Lista 9 [AD-UFPE-2019]

Antonio Fernandes

11 de maio de 2019

Conteúdo

Apresentação

Este documento apresenta as respostas da lista de exercícios 9 da disciplina de Análise de dados.

O link está disponível no GitHub: https://github.com/alvesat/AD_9

Questão 1

a)

Caso a correlação entre Zi e Xi é igual a 0, não acontece nenhum efeito em β, mas a conclusão acerca da variável dependente estará distorcida.

b)

Se Zi e Xi estão correlacionados, então Zi é uma variável omitida (viés de variável omitida) e deve ser incluída no modelo. Dado que a correlação entre Zi e Xi é positiva, significa que ambas as variáveis explicam uma mesma parcela da variação em Y.

c)

Do mesmo modo, caso Zi e Xi estejam negativamente correlacionadas, então Zi é uma variável omitida (viés de variável omitida) e deve ser incluída no modelo. Dado que a correlação é negativa, significa que a inclusão de Zi no modelo irá reduzir o efeito de Xi em β.

Questão 2

A tabela 9.4 apresenta o resultado de três modelos de regressão de salários de professores nos estados americanos e no distrito de colúmbia.

O modelo A contém apenas a variável *porcentagem de residentes no estado que possuem ensino superior*. O efeito dessa variável na VD é de 704.02 (com p-valor < 0.05). O R^2 do modelo é de 0.34.

Já o modelo B apresenta apenas a variável *Renda per capita*. O efeito dessa variável na VD é de 0.68 (com p-valor < 0.05). O R^2 do modelo é de 0.47.

Questão 3

O modelo C da tabela 9.4 contém as duas variáveis independentes (do modelo A e do modelo B). O efeito da variável *porcentagem de residentes no estado que possuem ensino superior* é de 24.56 mas não é significativo (p-valor > 0.05) enquanto que o efeito da variável *Renda per capita* é de 0.66 e significativo (p-valor > 0.05). Nesse caso, podemos concluir que as VIs estão correlacionadas, ocorrendo o viés de variável omitida tanto no modelo A quanto no modelo B.

Questão 4.1

a)

Modelo com as variáveis do banco *worldrecall.txt*

```r
# lendo banco worldrecall

worldrecall <- read.delim("~/Dados/Listas/AD_9/AD_9/worldrecall.txt")

  # modelo linear

Linear <- lm(prop ~ time, data = worldrecall)

  # resumo do modelo

summary(Linear)
```

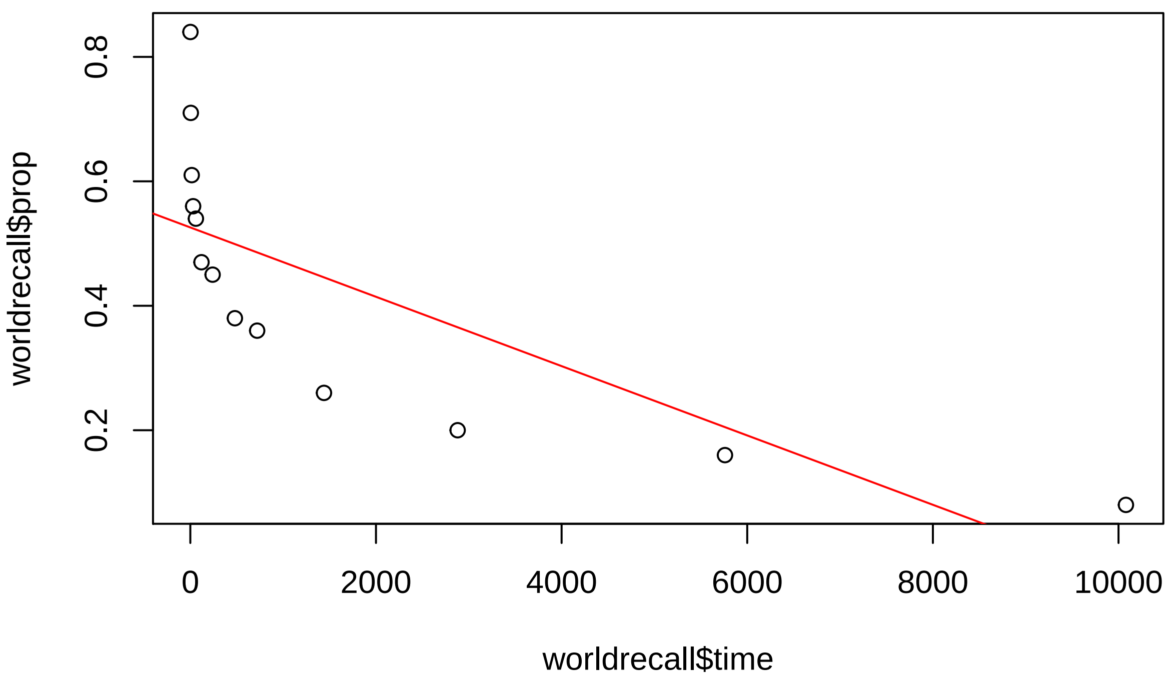
```
##
## Call:
## lm(formula = prop ~ time, data = worldrecall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18564 -0.11913 -0.04495  0.08496  0.31418
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.259e-01  4.881e-02  10.774 3.49e-07 ***
## time        -5.571e-05  1.457e-05  -3.825  0.00282 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1523 on 11 degrees of freedom
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5318
## F-statistic: 14.63 on 1 and 11 DF,  p-value: 0.002817
```

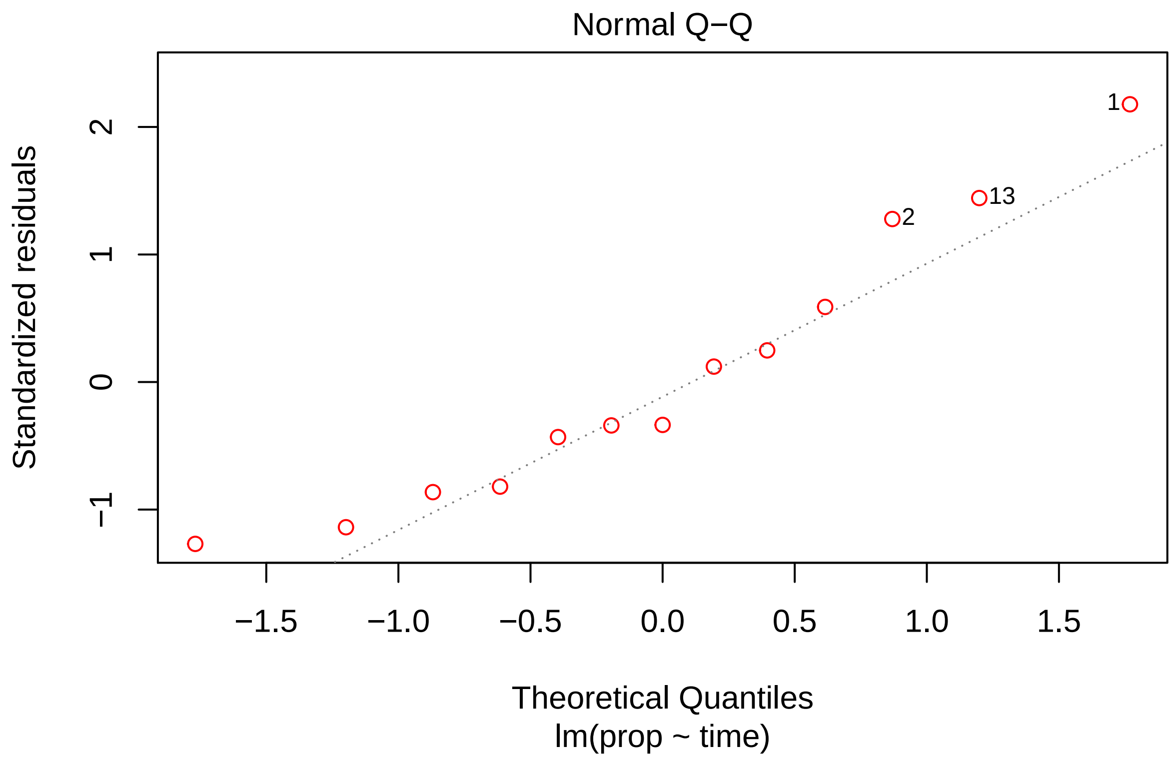
O resutado do modelo mostra um efeito de -0.00005571 da VI na VD. Entretanto o resultado apenas se mantém significativo com o P-valor < 0.1. O R^2 do modelo é de 0.53.

Entretanto, é necessário verificar se o modelo está devidamente ajustado.

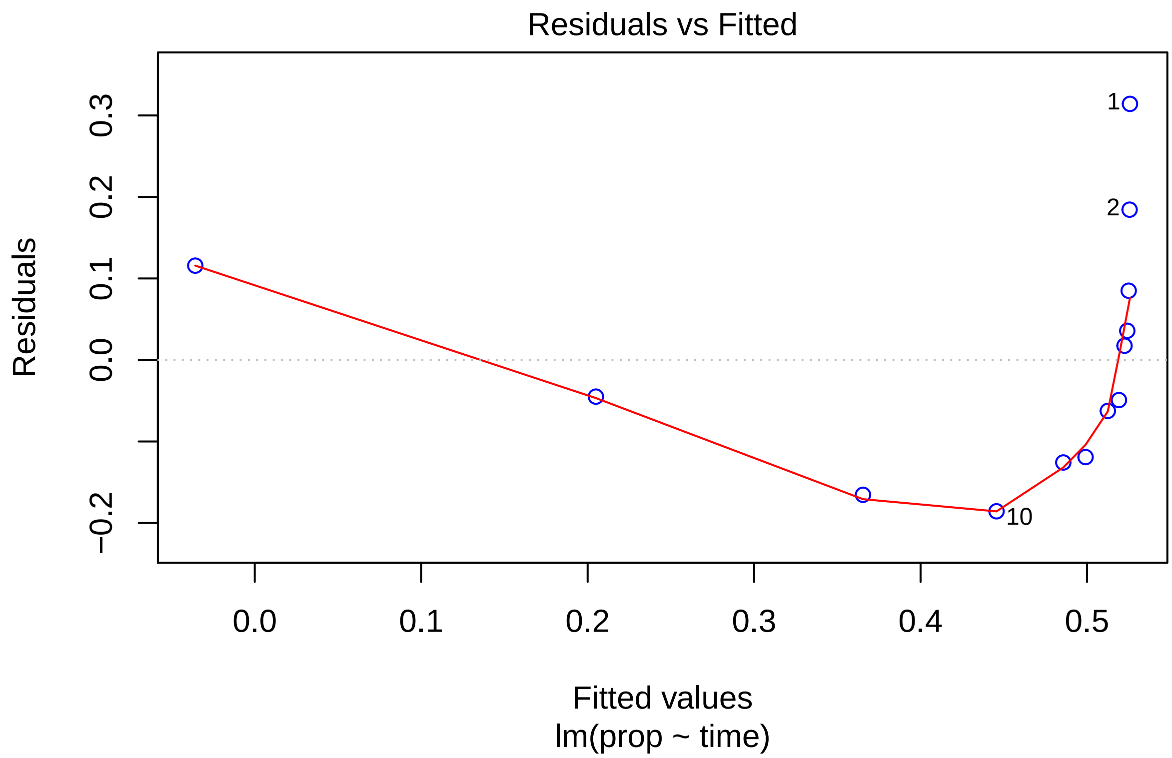
```r
  # ajuste do modelo

plot(worldrecall$time, worldrecall$prop)
abline(lm(prop ~ time, data = worldrecall), col = 'red')
```




```r
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Com base nos gráficos acima, podemos concluir que a relação entre as variáveis do modelo não é linear. A linha de regressão do gráfico mostra que as observações da variável independente não possuem um comportamento linear, do mesmo modo que o gráfico de residuals vs fitted.

Por isso, a variável independente do modelo será transformada em log

