
140.0 156.0 170.5 184.0 199.0 
```

Quartis de peso de todas as pessoas:

Hide

```
quantile(imc$weight)
```

```
  0%  25%  50%  75% 100% 
  50   80  106  136  160 
```

Quartis do índice de obesidade:

Hide

```
quantile(imc$index)
```

```
  0%  25%  50%  75% 100% 
   0    3    4    5    5 
```

Com os quartis mais uma vez destaca-se os valores dos índices, onde já é possível identificar que pelo menos 25% da população possui o índice de obesidade 5, e mais da metade da população possui um índice de 4 ou mais.

_____ Desvio Padrão

Com a função “sd()” é possível identificar o desvio padrâo dos pesos, alturas e índices:

Desvio padrão das alturas de todas as pessoas:

Hide

```
sd(imc$height)
```

```
[1] 16.37526
```

Desvio padrão de peso de todas as pessoas:

Hide

```
sd(imc$weight)
```

```
[1] 32.38261
```

Desvio padrão do índice de obesidade:

Hide

```
sd(imc$index)
```

```
[1] 1.355053
```

Com o devio padrão, é possível identificar para análises que o conjunto de dados possui uma maior diferença de valores em relação a média na varíavel do peso.

Variância

Com a função “var()” é possível identificar a variância dos pesos, alturas e índices:

Variância das alturas de todas as pessoas:

Hide

```
var(imc$height)
```

```
[1] 268.1492
```

Variância dos pesos de todas as pessoas:

Hide

```
var(imc$weight)
```

```
[1] 1048.633
```

Variância do índice de obesidade:

Hide

```
var(imc$index)
```

```
[1] 1.836168
```

É possível visualizar um resumo dos principais dados estatísticos do dataset com o comando “summary”.

Hide

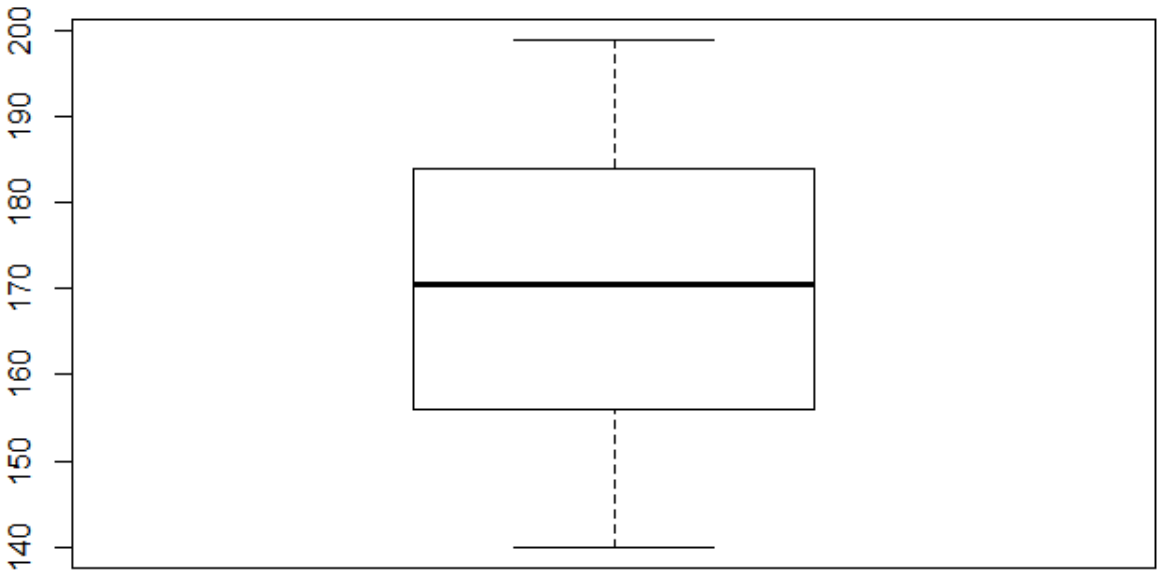
```
summary(imc)
```

