Bioestatística - Regressão Linear e Regressão Logística

2022-07-10

Regressão Linear

Carregando os dados

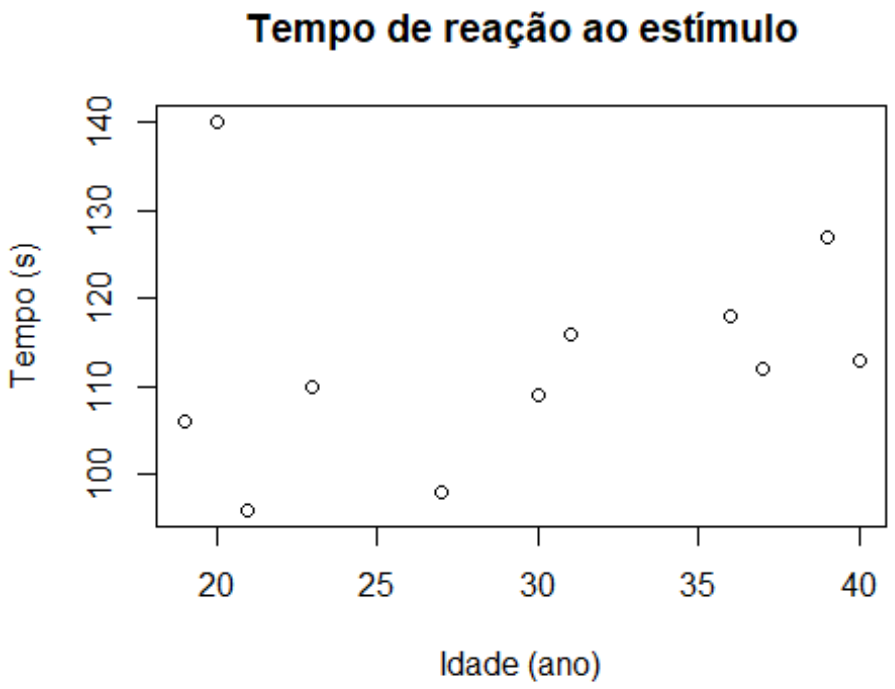
Lembra que podemos colocar os nosso dados manualmente da seguinte forma:

```
idade = c(21,19,27,23,20,31,30,37,36,40,39)
tempo = c(96,106,98,110,140,116,109,112,118,113,127)
```

Gráfico de dispersão

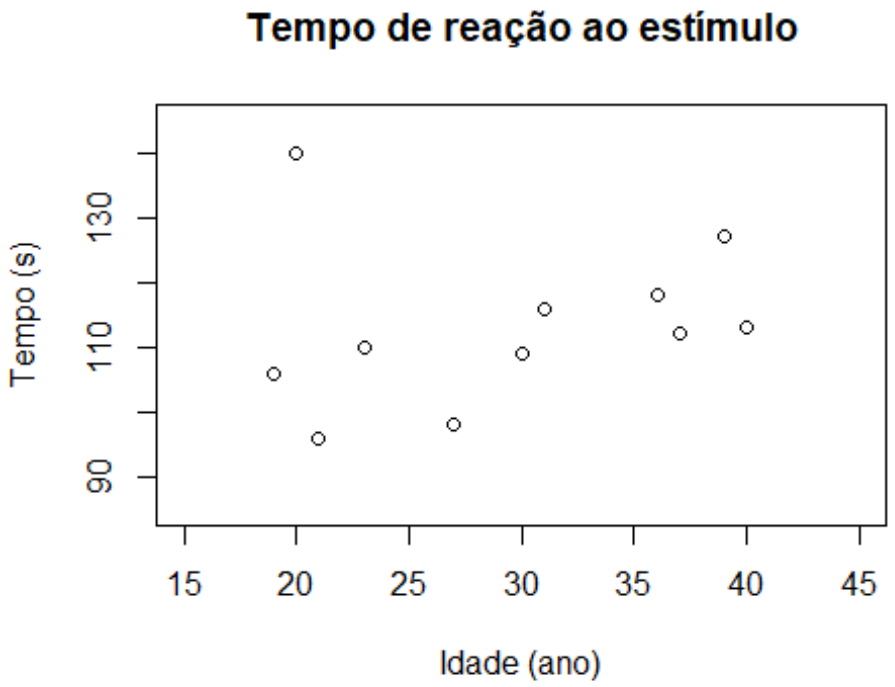
O primeiro passo que temos que fazer é gerar o gráfico de dispersão, com ele podemos já ter uma ideia inicial de como será nossa análise utilizando regressão linear. E se é possível aplicar uma regressão linear.

```
plot(idade, tempo,
     main = "Tempo de reação ao estímulo",
     xlab="Idade (ano)",
     ylab = "Tempo (s)")
```