```
  # colocando vi em log

worldrecall$timeln <- log(worldrecall$time)

  # modelo linear

Linear_ln <- lm(prop ~ timeln, data = worldrecall)

  # resumo do modelo

summary(Linear_ln)
```

```
##
## Call:
## lm(formula = prop ~ timeln, data = worldrecall)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036077 -0.015330 -0.006415  0.017967  0.037799
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
## timeln      -0.079227   0.002416  -32.80 2.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02339 on 11 degrees of freedom
## Multiple R-squared:  0.9899, Adjusted R-squared:  0.989
## F-statistic:  1076 on 1 and 11 DF,  p-value: 2.525e-12
```

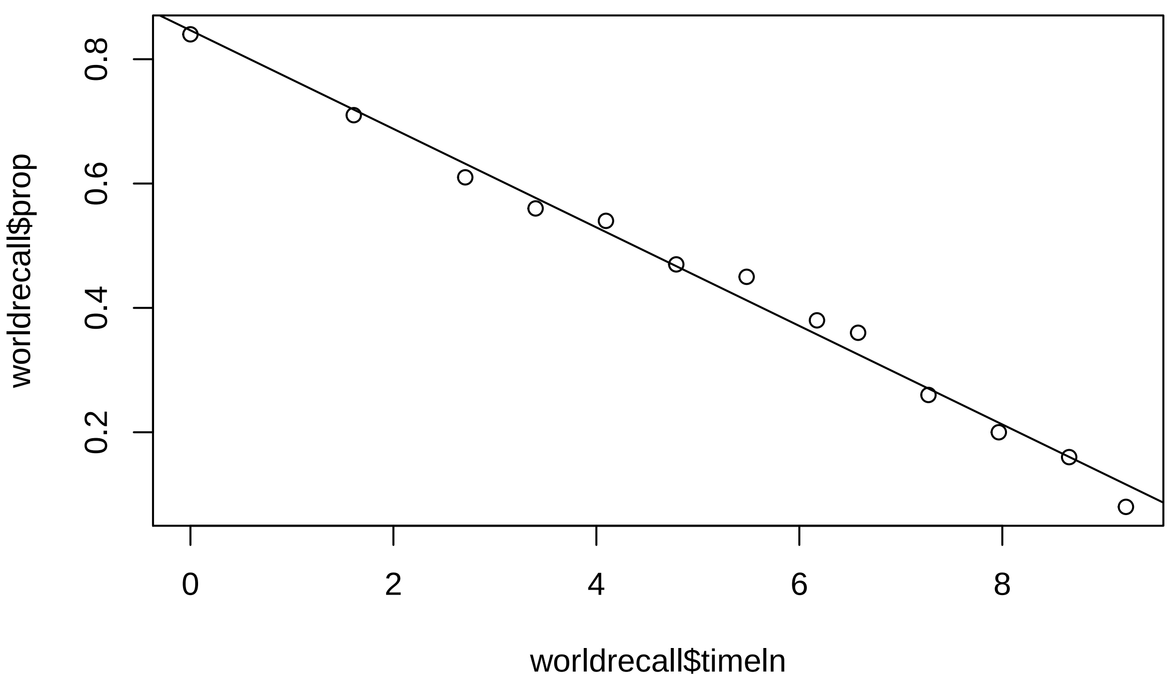
O modelo Level-log mostra que a variável independente tem um efeito negativo e estatisticamente significativo na VD (P-valor < 0.05). O R^2 do modelo é de 0.989, mostrando que a VI explica quase que completamente a variância da VD. Para interpretar o modelo, é necessário efetuar uma transformação:

```r
(-0.079227/100)
```

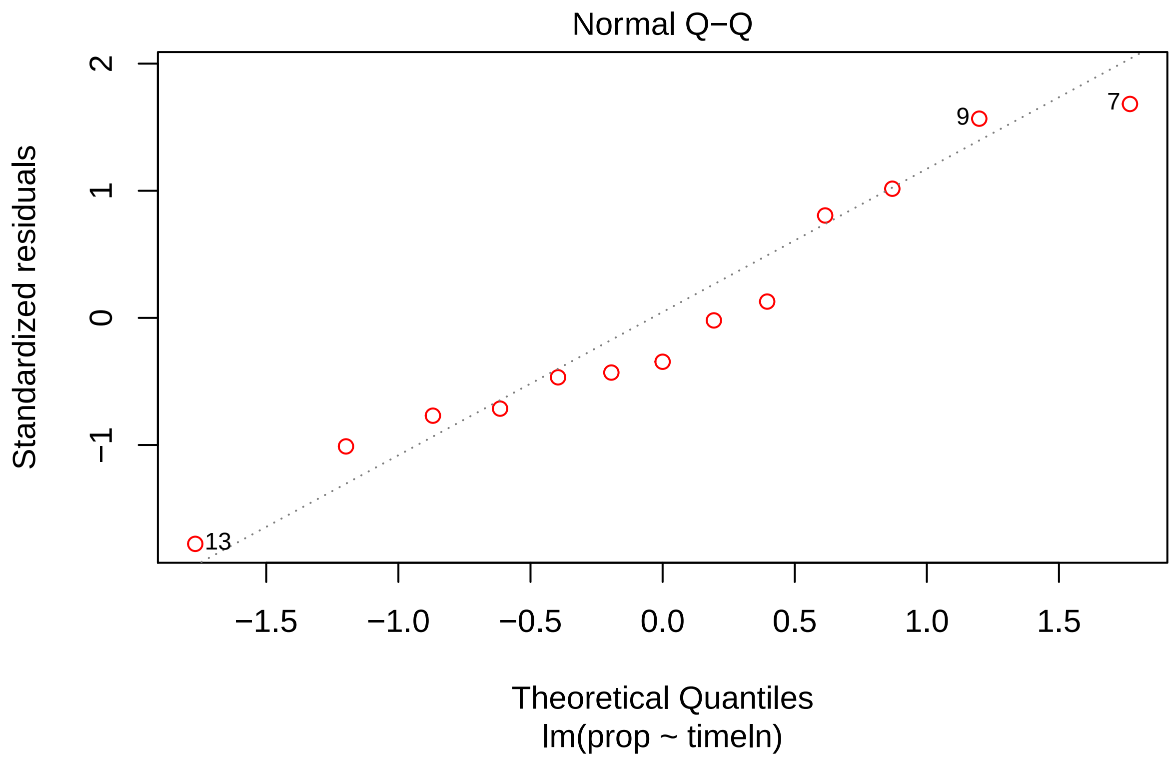
```
## [1] -0.00079227
```

Isso significa que 1% de variação em Xi implica em -0.00079227 de variação em Yi. Agora vamos verificar o ajuste do modelo

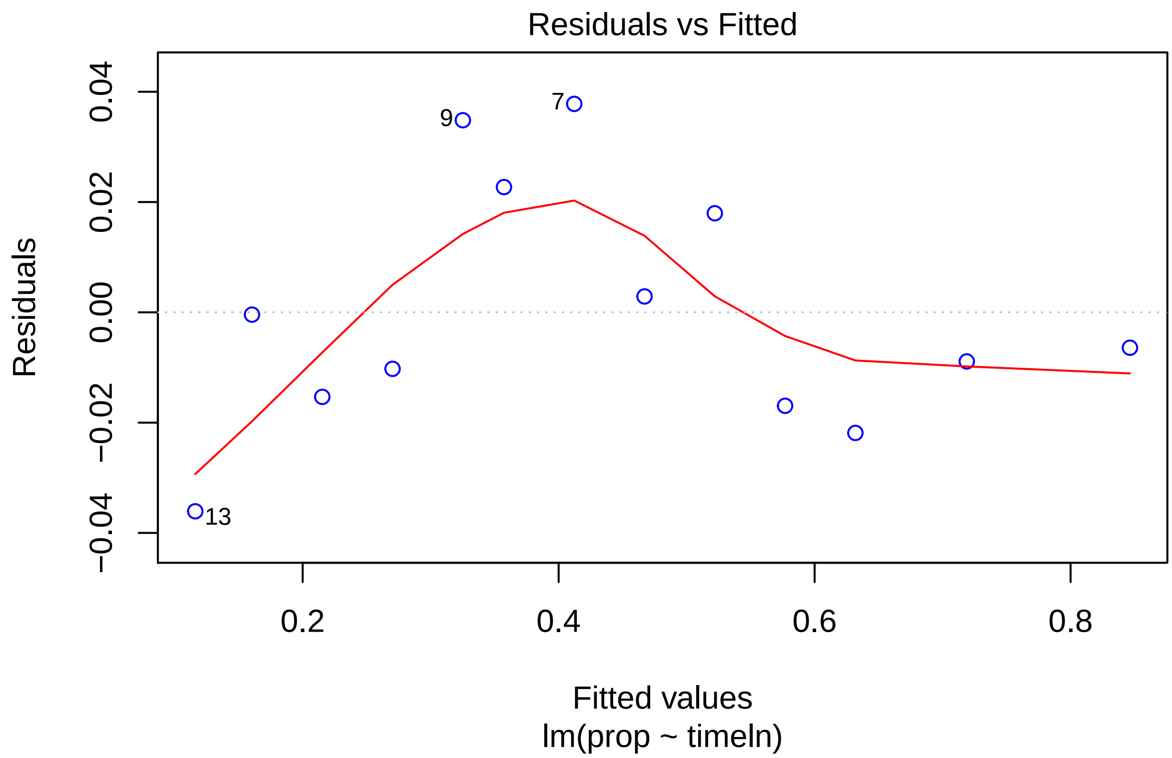
```r
# ajuste do modelo

plot(worldrecall$timeln, worldrecall$prop)
abline(lm(prop ~ timeln, data = worldrecall))
```