```
   gender        height          weight        index      
 Female:255   Min.   :140.0   Min.   : 50   Min.   :0.000  
 Male  :245   1st Qu.:156.0   1st Qu.: 80   1st Qu.:3.000  
              Median :170.5   Median :106   Median :4.000  
              Mean   :169.9   Mean   :106   Mean   :3.748  
              3rd Qu.:184.0   3rd Qu.:136   3rd Qu.:5.000  
              Max.   :199.0   Max.   :160   Max.   :5.000
```

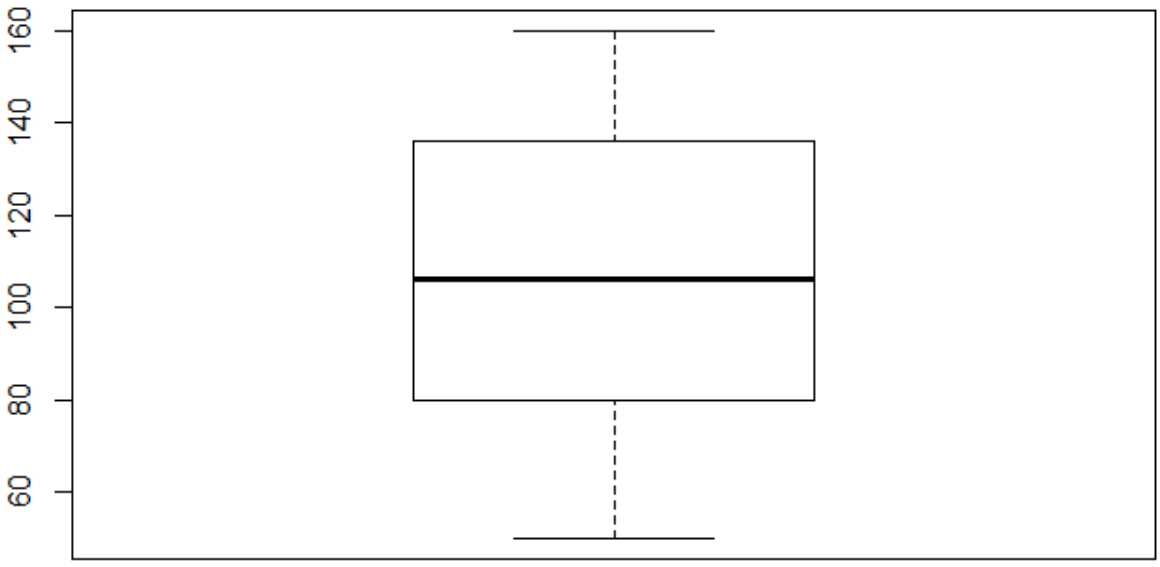
Com o summary, é possível identificar que o dataset possui registros de 255 mulheres e 245 homens. Também apresenta os valores mínimos, máximos, quartis, médias e medianas dos pesos, alturas e índices.

Também é possível visualizar graficamente as informações da mediana, primeiro e terceiro quartil, maior e menor elemento da distribuição com a função “boxplot()”

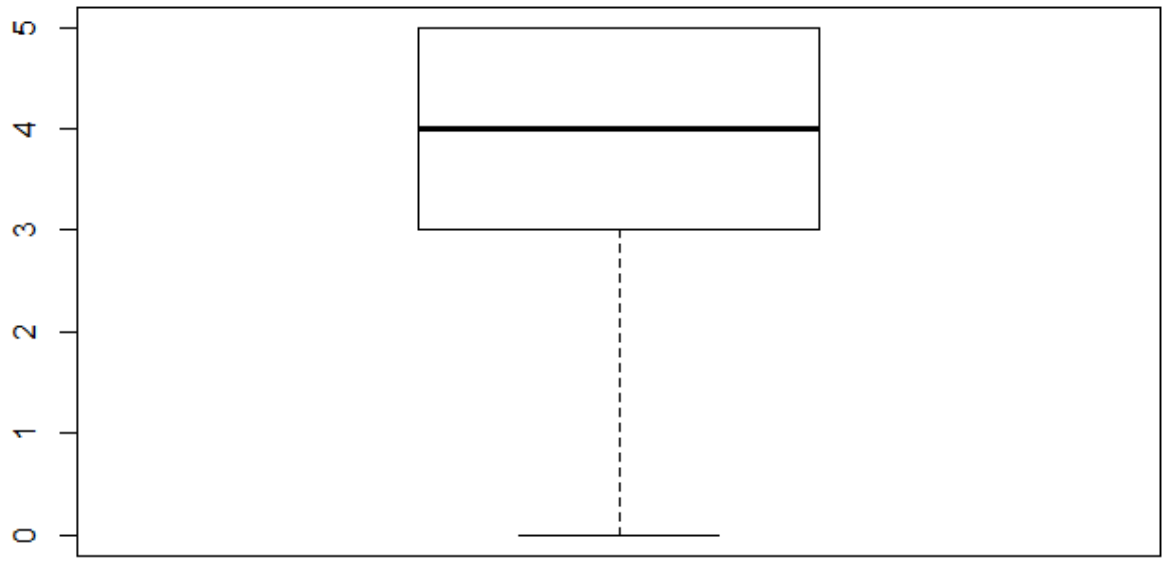