Olhando inicialmente, talvez não conseguimos ver nenhuma relação entre o tempo de reação e a idade. Até conseguimos ver que parece que o tempo de reação está aumentando conforme a idade. Vamos tentar mudar o gráfico para que os eixos tenham outro limite e ver se fica mais aparente essa relação.

```
plot(idade, tempo,
     ylim = c(85, 145),
     xlim = c(15, 45),
     main = "Tempo de reação ao estímulo",
     xlab = "Idade (ano)",
     ylab = "Tempo (s)")
```

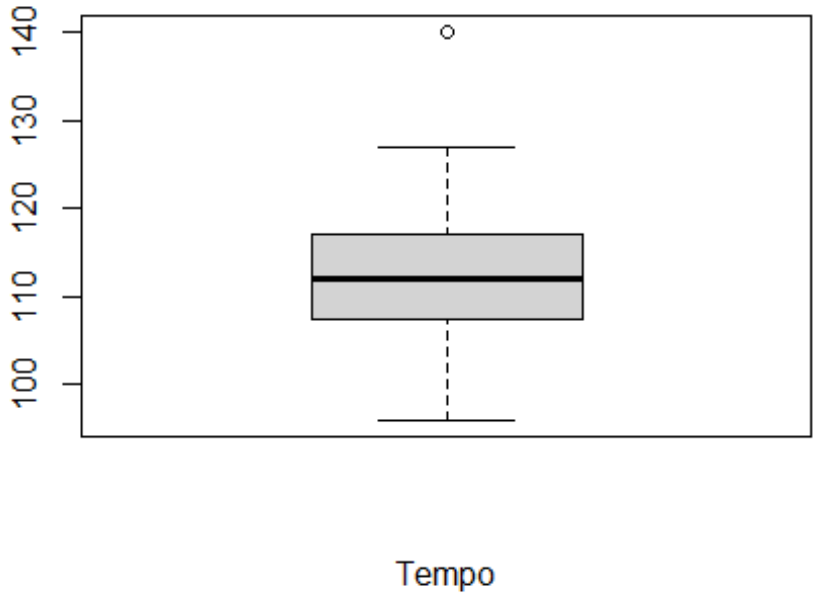


Agora parece ter ficado um pouco mais claro que realmente parece aumentar conforme a idade.