```r
plot(Linear_ln, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear_ln, which=1, col=c("blue")) # Q-Q plot
```



Ao observar os gráficos, percebemos que, com a transformação da variável independente em log, o modelo apresenta um ajuste adequado e uma relação linear entre as variáveis.

b)

Modelo com as variáveis do banco *worldrecall.txt*

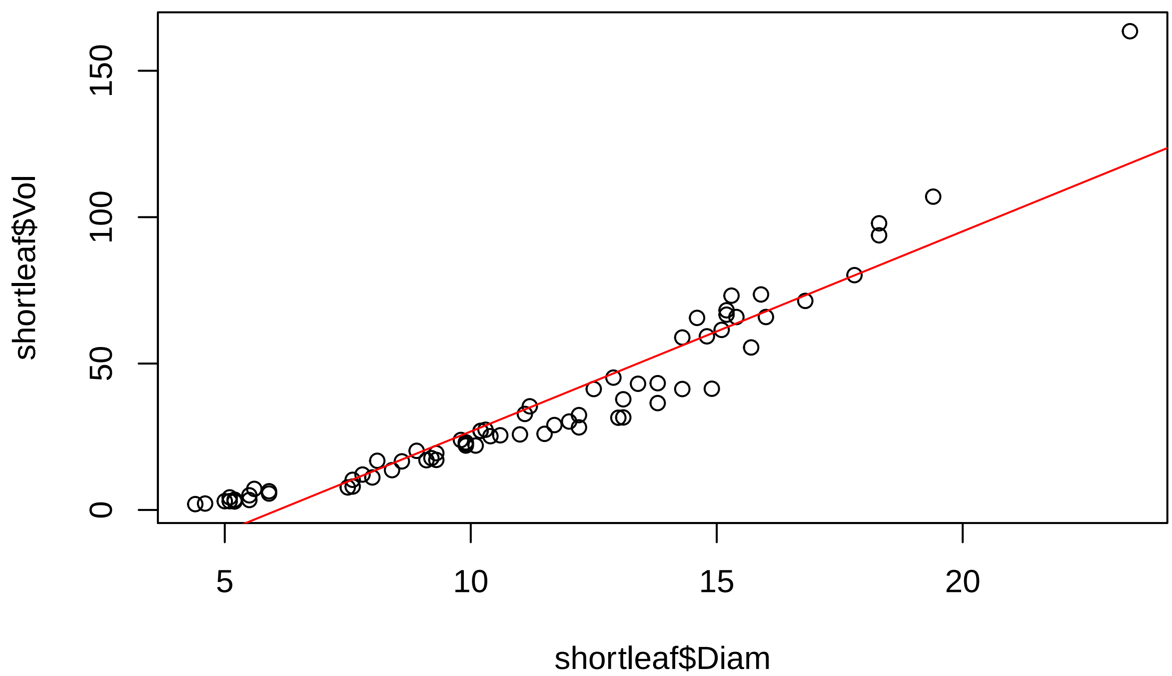
```r
# lendo banco shortleaf

shortleaf <- read.delim("~/Dados/Listas/AD_9/AD_9/shortleaf.txt")

# modelo linear

Linear <- lm(Vol ~ Diam, data = shortleaf)

# resumo do modelo

summary(Linear)
```

```
##
## Call:
## lm(formula = Vol ~ Diam, data = shortleaf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.899  -4.768  -1.438   6.740  45.089
```

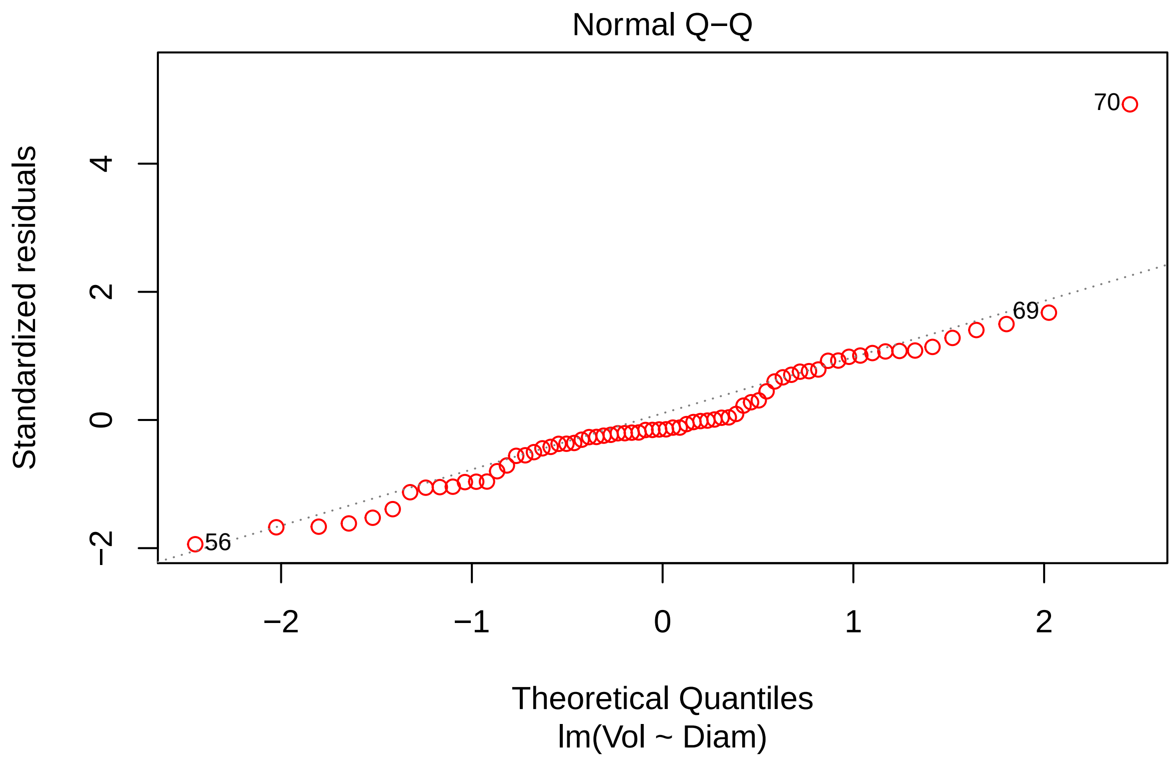
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41.5681     3.4269  -12.13   <2e-16 ***
## Diam          6.8367     0.2877   23.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.875 on 68 degrees of freedom
## Multiple R-squared:  0.8926, Adjusted R-squared:  0.891
## F-statistic: 564.9 on 1 and 68 DF,  p-value: < 2.2e-16
```

Os resultados mostram que o aumento de 1 unidade na VI leva a um aumento de 6.84 na VD. O resultado é estatisticamente significante (p-valor < 0.05) e o R^2 do modelo é de 0.89.

