Análise de Dados - UFPE/2019 - Lista 9

Maria Eduarda R. N. Lessa

11 de junho de 2019

Questão 1:

Xi e Yi - Correlação positiva

Zi e Yi - Correlação negativa

letra a)

Se a correlação entre Xi e Zi é zero, esta última não terá efeito sobre a análise da relação entre Xi e Yi e, portanto, é possível omiti-la. Esta situação, no entanto, é improvável de ocorrer.

letra b)

Se a correlação entre Xi e Zi for positiva, é necessário analisar a relação entre Xi e Yi controlando para Zi, visto que a correlação positiva entre estas duas VI's pode superestimar o verdadeiro efeito de Xi em Yi.

letra c)

Se a correlação entre Xi e Zi for negativa, o efeito de Xi sobre Yi pode ter sido subestimado; assim como no item anterior, é necessário controlar para Zi para ter uma melhor aproximação do real efeito de Xi sobre Yi.

Questão 2:

No primeiro modelo (A), o salário médio dos professores de escolas públicas nos EUA (nos níveis de elementary e secondary school) é a variável dependente. A porcentagem de residentes dos estados americanos com diploma universitário é a variável independente. O efeito encontrado é significativo, com o ^alfa = 28768.01; ^beta = 704.02; erro padrão = 140.22 e p-valor $<$ 0.05. Estes resultados apontam que quando a VI = 0, o valor da VD é representado por ^alfa e que o aumento de 1 unidade na VI é responsável por um aumento de 704.2 unidades da VD (neste caso, provavelmente, a unidade da VD é dólar americano). É possível concluir que o aumento de 1% no total(%) de residentes com diploma universitário em um estado é responsável por um aumento de cerca de 704.02 dólares na média salarial dos professores deste mesmo estado. O R^2 aponta que 0.34 da variação na VD (salário médio) deve-se ao efeito da VI (% de residentes com diploma).

No segundo modelo (B), a VD é a mesma (o salário médio dos professores) e a VI é a renda per capita. O efeito encontrado é significativo, com o ^alfa = 21168.11; ^beta = 0.68; erro padrão = 0.11 e p-valor < 0.05. Estes resultados apontam que quando a VI = 0, o valor da VD é representado por ^alfa e que o aumento de uma unidade na VI é responsável por um aumento de 0.68 na VD. O R^2 aponta que 0.47 da variação na VD (salário médio) deve-se ao efeito da VI (renda per capita).

Questão 3:

No terceiro modelo (C) são analisados os efeitos de cada uma das VIs sobre a VD, ou seja, o efeito do percentual de habitantes com diploma universitário sobre o salário médio dos professores, controlado pela renda per capita; assim como o efeito da renda per capita sobre o salário médio dos professores, controlado pelo percentual de habitantes com diploma. Para a primeira VI, o efeito encontrado não foi significativo, já para a segunda (renda), o ^beta de 0.66 apresentou p-valor < 0.05. O R^2 apresentou o mesmo valor encontrado no segundo modelo, de 0.47.

É possível notar, ao comparar os modelos, que ao controlar pela renda, o efeito da primeira VI (%diploma) deixa de ser significativo e também que a sua inclusão, neste terceiro modelo (C), não alterou a capacidade explicativa obtida no segundo (B). Para a VI renda, com o controle para a VI % diploma, o valor do ^beta diminuiu um pouco, o que é um indício de que o valor do ^beta no segundo modelo havia sido discretamente superestimado, devido à provável correlação positiva entre estes duas VIs. Finalmente, a capacidade explicativa do modelo C é maior que a do modelo A e igual a do modelo B, o que aponta que no modelo A o efeito da VI analisada era, na verdade, o efeito compartilhado por renda e % diploma.

Questão 4:

4.1

letra a)

```r
# Carregar base:
require(ggplot2)
require(readr)
setwd("C:/Users/Duda/Desktop/PPGCP/Análise de Dados/lista_09")
WordRecall <- read_tsv("wordrecall.txt")
```

```r
head(WordRecall)
```