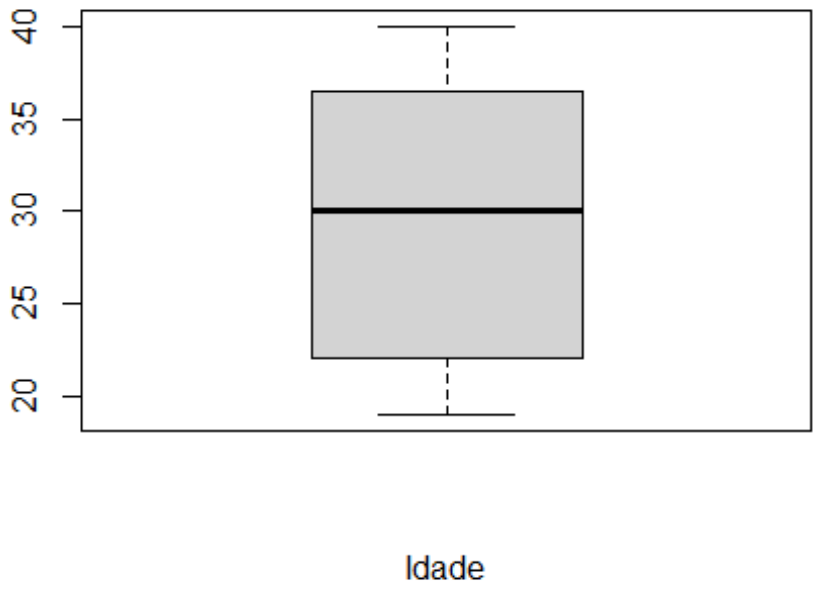
Vejam que tem um valor que parece um pouco estranho dos demais, vamos tentar verificar isso. Utilizando o boxplot.

Boxplot

```
boxplot(tempo,
        xlab = 'Tempo')
```




```
boxplot(idade,
        xlab = 'Idade')
```



Podemos ver que temos um possível outlier no boxplot do tempo. Não necessariamente deveriamos concluir que era preciso tirar esse valor, mas perceba o seguinte, enquanto parece que os valores vão aumentando conforme a idade, esse é o único que parece estar muito acima dos outros valores com idade próxima. Talvez realmente devemos tirar essa observação.

Para isso, podemos fazer de duas formas:

Removendo Outlier Método 1

Se eu sei onde estão esses valores dentro dos meus dados (idade e tempo). Nesse caso, olhando eu consigo saber que eles estão na quinta observação, logo podemos fazer da seguinte forma:

```
tempo2 = tempo[-5]
idade2 = idade[-5]
```

Perceba que temos que tirar tanto da variavel tempo quanto idade, pois estamos tirando uma observação.

Se os seus dados são em formato de tabela, pode ser feito da mesma forma:

‘tabela = tabela[-5]’ -> Voce está retirando a linha 5 da tabela

Se quiser retirar mais linhas, podem fazer da seguinte forma:

‘tabela = tabela[-c(5,7)]’ -> Voce está retirando a linha 5 e 7 da tabela

‘tempo2 = tempo[-c(5,7)]’

‘idade2 = idade[-c(5,7)]’

Agora, se você não sabe onde está a observação, porque tem muitos dados: ### Removendo Outlier Método 2

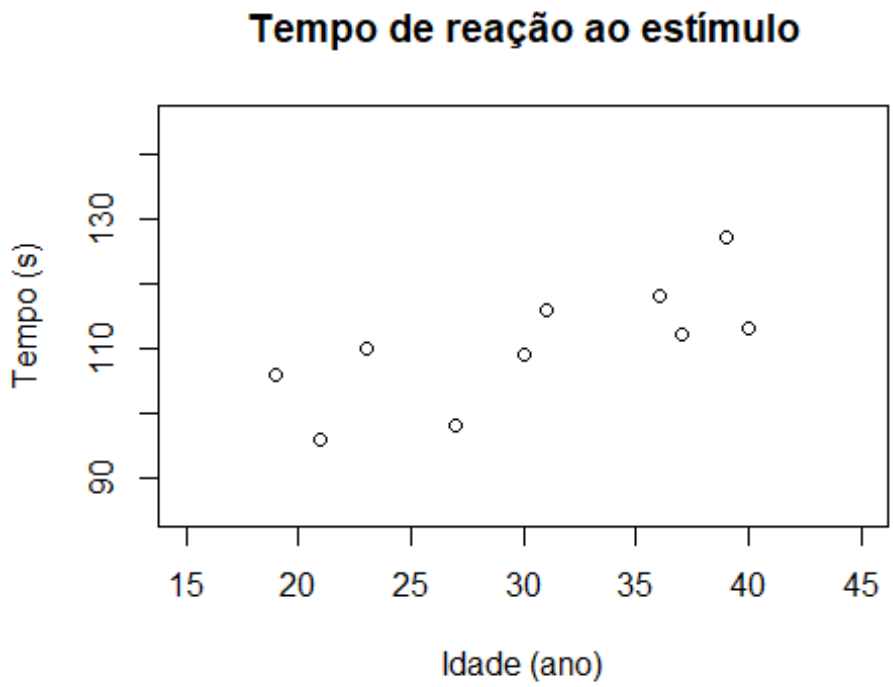
Dessa forma vocês podem retirar a observação sabendo o valor dele.