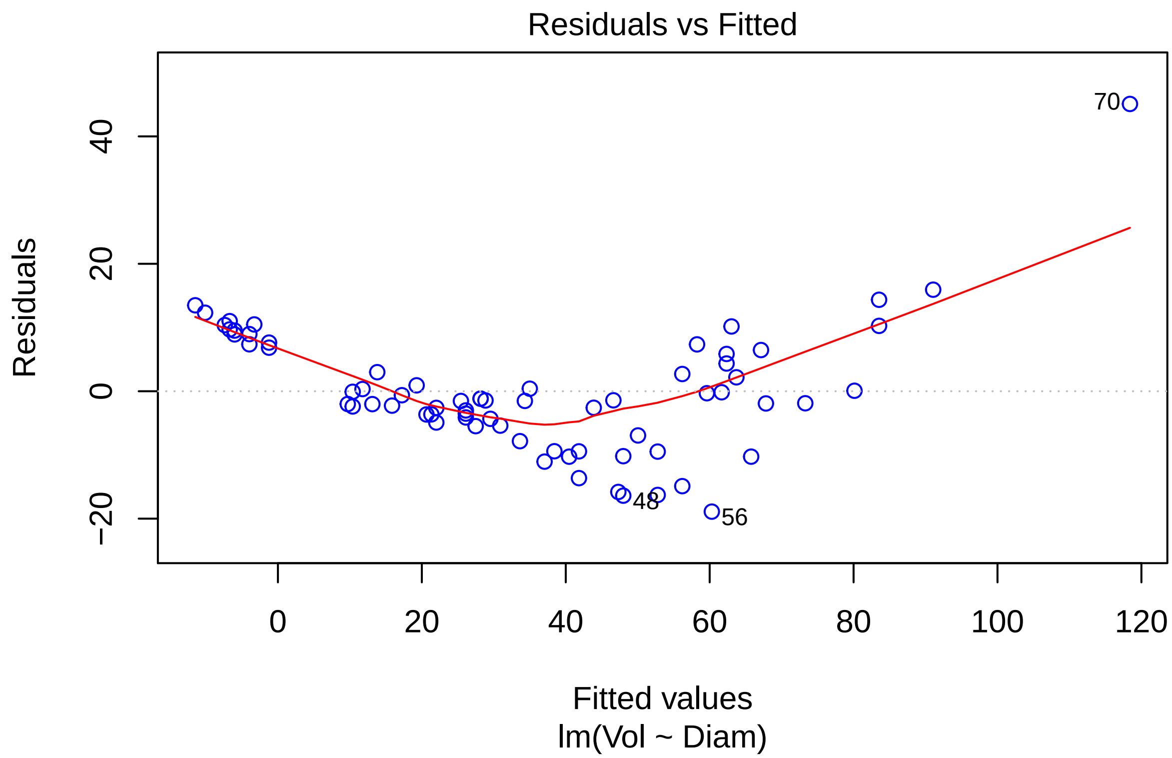
```r
# ajuste do modelo

plot(shortleaf$Diam, shortleaf$Vol)
abline(lm(Vol ~ Diam, data = shortleaf), col = 'red')
```




```r
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```

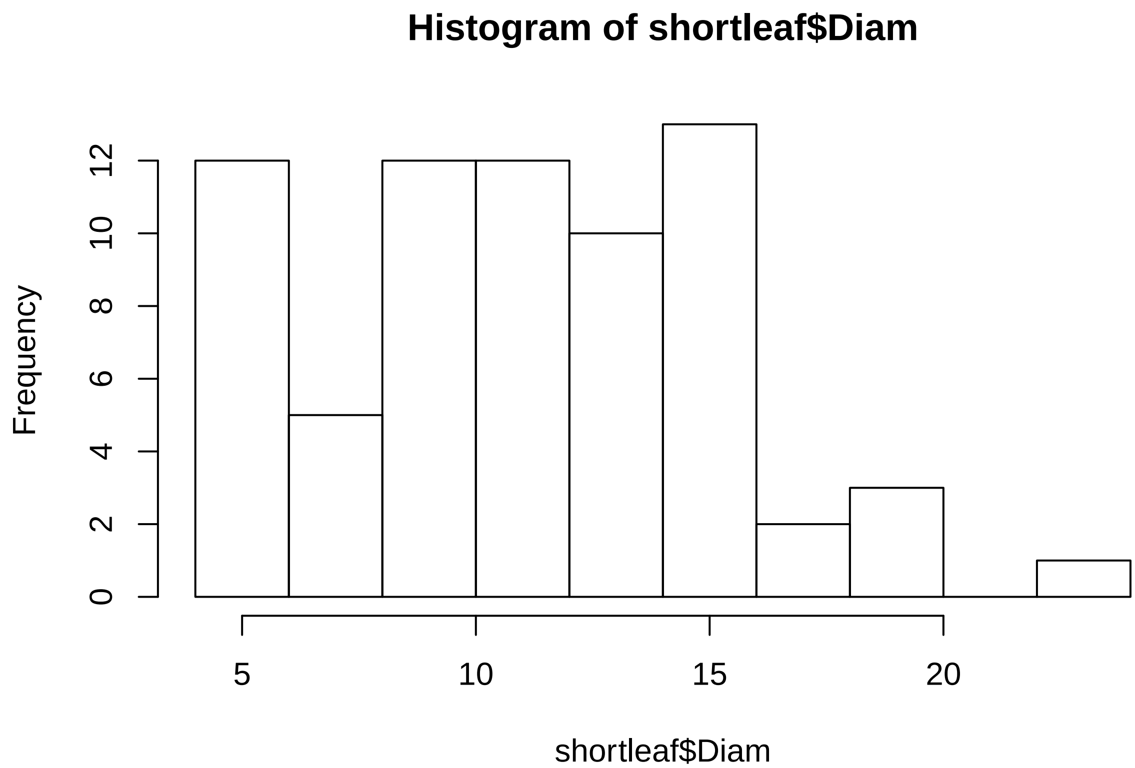



```
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```

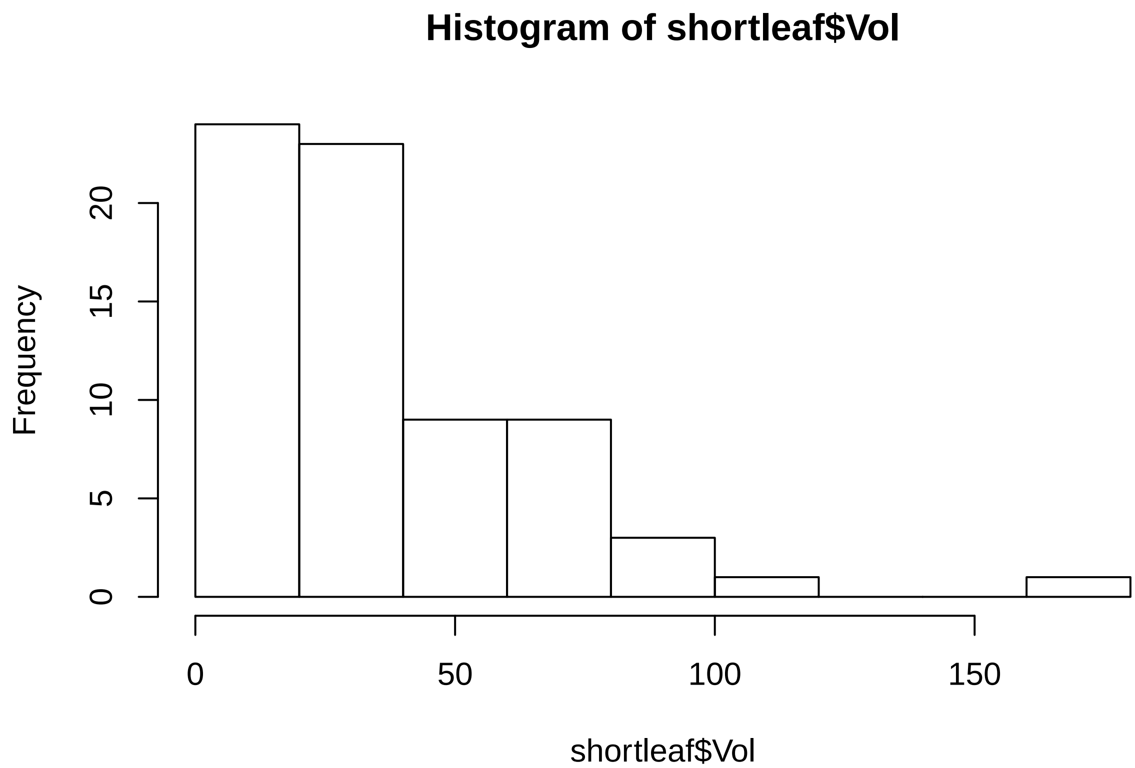


Os gráficos para analisar o ajuste do modelo mostram que a relação entre as duas variáveis não é linear. Ou seja, é necessário executar transformações na VD, VI ou em ambas. Para verificar isso, serão analisados o histograma de cada variável.

```
hist(shortleaf$Diam)
```




```
hist(shortleaf$Vol)
```



Observando os histogramas das duas variáveis, observa-se que ambas não apresentam uma distribuição próxima da normal. Assim, ambas serão transformadas em log.

```
  # logaritmo das duas variáveis

shortleaf$diam_log <- log(shortleaf$Diam)
shortleaf$Vol_log <- log(shortleaf$Vol)
```

Com as variaveis transformas em log, vamos executar outro modelo linear:

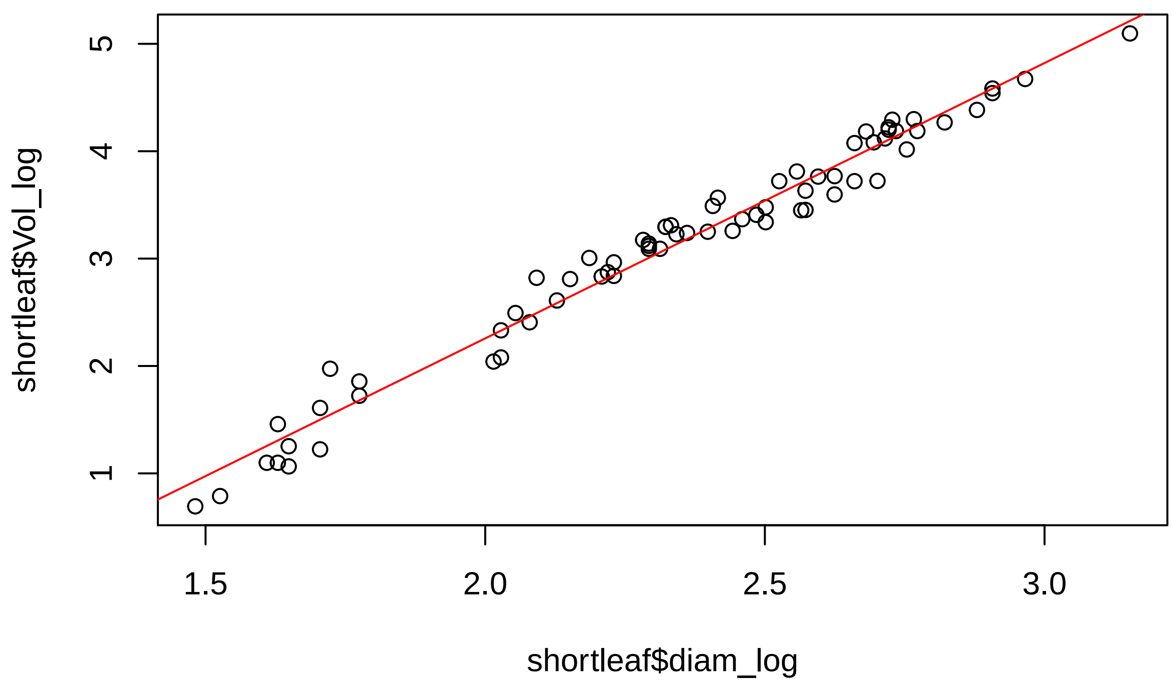
```
# modelo linear

Linear <- lm(Vol_log ~ diam_log, data = shortleaf)

# resumo do modelo

summary(Linear)
```

```
##
## Call:
## lm(formula = Vol_log ~ diam_log, data = shortleaf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3323 -0.1131  0.0267  0.1177  0.4280
##
```