Boxplot das alturas:



Boxplot das alturas:



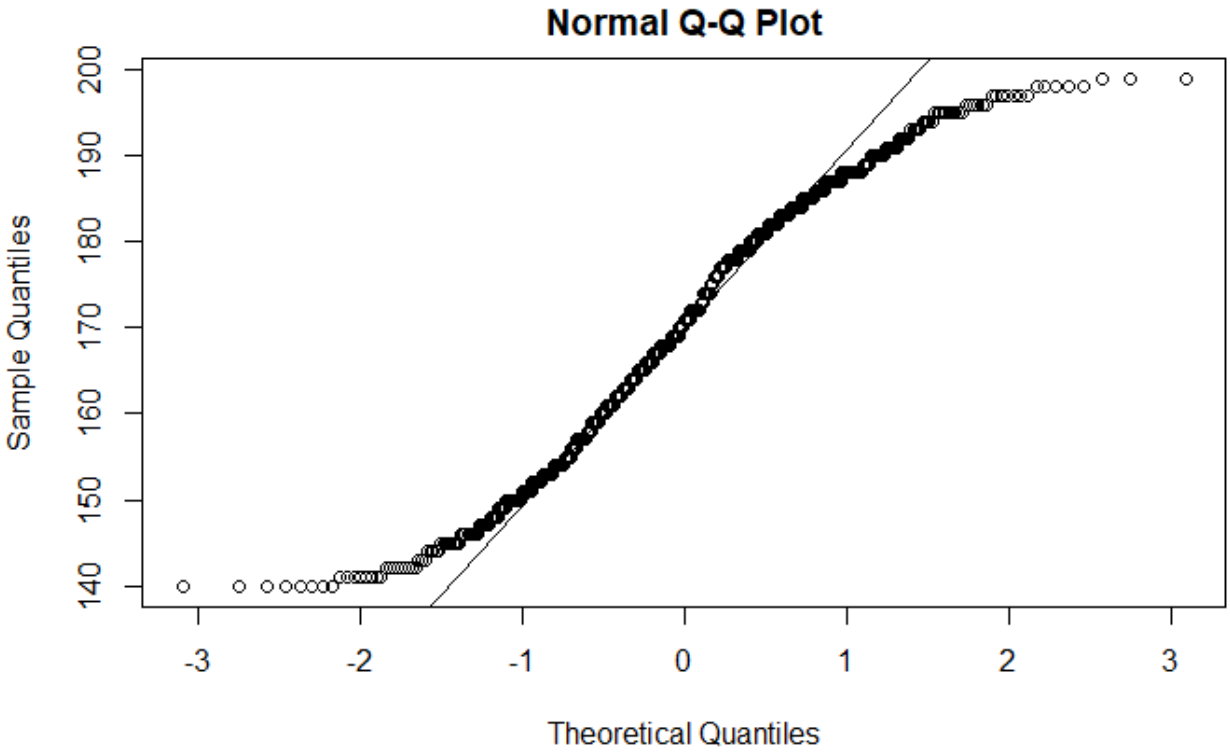
Boxplot dos índices:



DISTRIBUIÇÃO DOS DADOS

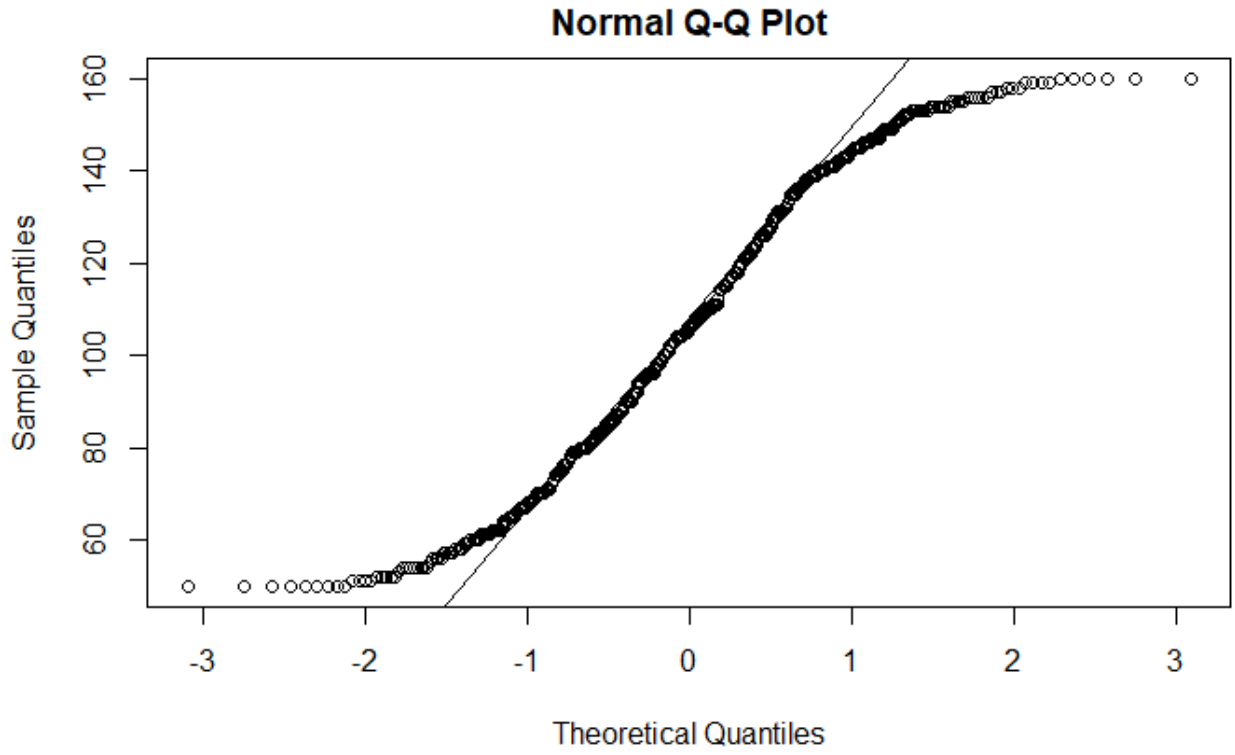
Com as funções “qqnorm()” e “qqline()” é possível verificar se a distribuição dos dados é normal.

Gráfico de normalidade e linha de melhor ajuste das alturas:



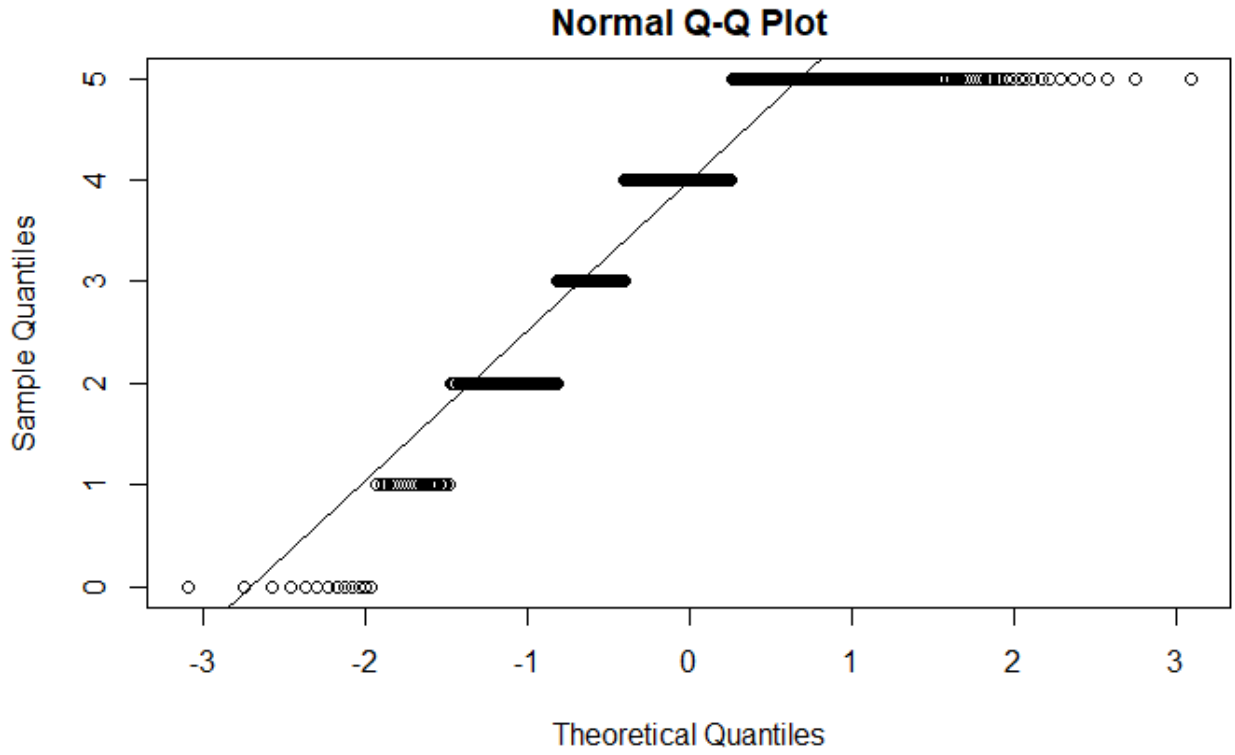
É possível verificar com o gráfico que os dados não estão bem distribuídos.

Gráfico de normalidade e linha de melhor ajuste dos pesos:



É possível verificar com o gráfico que os dados não estão bem distribuídos.

Gráfico de normalidade e linha de melhor ajuste dos índices:



Teste de Shapiro

Com o teste de Shapiro-Wilk, também é possível verificar se há uma distribuição normal dos dados. É utilizada a função “shapiro.test()” para criar um Teste de Shapiro.

Teste de Shapiro para alturas:

Hide