Nesse caso queremos tirar a observação que tem um tempo de 140 s e 20 anos de idade:

```
remover = (tempo == 140)&(idade==20)
tempo2 = tempo[!remover]
idade2 = idade[!remover]

#tabela = tabela[!remover]' #Caso fosse uma tabela
```

Refazendo o gráfico

```
plot(idade2, tempo2,
     ylim = c(85, 145),
     xlim = c(15, 45),
     main = "Tempo de reação ao estímulo",
     xlab = "Idade (ano)",
     ylab = "Tempo (s)")
```



Agora que já “limpamos” nossos dados, vamos aplicar a regressão.

Regressão Linear

Para criarmos o nosso modelo de regressão linear:

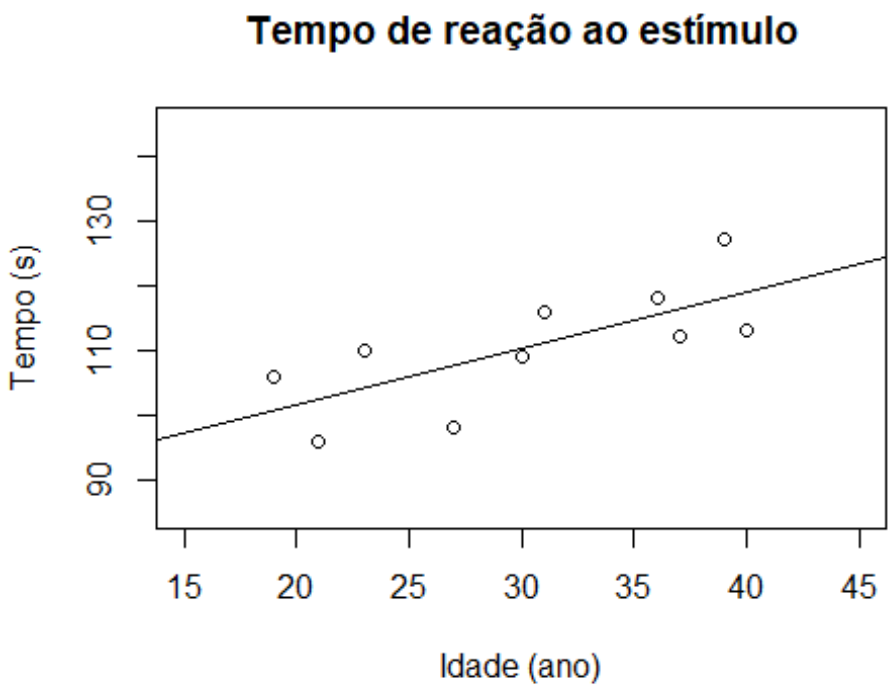
```
modelo = lm(tempo2~idade2)
#modelo = lm(avg_glucose_level ~ age, data = tabela) #caso em que os nosso da
dos estão em formato de tabela
```

Vejam que para criar o nosso modelo utilizando um dado em formato de tabela pode ser feito como apresentado acima.

Vocês tem que dizer a coluna que vocês querem fazer, percebam a ordem, e indicar a tabela na parte de 'data ='.