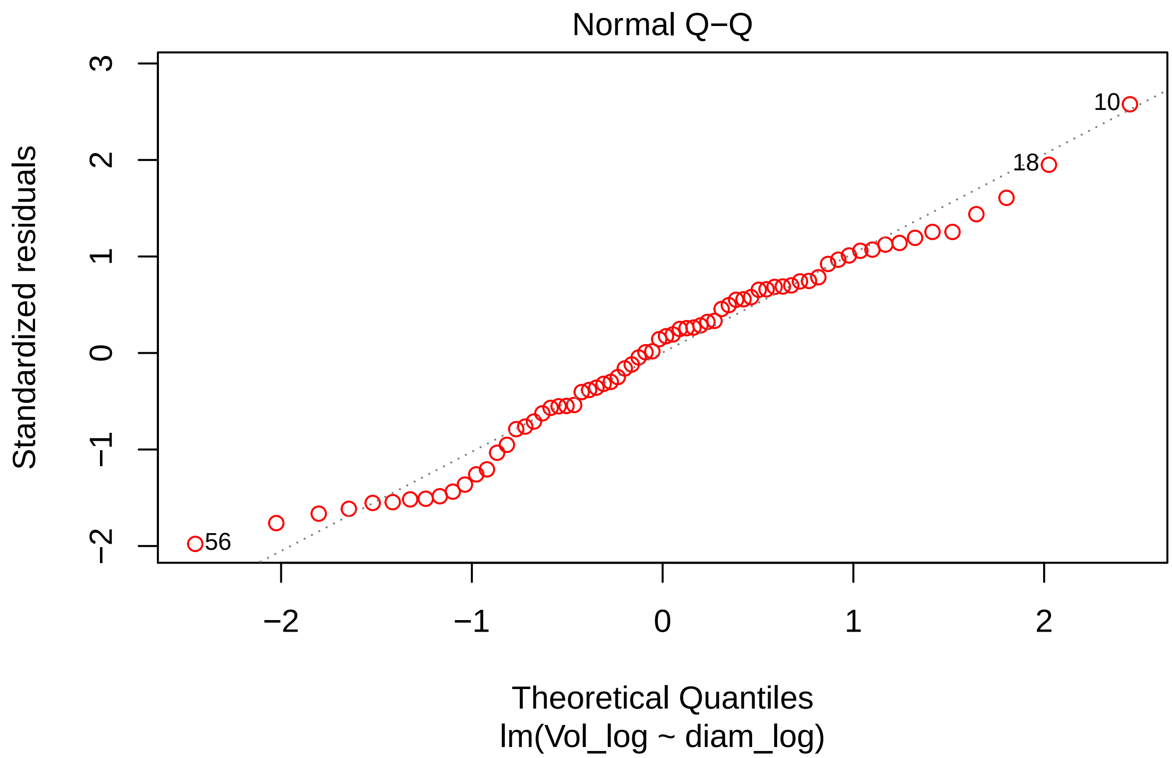
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.8718     0.1216  -23.63   <2e-16 ***
## diam_log      2.5644     0.0512   50.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1703 on 68 degrees of freedom
## Multiple R-squared:  0.9736, Adjusted R-squared:  0.9732
## F-statistic:  2509 on 1 and 68 DF,  p-value: < 2.2e-16
```

Os resultados do modelo mostra que 1% de variação na Variável Independente implica em 2,56% de variação na VD. Esse resultado é estatisticamente significativo (p-valor < 0.05) O R^2 do modelo é de 0.97. Agora, iremos verificar o ajuste do modelo:

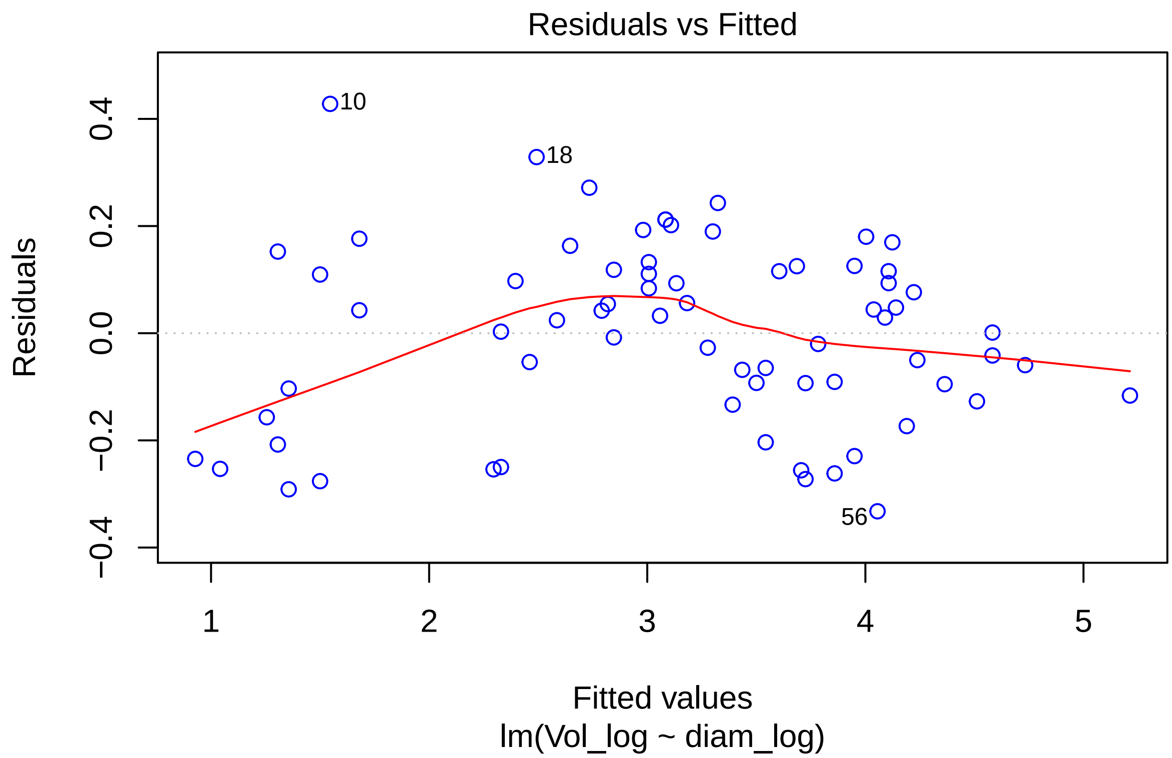
```
# ajuste do modelo

plot(shortleaf$diam_log, shortleaf$Vol_log)
abline(lm(Vol_log ~ diam_log, data = shortleaf), col = 'red')
```




```
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Como é possível observar, os gráficos mostram que o modelo linear com as variáveis em log apresentam uma relação linear, mostrando que o modelo é adequado.

c)

Modelo com as variáveis do banco *mammgest*

```
library("readxl")

# lendo banco mammgest

mammget <- read_excel("mammget.xlsx", col_types = c("numeric","numeric"))
```

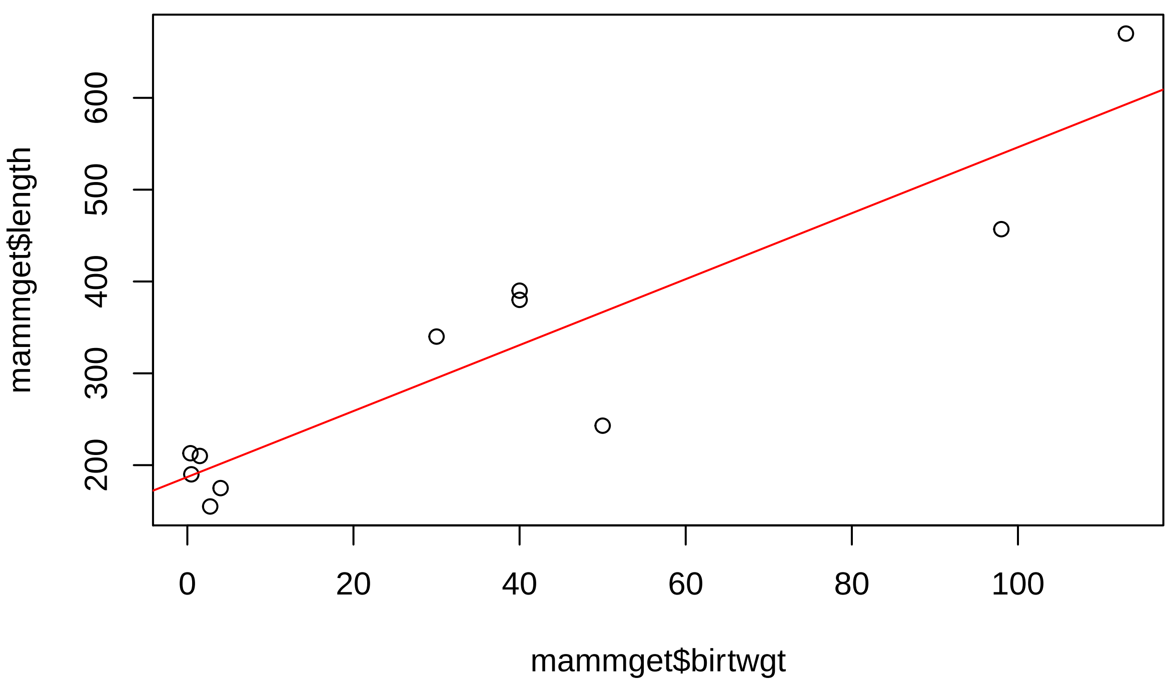
```
# modelo linear

Linear <- lm(length ~ birtwgt, data = mammget)

# resumo do modelo

summary(Linear)
```

```
##
## Call:
## lm(formula = length ~ birtwgt, data = mammget)
##
```