```
shapiro.test(imc$height)
```

```
    Shapiro-Wilk normality test

data:  imc$height
W = 0.96065, p-value = 2.665e-10
```

Teste de Shapiro para pesos:

Hide

```
shapiro.test(imc$weight)
```

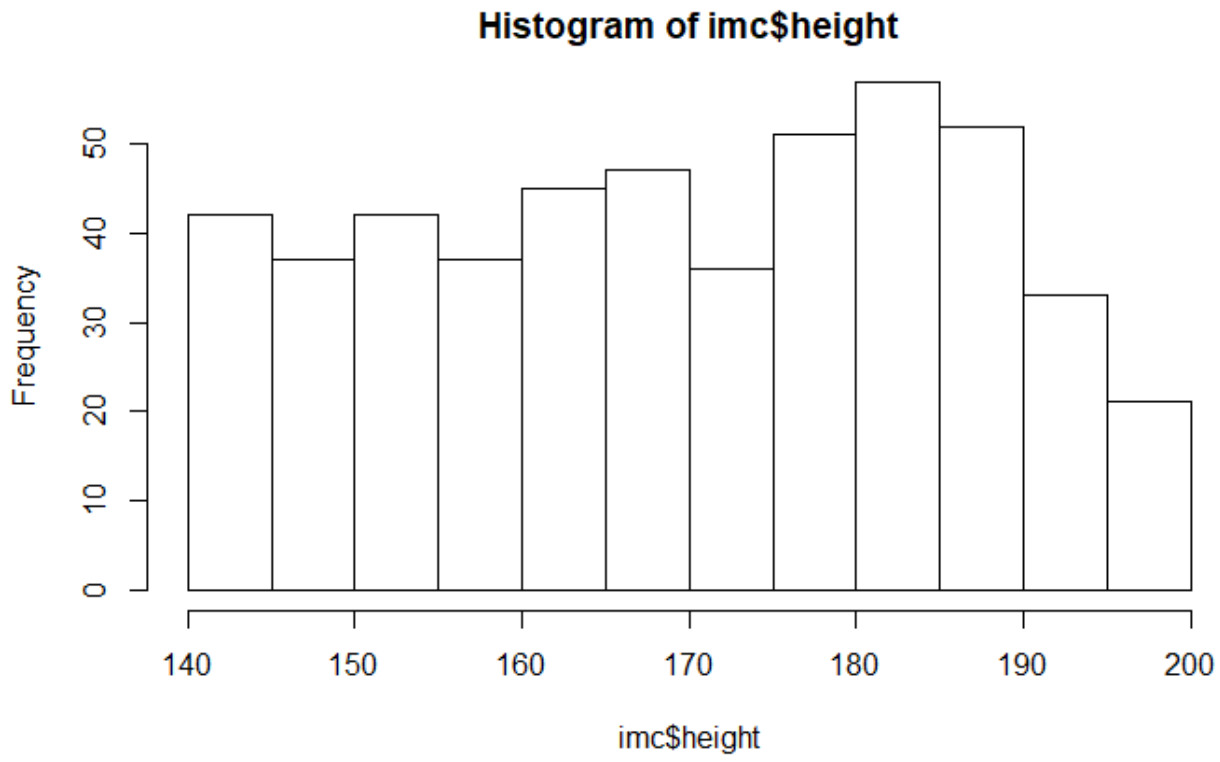
```
    Shapiro-Wilk normality test