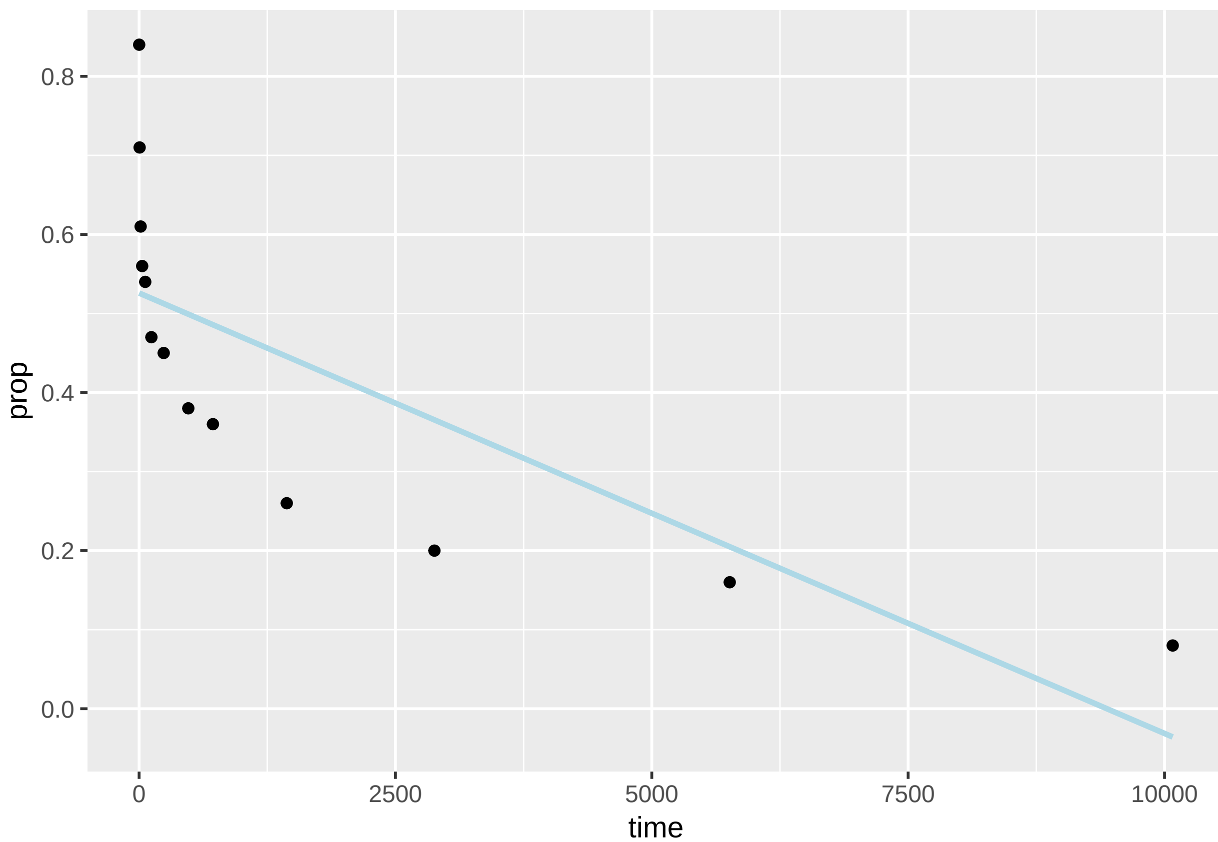
```
## # A tibble: 6 x 2
##    time  prop
##   <dbl> <dbl>
```

```
## 1      1  0.84
## 2      5  0.71
## 3     15  0.61
## 4     30  0.56
## 5     60  0.54
## 6    120  0.47
```

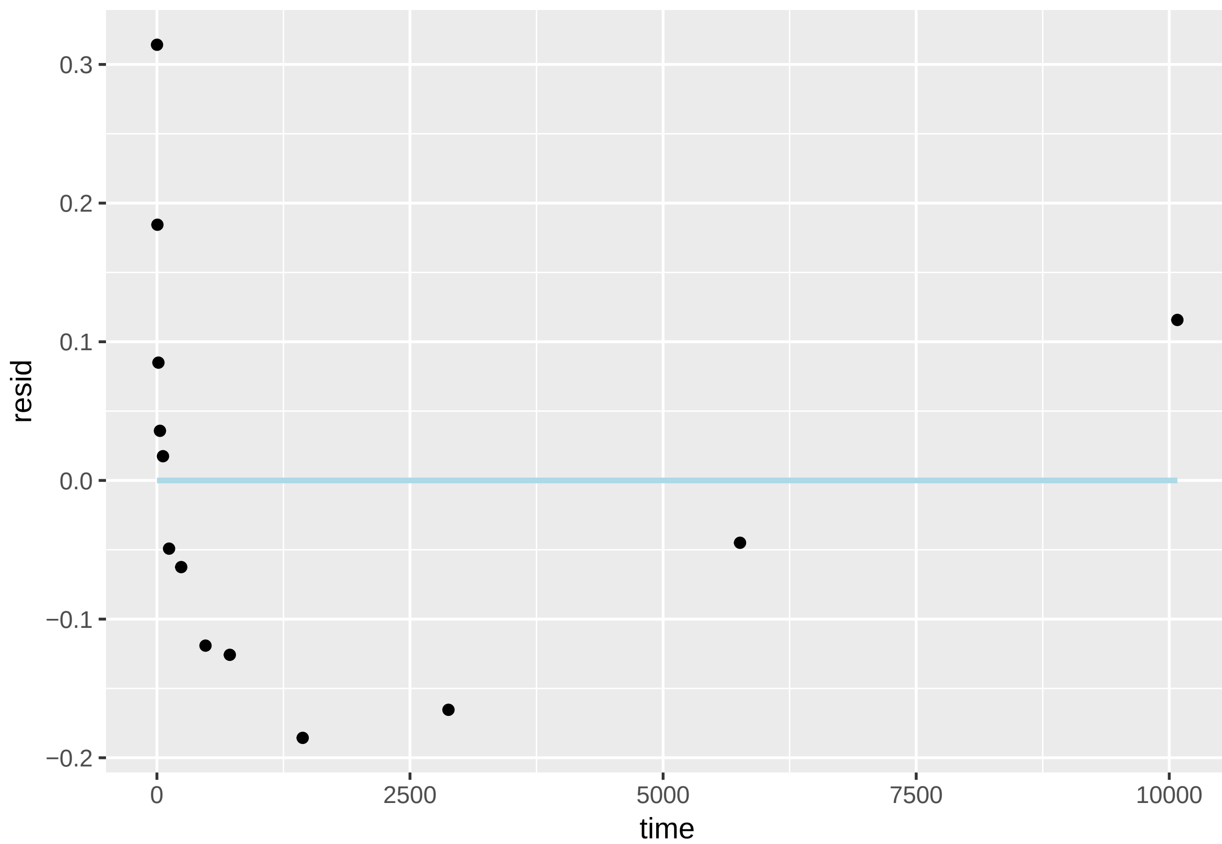
```r
# Modelo linear:
reg <- lm(prop ~ time, data = WordRecall)

# Analisar regressão:
summary(reg)
```

```
##
## Call:
## lm(formula = prop ~ time, data = WordRecall)
##
## 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
```