Pronto, agora vocês já tem o modelo de regressão linear. Vamos verificar agora como ficou:

```
plot(idade2, tempo2,
     ylim = c(85, 145),
     xlim = c(15, 45),
     main = "Tempo de reação ao estímulo",
     xlab = "Idade (ano)",
     ylab = "Tempo (s)")
abline(modelo) #Adiciona a linha do modelo de regressão linear
```



Parece bom, vamos agora analisar esse resultado.

```
summary(modelo)
```

```
##
## Call:
## lm(formula = tempo2 ~ idade2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6428 -5.4990  0.6623  5.1862  8.9675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.2661     9.1389   9.221 1.55e-05 ***
## idade2        0.8658     0.2933   2.952   0.0184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.728 on 8 degrees of freedom
## Multiple R-squared:  0.5213, Adjusted R-squared:  0.4615
## F-statistic: 8.712 on 1 and 8 DF,  p-value: 0.01838
```

Vejam que temos os coeficientes beta0 e beta1. -> Estimate -> (Intercept, idade2)

E já temos o teste de hipótese para cada um desses coeficientes. -> Pr(>|t|)

Temos o desvio padrão do resíduos. -> Residual standard error

Temos também o nosso coeficiente de Determinação. -> Multiple R-squared

O que vocês concluem disso?

A correlação de Pearson pode ser feita da seguinte forma:

```
cor(idade2,tempo2)
```

```
## [1] 0.7220209
```

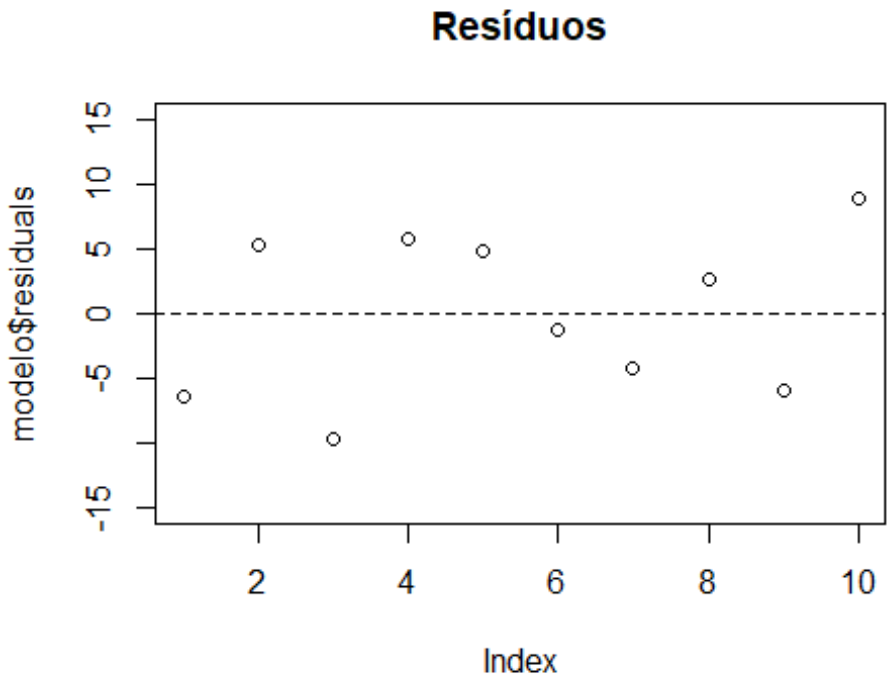
Lembrando que podemos elevar esse valor para obter o coeficiente de Determinação:

```
cor(idade2,tempo2)^2
```

```
## [1] 0.5213142
```

Como eu falei, temos que ver também os resíduos para podermos avaliar esse modelo. Para isso vamos fazer graficamente, da seguinte forma:

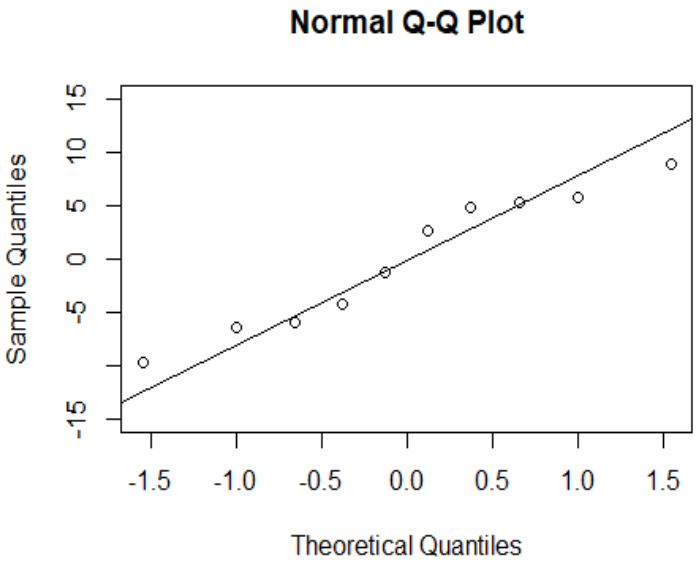
```
plot(modelo$residuals,
     main = "Resíduos",
     ylim = c(-15,15))

abline(h=0,lty=2) #adiciona uma linha em 0
```



Parece ok.

Uma outra análise que podia ser feita, para saber se os residúos seguem uma distribuição normal.

```
qqnorm(modelo$residuals,ylim = c(-15,15))
qqline(modelo$residuals)
```



Regressão Logística

Carregando os Dados:

Para a regressão logística, vamos pegar outros dados.

Vamos separar e vamos pegar apenas os que tem idade acima de 18 e gênero masculino.

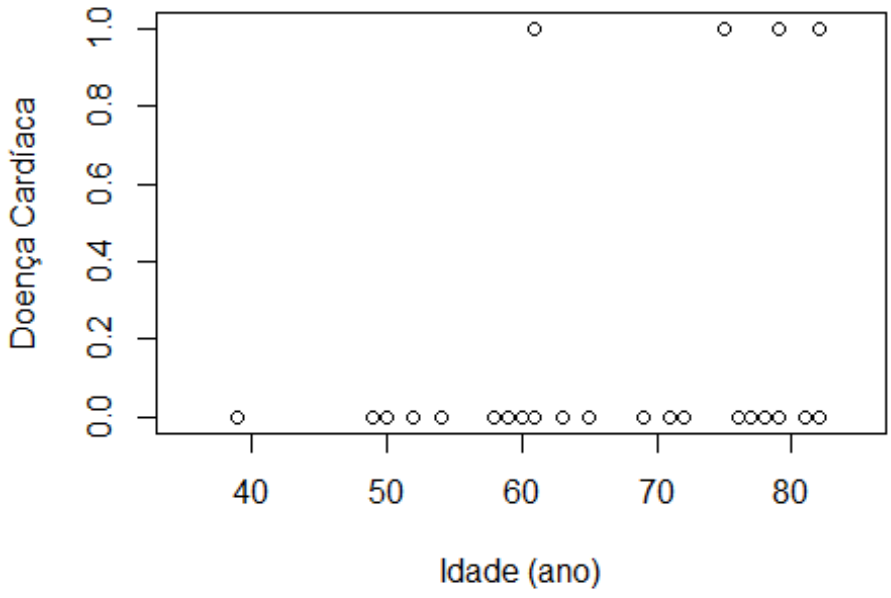
```
dados = read.table(file="D:\\Macae Disciplinas\\2022\\R\\dados_pratica.csv",
sep=",", header=TRUE)
dados = dados[!dados$gender=="Male",]
dados = dados[dados$age>18,]
```

Se eu estivesse querendo fazer uma regressão logística com a informação se são cassados ou não, teriamos que fazer a seguinte transformação:

```
#dados$ever_married[dados$ever_married == "Yes"] = 1
#dados$ever_married[dados$ever_married == "No"] = 0
```