data:  imc$weight
W = 0.95298, p-value = 1.596e-11
```

Ambos dão um valor muito baixo, abaixo de 0,5 (padrão), comprovando que os dados não estão bem distribuídos.

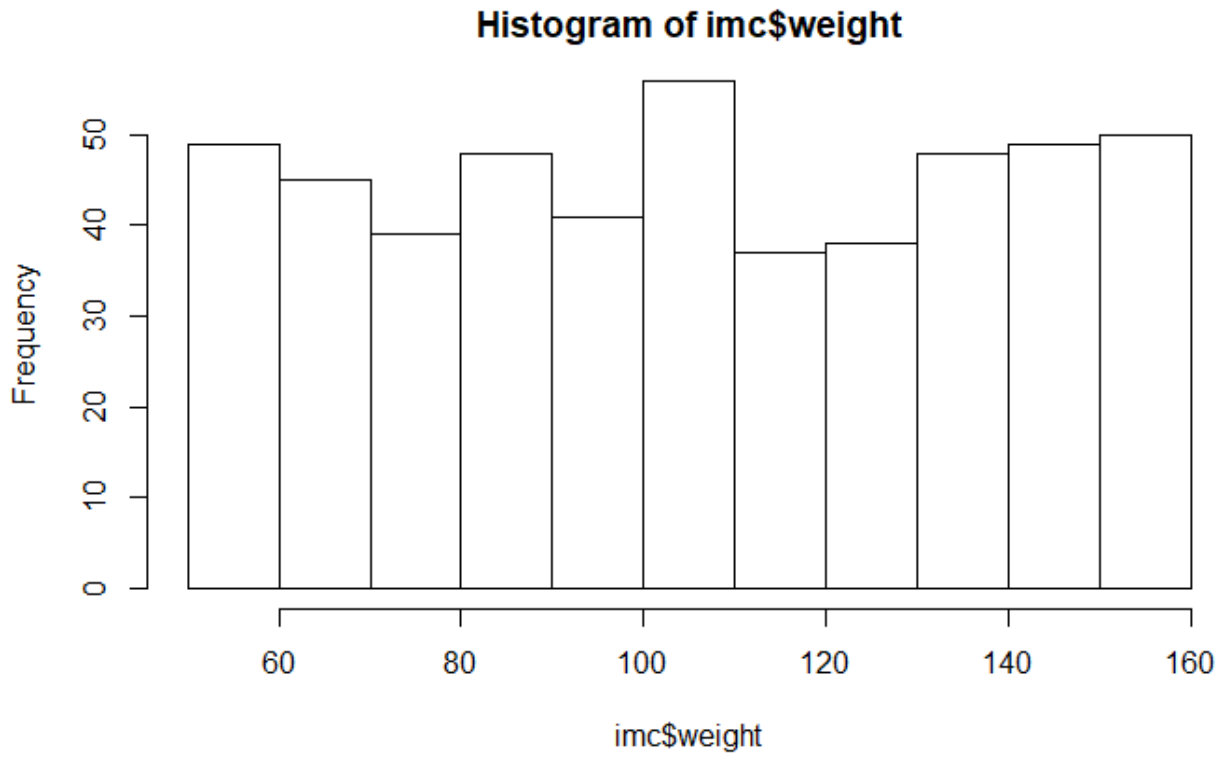
Histograma

O Histograma é interessante para visualizar rapidamente a frequência de dados em determinadas classes. Com a função “hist()” é possível criar um histograma.

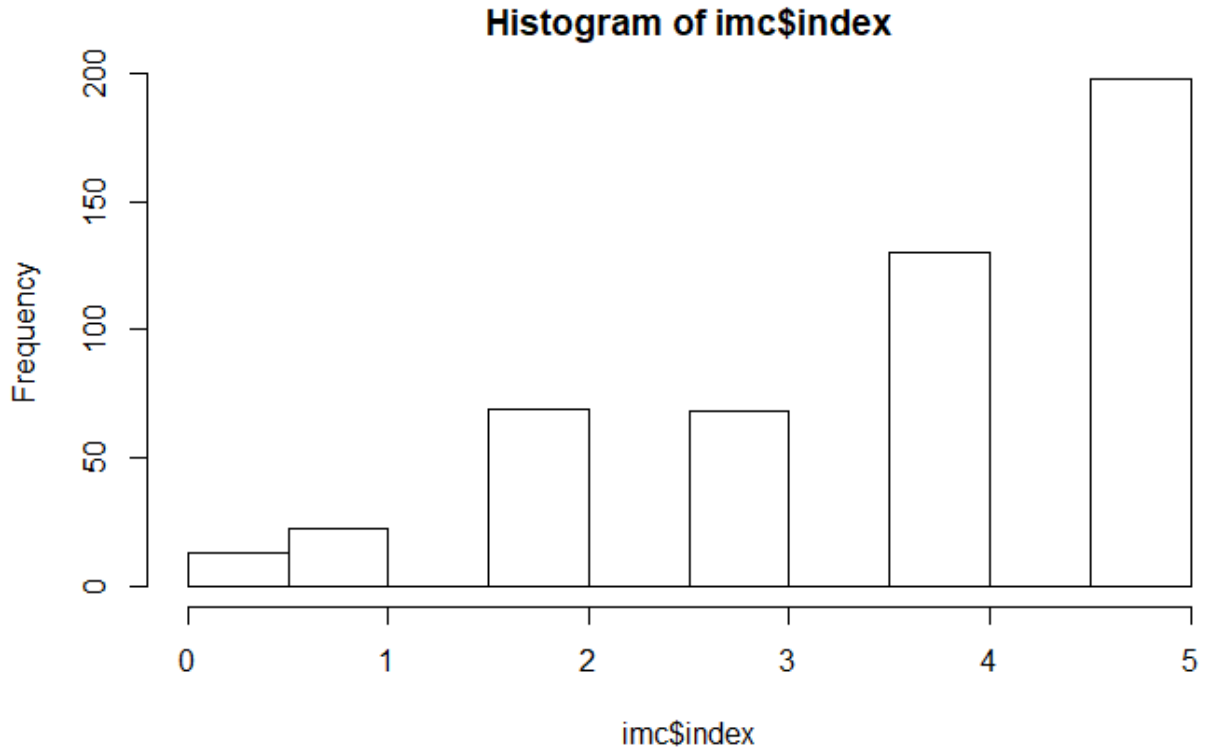
Histograma das alturas:



Histograma dos pesos:



Histograma dos índices de imc:



Com os histogramas é possível identificar uma grande parcela da população com a altura no intervalo de 175cm e 190cm. Para o histograma de pesos, destaca-se a alta frequeência de pessoas com peso acima de 100kg, sendo um alto número acima de 130kg, o que ajuda a justificar a falta de normalidade dos dados. Para o índice, destaca-se a grande frequência de pessoas com índice de IMC = 5.

Intervalo de Confiança

Com a função “t.test()” é possível descobrir o intervalo de confiança de um conjunto de dados.

Hide

```
t.test(imc$height)
```

```
    One Sample t-test

data:  imc$height
t = 232.06, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 168.5052 171.3828
sample estimates:
mean of x
  169.944
```

Intervalo de confiança da variável altura que possui média 169,944 é de 168,50 a 171,38.

Hide

```
t.test(imc$weight)
```

```
    One Sample t-test

data:  imc$weight
t = 73.195, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 103.1547 108.8453
sample estimates:
mean of x
      106
```

Intervalo de confiança da variável peso que possui média 106 é de 103,15 a 108,84.

Hide

```
t.test(imc$index)
```

```
    One Sample t-test

data:  imc$index
t = 61.848, df = 499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.628938 3.867062
sample estimates:
mean of x
    3.748
```

Intervalo de confiança da variável index que possui média 3,75 é de 3,63 a 3,87.

Como foi possível visualizar na saída da função, o intervalo de confiança padrão da função t.test é de 95%.

Exercício: Fazer o intervalo de confiança para cada uma das variáveis.

Intervalo de confiança de 90% da altura

Intervalo de confiança de 85% do peso

Intervalo de confiança de 99% do índice

Correlação

Para a calcular a correção, usa-se a função “cor()”, sendo que a função permite apenas variáveis númericas.

Correlação entre altura e peso:

Hide