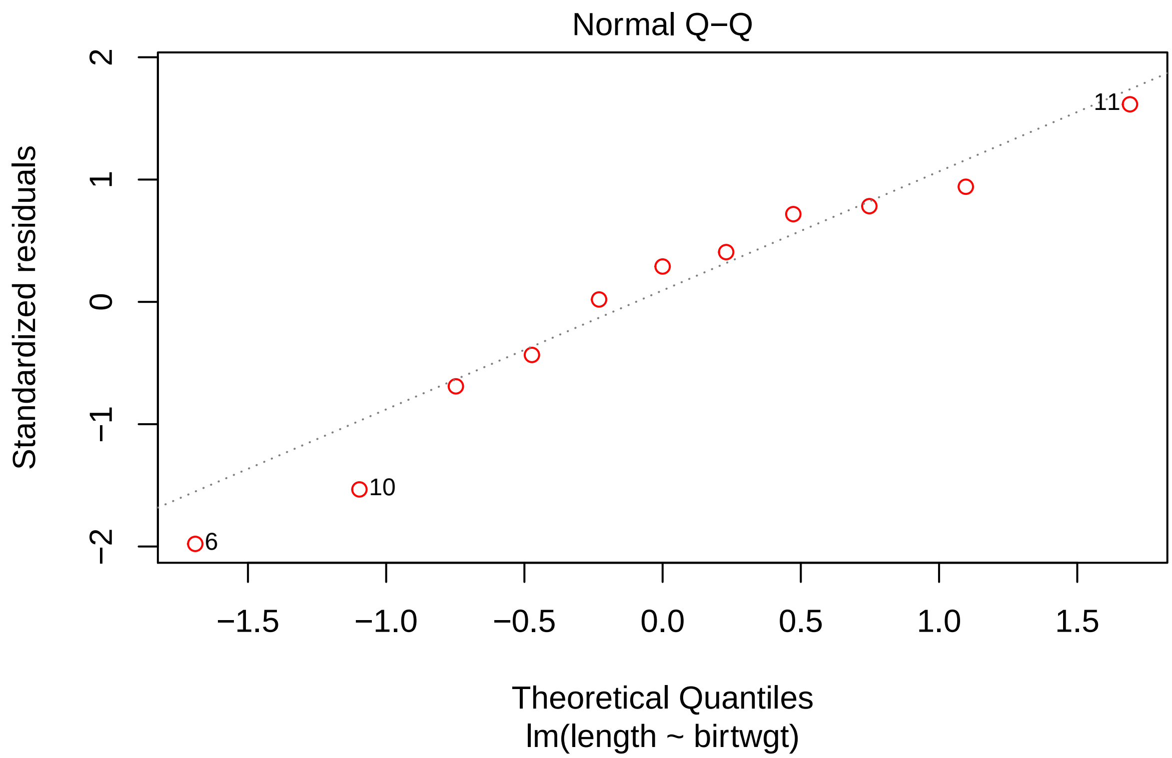
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.65  -34.20   17.53   47.22   77.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 187.0837    26.9426   6.944 6.73e-05 ***
## birtwgt       3.5914     0.5247   6.844 7.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.09 on 9 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.8209
## F-statistic: 46.84 on 1 and 9 DF,  p-value: 7.523e-05
```

Os resultados mostram que o aumento de 1 unidade na VI leva a um aumento de 3.59 na VD. O resultado é estatisticamente significante (p-valor < 0.05) e o R^2 do modelo é de 0.82.

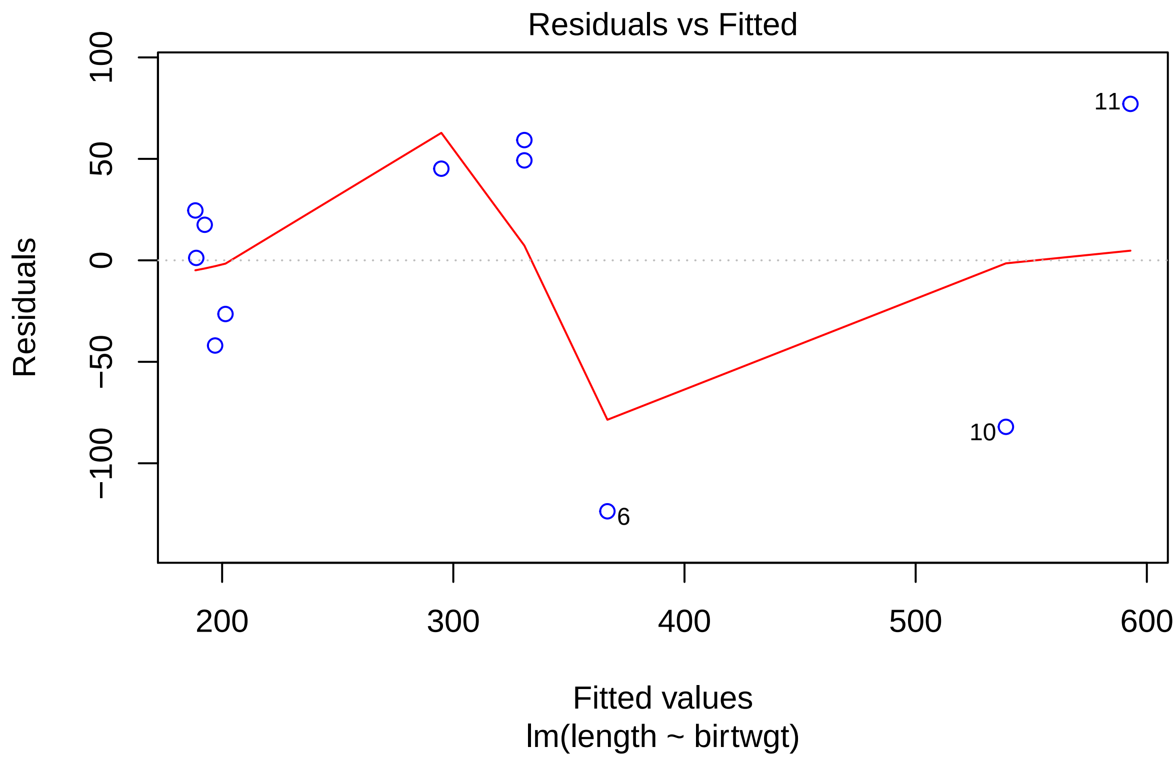
```r
# ajuste do modelo

plot(mammget$birtwgt, mammget$length)
abline(lm(length ~ birtwgt, data = mammget), col = 'red')
```




```r
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Os gráficos para analisar o ajuste do modelo mostram que a relação entre as duas variáveis não é linear. Ou seja, é necessário executar uma transformação logaritmica na VI.

```
  # logaritmo da VI

mammget$birtwgt_log <- log(mammget$birtwgt)
```

Com as variaveis transformas em log, vamos executar outro modelo linear:

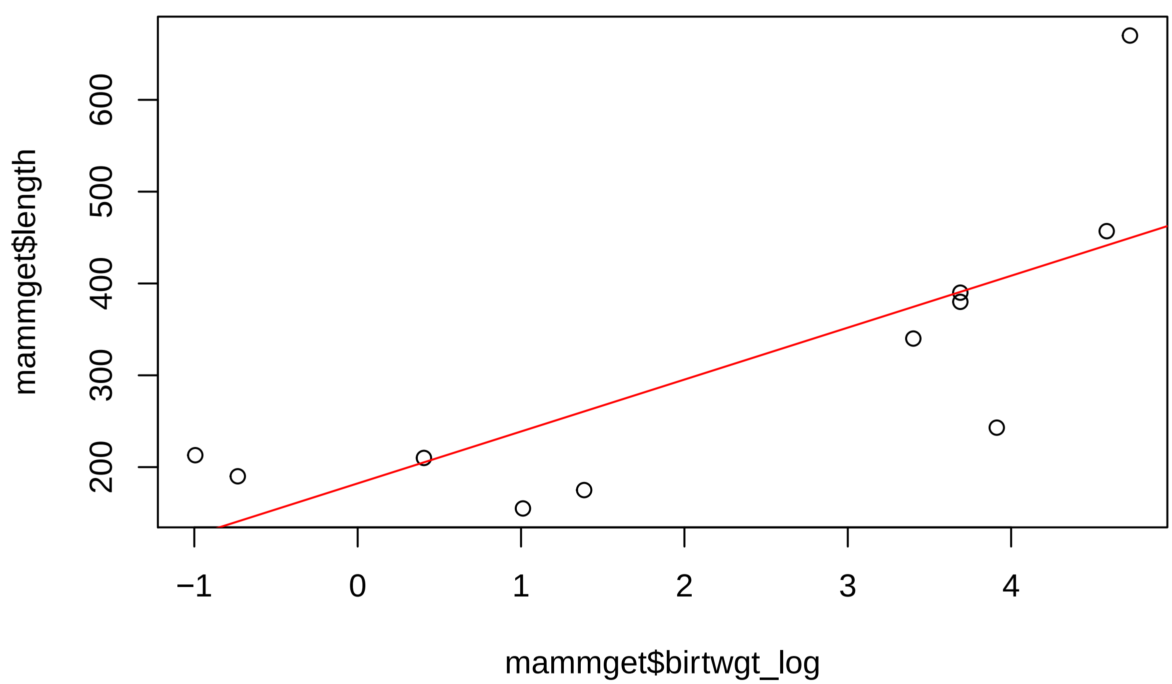
```
# modelo linear

Linear <- lm(length ~ birtwgt_log, data = mammget)

# resumo do modelo

summary(Linear)
```

```
##
## Call:
## lm(formula = length ~ birtwgt_log, data = mammget)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -160.46  -59.53   -0.84   32.35  220.45
##
## Coefficients:
```