Vamos gerar os gráficos da relação idade e se possuem alguma doença cardíaca:

```
plot(dados$age, dados$heart_disease,
     xlim = c(35, 85),
     xlab = "Idade (ano)",
     ylab = "Doença Cardíaca")
```



Vamos criar nosso modelo de regressão logística e apresentar o resultado:

```
modelo_log = glm(heart_disease ~ age, data = dados, family = binomial)
summary(modelo_log)
```

```
##
## Call:
## glm(formula = heart_disease ~ age, family = binomial, data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9490  -0.7553  -0.4032  -0.2371   2.2584
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.00175    4.46915  -1.790   0.0734 .
## age          0.09070    0.05987   1.515   0.1298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26.662  on 28  degrees of freedom
## Residual deviance: 23.437  on 27  degrees of freedom
## AIC: 27.437
##
## Number of Fisher Scoring iterations: 5
```

Vejam que temos nossos coeficientes e nossos teste de hipótese.

Vamos tentar utilizar esse modelo para termos uma noção de probabilidade de vir a ter uma doença com base na idade. Vamos tentar de 35 até 85 anos, que são valores próximos dos extremos que temos nos nossos dados.

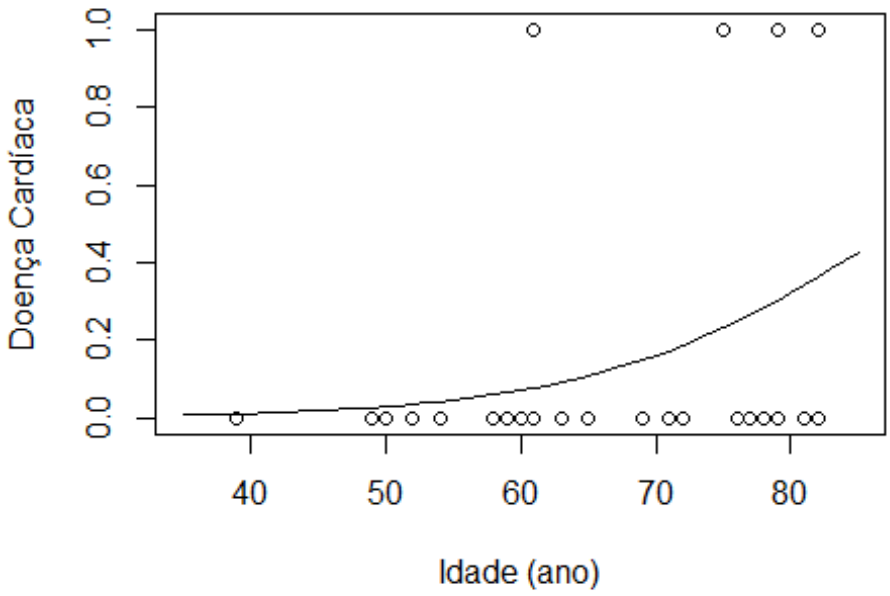
```
dados_plot = data.frame(age=seq(35, 85)) #Idade de 35 a 85 anos
dados_plot$heart_disease = predict(modelo_log, dados_plot, type = 'response')
```

Se vocês abrirem a tabela dados_plot, vocês podem ver a idade e a probabilidade correspondente.

age	heart_disease
77	0.265467060
78	0.283525063
79	0.302305885
80	0.321772048
81	0.341877468
82	0.362567573
83	0.383779644
84	0.405443374
85	0.427481654

Vamos gerar um gráfico com essas probabilidades e com os valores reais dos nossos dados.

```
plot(dados$age, dados$heart_disease,
    xlim = c(35, 85),
    xlab = "Idade (ano)",
    ylab = "Doença Cardíaca")
lines(dados_plot$age, dados_plot$heart_disease) #Resultado do nosso modelo de
regressão logística
```



Podemos ver um aumento nas chances de vir a ter uma doença conforme a idade vai aumentando.