```
cor(imc$height,imc$weight)
```

```
[1] 0.0004459451
```

O resultado da correlação entre altura e peso é positiva, próxima a 0. Neste caso é uma correlação positiva fraca.

Correlação entre altura e índice:

Hide

```
cor(imc$height,imc$index)
```

```
[1] -0.4222229
```

O resultado da correlação entre altura e peso é -0,42. Neste caso é uma correlação negativa fraca.

Correlação entre peso e índice:

Hide

```
cor(imc$weight,imc$index)
```

```
[1] 0.8045691
```

Como é esperado, a correlação entre peso e índice é alta, de 0,80, sendo considerada uma correlação positiva forte. Ou seja, quanto maior o peso de uma pessoa, maior a tendência de ela possuir uma índice de obesidade alto.

Regressão linear Simples

Para a regressão linear simples deve ser utilizada a função “lm()” (linear model).

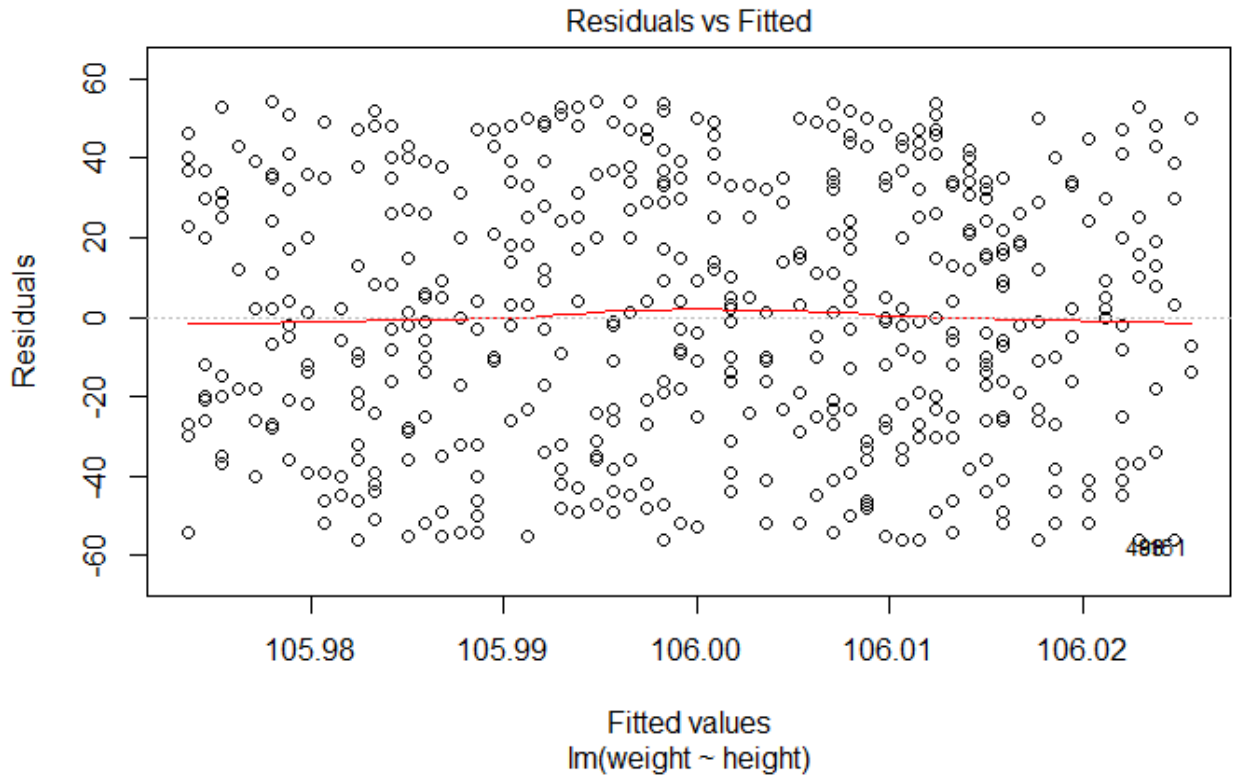
Hide

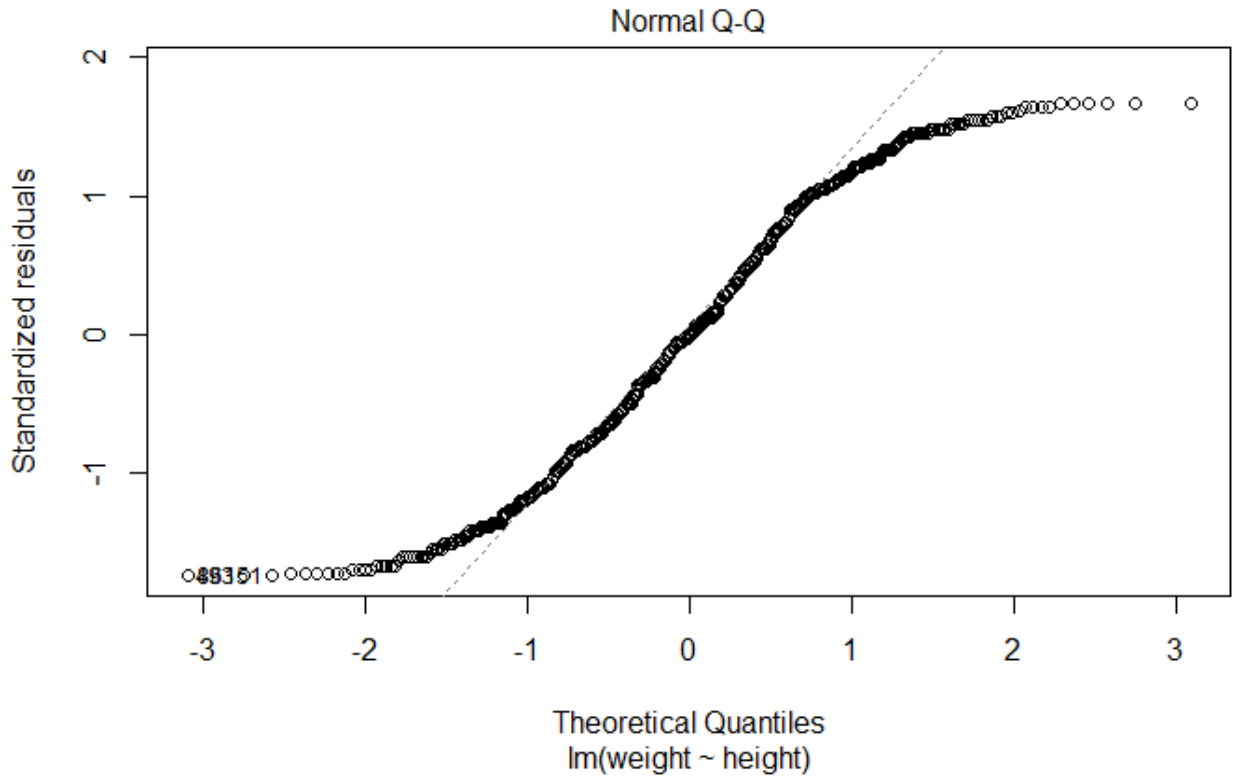
```
modelo = lm(weight ~ height, data=imc)
modelo
```

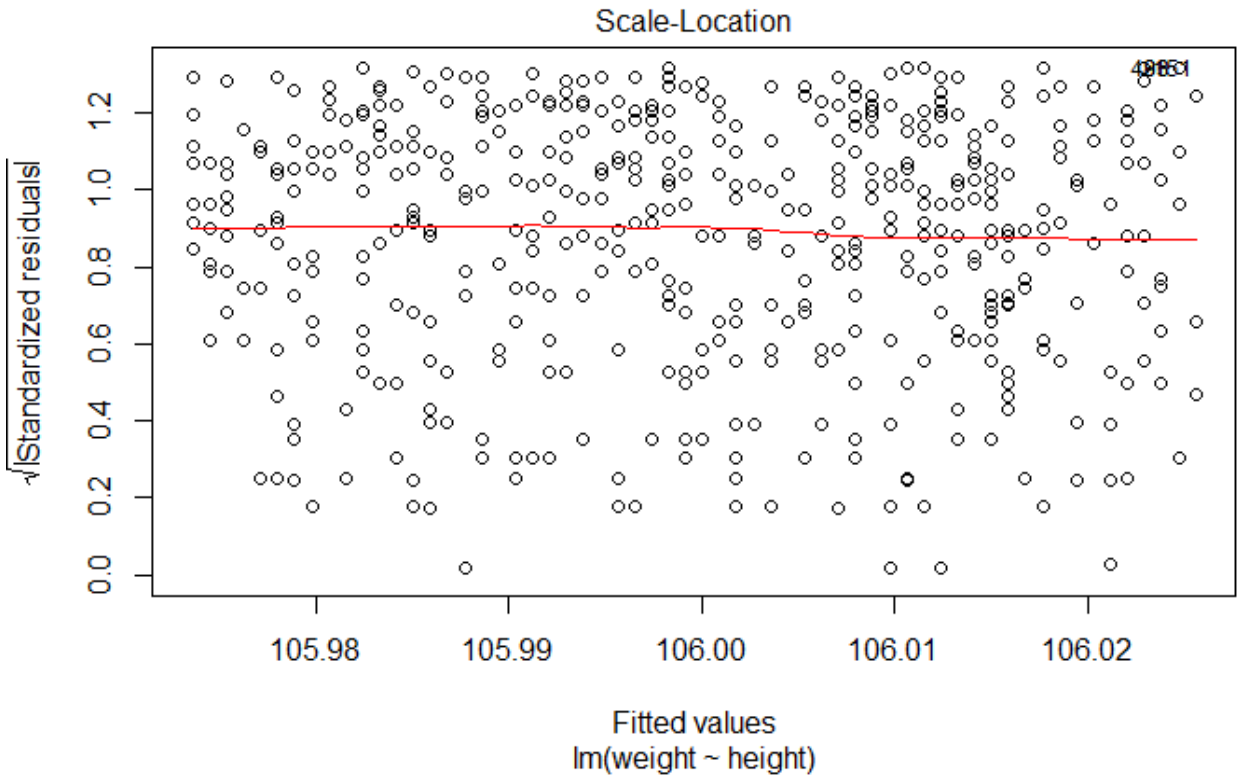
```