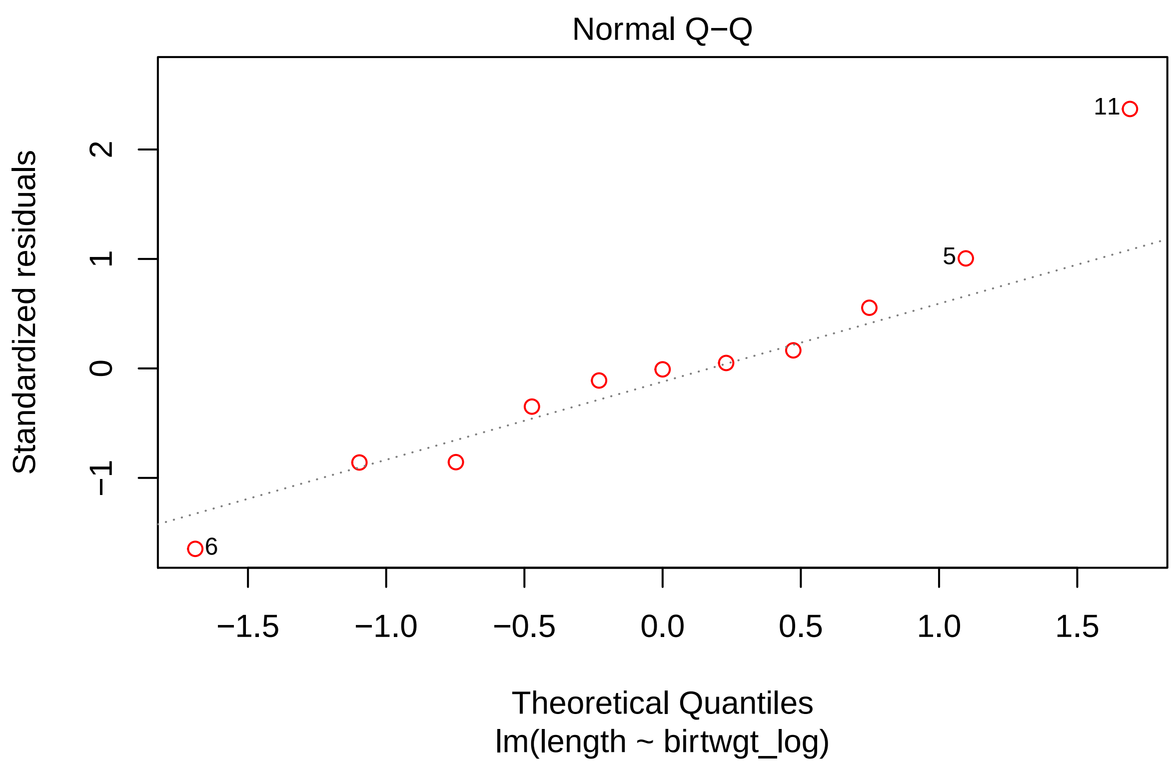
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   182.29      48.00   3.797  0.00423 **
## birtwgt_log    56.53      15.76   3.588  0.00586 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105.6 on 9 degrees of freedom
## Multiple R-squared:  0.5885, Adjusted R-squared:  0.5428
## F-statistic: 12.87 on 1 and 9 DF,  p-value: 0.005857
```

Os resultados do modelo mostra que 1% de variação na Variável Independente implica em 0.56 de variação na VD. Esse resultado é estatisticamente significativo (p-valor < 0.10) O R^2 do modelo é de 0.54. Agora, iremos verificar o ajuste do modelo:

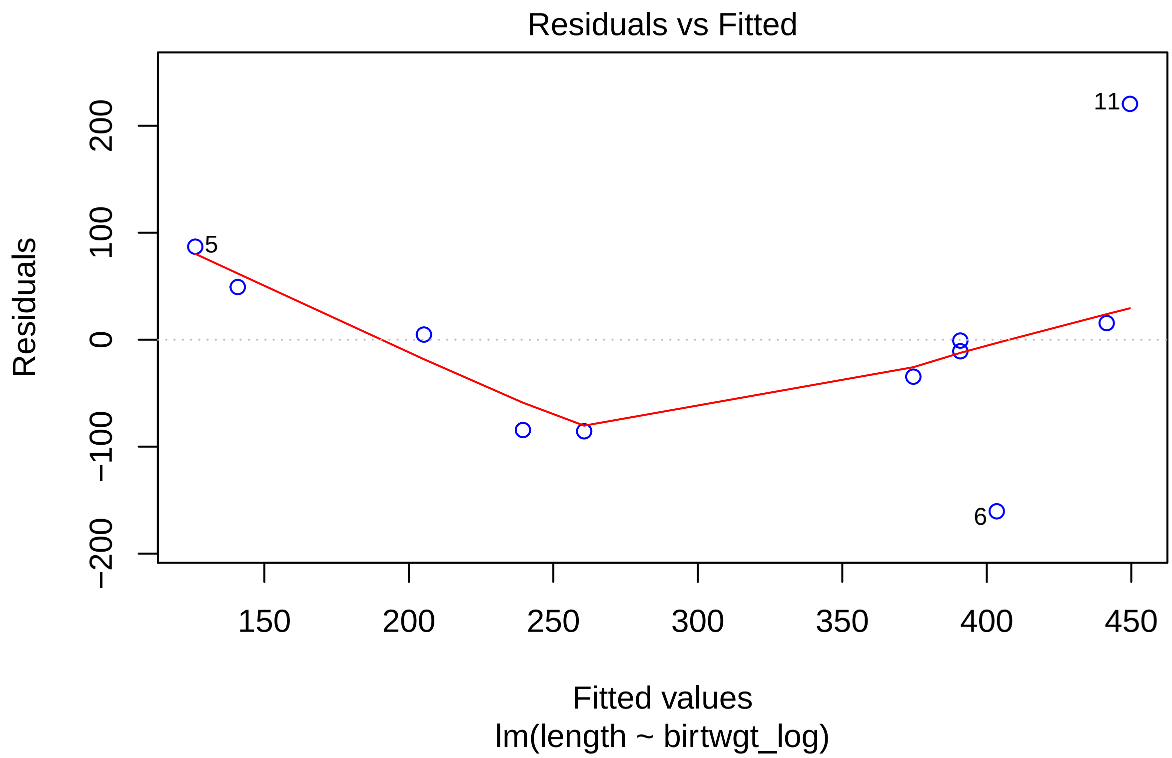
```
# ajuste do modelo

plot(mammget$birtwgt_log, mammget$length)
abline(lm(length ~ birtwgt_log, data = mammget), col = 'red')
```



```r
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Como é possível observar, os gráficos mostram que o modelo linear com a VI em log apresenta uma relação linear um pouco melhor que o modelo anterior, sendo esse modelo mais adequado para análise.

4.2

a)

Os modelos polinomiais apresentam uma distribuição dos dados não linear, que pode apresentar variações de acordo com o grau do modelo (quadrático, cúbico, etc). Desse modo, um modelo quadrático em um grafico de *residuals vs fitted* por exemplo tende a apresentar a distribuição dos dados de uma maneira não linear e não aleatório, mas como uma parábola. Assim, o arcabouço que temos acerca dos modelos lineares não é o apropriado para o desenvolvimento de modelos polinomiais. Os critérios de ajustes de um modelo linear não são os mesmos para um modelo polinomial.

b)

Modelo com as variáveis do banco *bluegills*

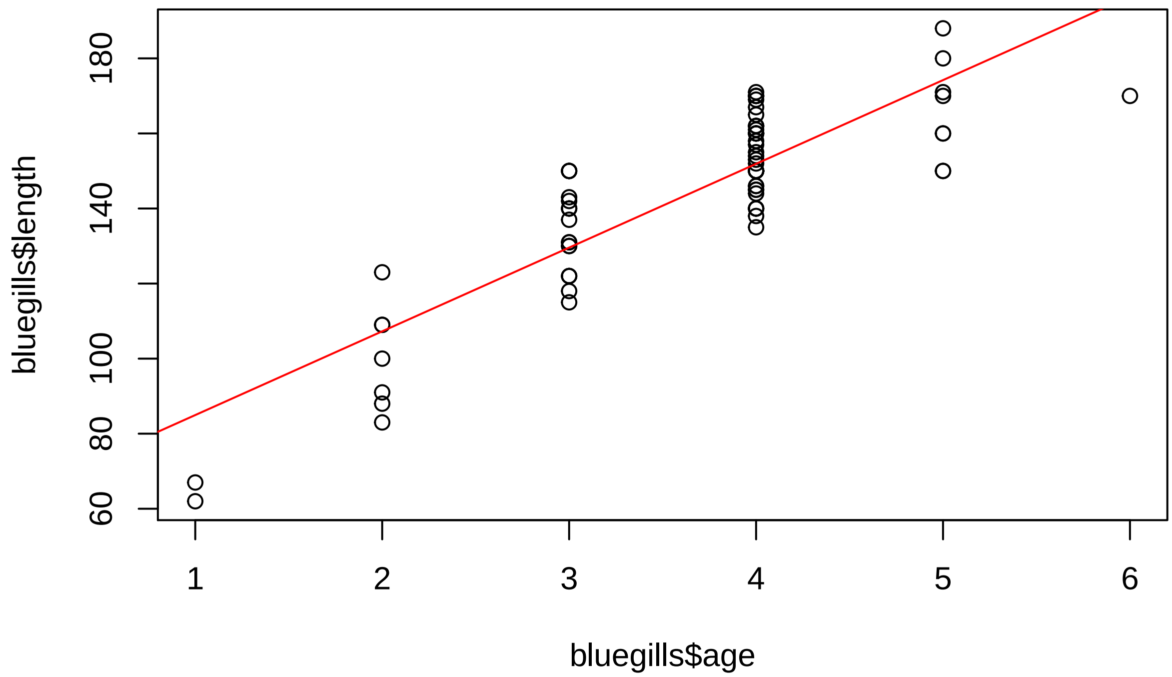
```
# lendo banco bluegills

bluegills <- read.delim("~/Dados/Listas/AD_9/AD_9/bluegills.txt")

# modelo linear

Linear <- lm(length ~ age, data = bluegills)
```

```r
# resumo do modelo

summary(Linear)
```

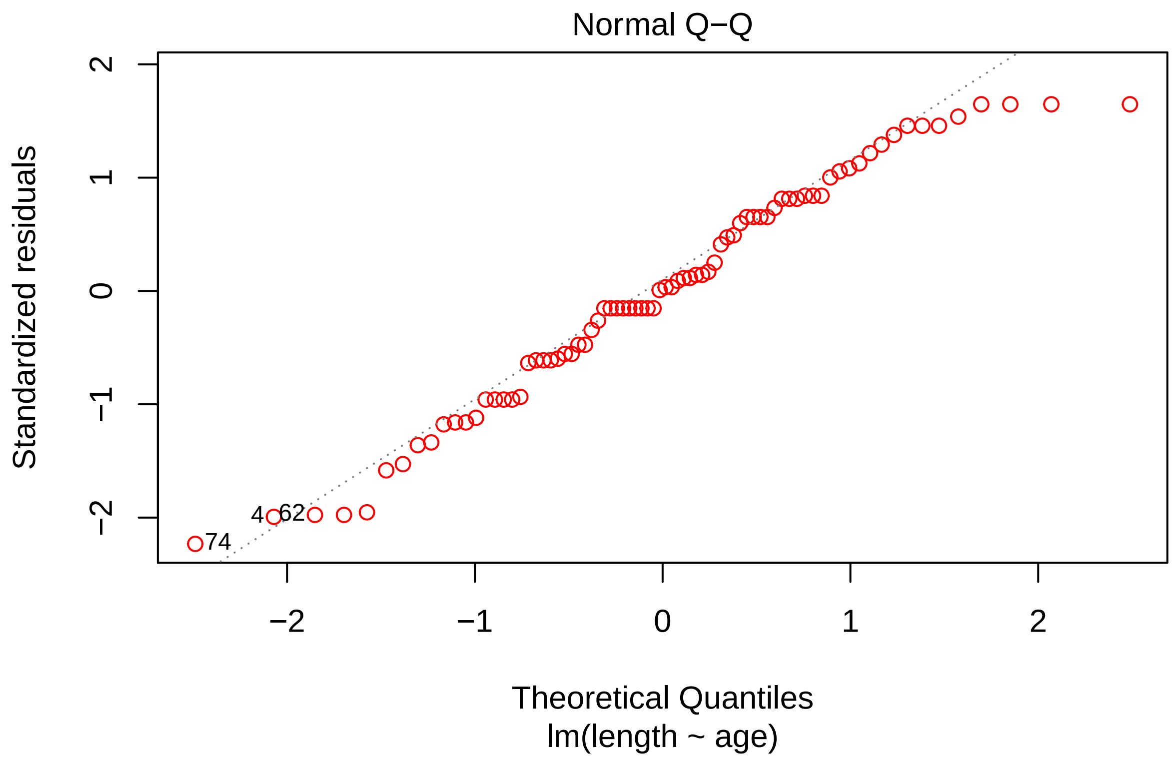
```
##
## Call:
## lm(formula = length ~ age, data = bluegills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.523  -7.586   0.258  10.102  20.414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.649      5.755   10.89   <2e-16 ***
## age           22.312      1.537   14.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.51 on 76 degrees of freedom
## Multiple R-squared:  0.7349, Adjusted R-squared:  0.7314
## F-statistic: 210.7 on 1 and 76 DF,  p-value: < 2.2e-16
```

O resultado do modelo linear apresenta um resultado estatisticamente positivo da VI sobre VD (p-valor < 0.05). O aumento de uma unidade na VI representa o aumento de 22.31 unidades na VD. O R^2 do modelo é de 0.73. Vamos ver o ajuste do modelo:

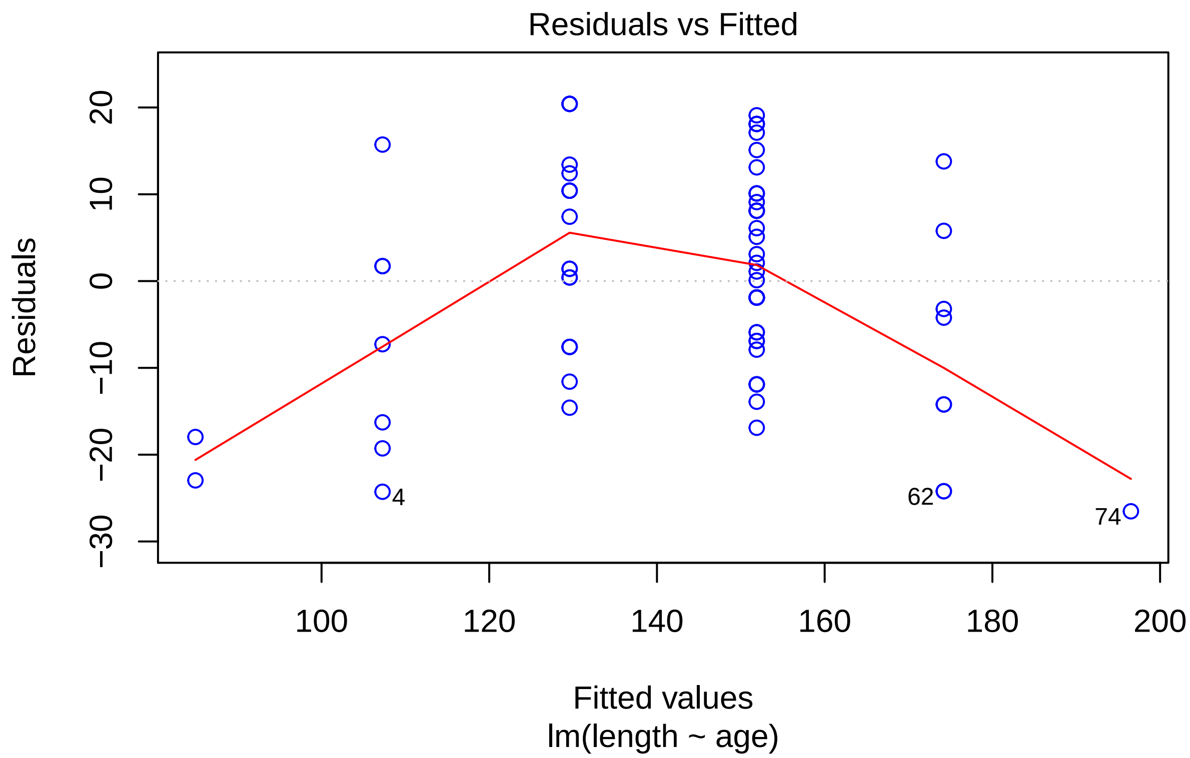
```r
# ajuste do modelo

plot(bluegills$age, bluegills$length)
abline(lm(length ~ age, data = bluegills), col = 'red')
```




```
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Os resultados do ajuste do modelo mostram claramente que as relações entre as duas variáveis não seguem uma relação linear. Desse modo, é necessário executarmos um modelo quadrático.

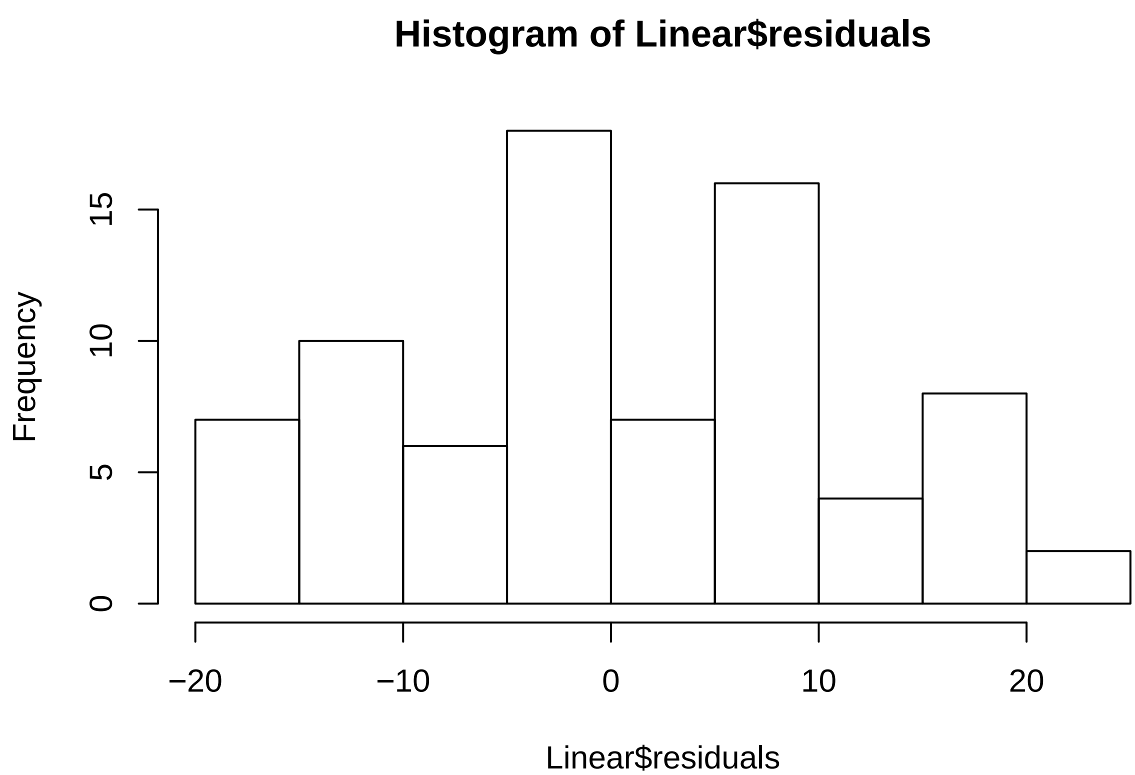
```
# elevando a VI ao quadrado

bluegills$age_sq <- bluegills$age^2

# rodando o modelo com a vi ao quadrado

Linear <- lm(length ~ age + age_sq, data = bluegills)

# modelo

summary(Linear)
```

```
##
## Call:
## lm(formula = length ~ age + age_sq, data = bluegills)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.846  -8.321  -1.137   6.698  22.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.622     11.016   1.237     0.22
## age           54.049      6.489   8.330 2.81e-12 ***
```

```
## age_sq        -4.719      0.944  -4.999 3.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.91 on 75 degrees of freedom
## Multiple R-squared:  0.8011, Adjusted R-squared:  0.7958
## F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

O sumário do modelo mostra que a capacidade explicativa do modelo é de 0.79 (R^2) em detrimento de 0.73 do modelo anterior.

```
# ajuste do modelo

hist(Linear$residuals)
```



Como podemos observar com o histrograma, os residuos do modelo se aproximam bastante de uma distribuição normal, mostrando uma maior adequação do modelo. Para verificar o efeito da VI na VD, não podemos interpretar como um modelo linear, mas é necessário levar em conta as duas VIs de mandeira simultanea. Para isso, vamos utilizar o comando abaixo que sumariza estes resultados:

```
Linear$fitted.values
```

```
##         1         2         3         4         5         6         7
##  62.95302  62.95302 102.84634 102.84634 102.84634 102.84634 133.30233
##         8         9        10        11        12        13        14
## 133.30233 133.30233 133.30233 133.30233 133.30233 133.30233 133.30233
```

```
##        15        16        17        18        19        20        21
## 133.30233 102.84634 133.30233 154.32099 154.32099 154.32099 154.32099
##        22        23        24        25        26        27        28
## 154.32099 154.32099 154.32099 154.32099 154.32099 154.32099 154.32099
##        29        30        31        32        33        34        35
## 154.32099 154.32099 154.32099 154.32099 154.32099 154.32099 154.32099
##        36        37        38        39        40        41        42
## 154.32099 154.32099 165.90232 154.32099 154.32099 154.32099 165.90232
##        43        44        45        46        47        48        49
## 102.84634 102.84634 154.32099 133.30233 154.32099 133.30233 154.32099
##        50        51        52        53        54        55        56
## 154.32099 154.32099 154.32099 133.30233 133.30233 133.30233 154.32099
##        57        58        59        60        61        62        63
## 154.32099 133.30233 154.32099 165.90232 154.32099 165.90232 154.32099
##        64        65        66        67        68        69        70
## 154.32099 133.30233 165.90232 165.90232 154.32099 165.90232 133.30233
##        71        72        73        74        75        76        77
## 154.32099 133.30233 154.32099 168.04632 154.32099 165.90232 154.32099
##        78
## 154.32099
```

Nesse caso, os resultados mostram que o aumento de uma unidade na VI não é o mesmo para todo aumento. Ou seja, o aumento de 0 para 1 unidade da VI resulta em um efeito na VD diferente do aumento da unidade 37 para a 38 da VI. Isso se dá devido a curva do modelo não ser linear, apresentado valores diferentes de acordo com o valor observado.