Call:
lm(formula = weight ~ height, data = imc)

Coefficients:
(Intercept)       height
  1.059e+02    8.819e-04
```

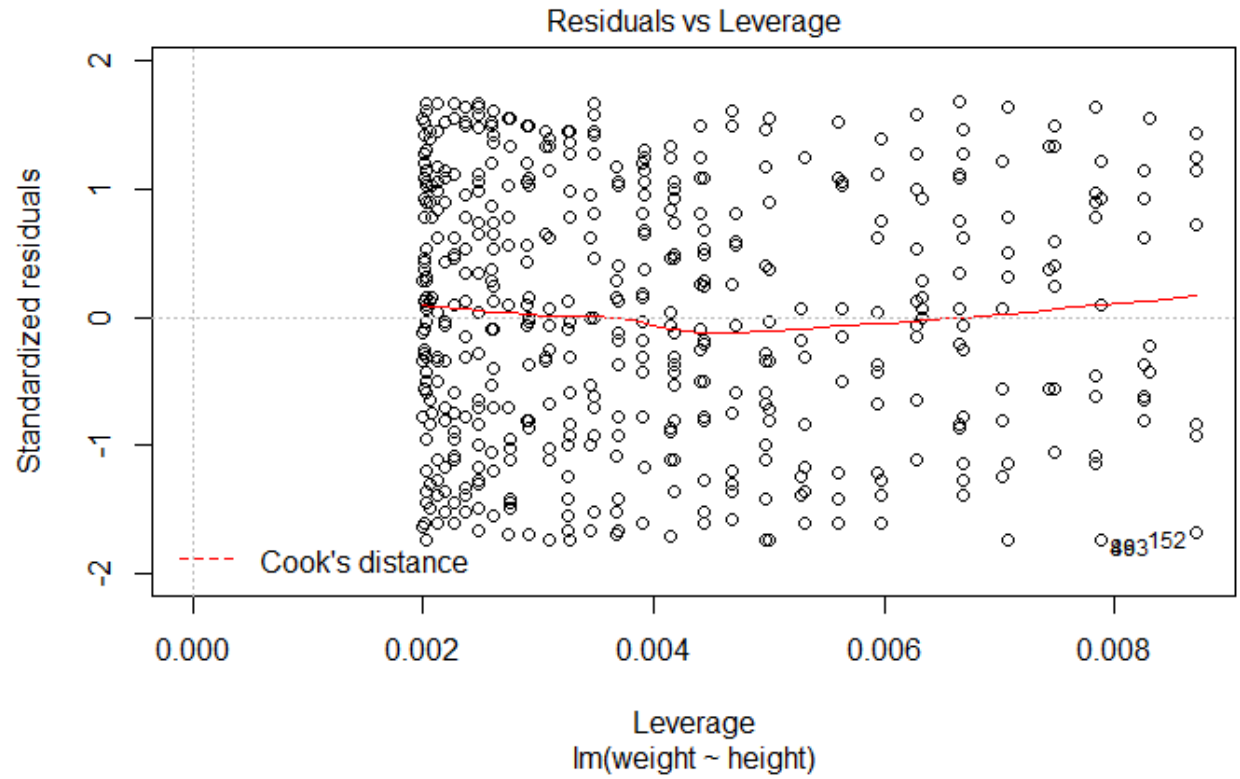
Nesta situação foi criada uma variável modelo para receber o resultado da função lm. Ao chamar a variável modelo, o resultado é a interceptação e a inclinação.

Para uma primeira análise do modelo pode ser utiliza a função “plot()” simplesmente chamando a variável “modelo”









O resultado é diversos gráficos, entre eles o gráfico de normalidade e o gráfico de residuais.

Também é possível utilizar a função “summary()” para visualizar as informações de residuais, coeficientes, inclinação, etc.

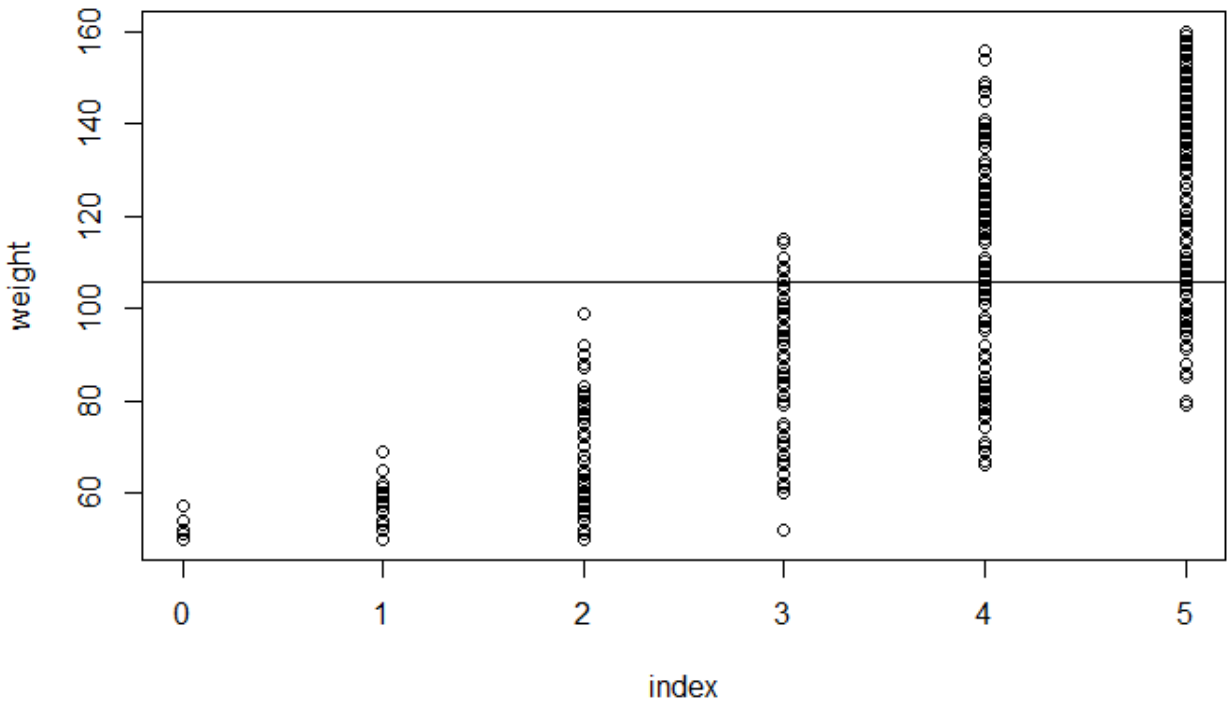
Hide

```
summary(modelo)
```

```
Call:
lm(formula = weight ~ height, data = imc)

Residuals:
    Min      1Q  Median      3Q     Max
-56.025 -26.016  -0.017  29.989  54.022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.059e+02  1.513e+01   6.996 8.51e-12 ***
height      8.819e-04  8.862e-02   0.010    0.992
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 32.42 on 498 degrees of freedom
Multiple R-squared:  1.989e-07, Adjusted R-squared:  -0.002008
F-statistic: 9.904e-05 on 1 and 498 DF,  p-value: 0.9921
```

Linha de melhor ajuste

Para definir a linha de melhor ajuste é utilizada a função “abline()”.



Previsão

Para o cálculo da previsão é utilizada a função “predict()”

Exercício: Utilizando a função predict e a variável “modelo”, faça a previsão de quanto seria o peso de uma pessoa de 180cm neste dataset:

Outliers

Para identificação de outliers pode ser utiliza novamente o “boxplot()”.

Hide

```
boxplot.stats(imc$height)$out
```

```
integer(0)
```

Hide

```
boxplot.stats(imc$weight)$out
```

```
integer(0)
```

Hide

```
boxplot.stats(imc$index)$out
```

```
integer(0)
```

Ou seja, este dataset não possui outliers para nenhuma das variáveis.