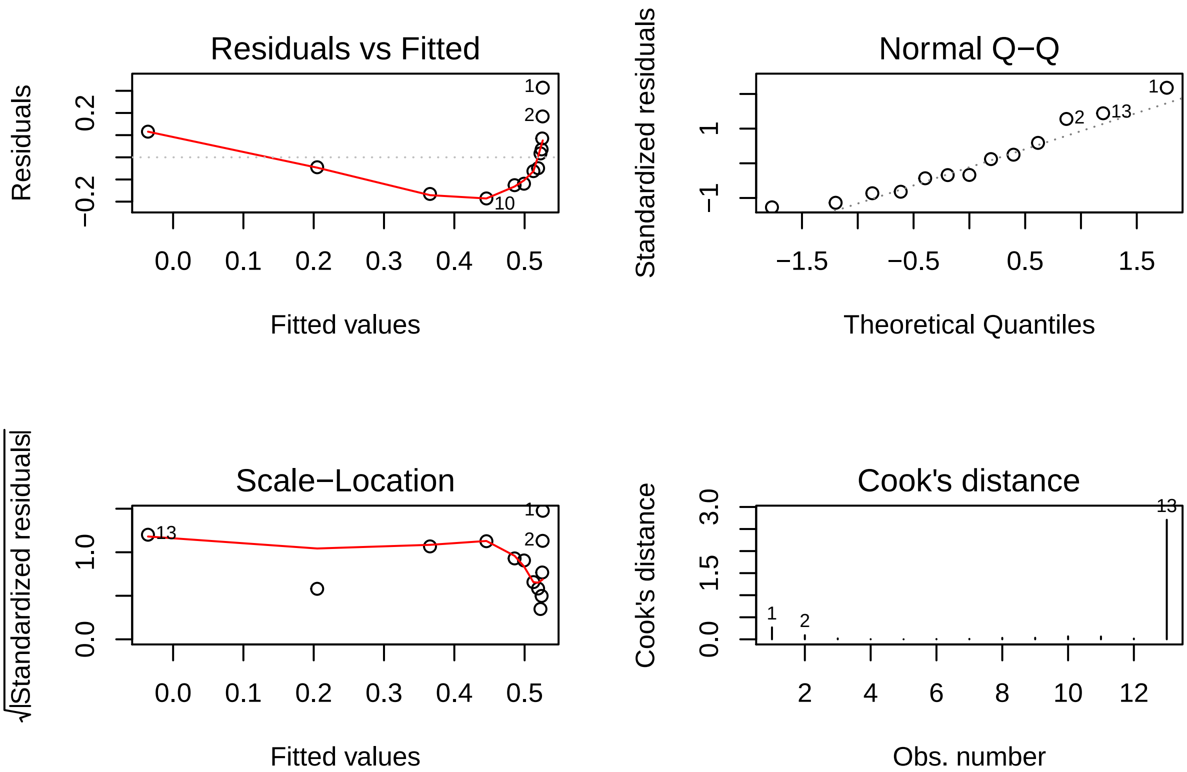
```r
# Plotar:
ggplot(data = WordRecall, aes(y = prop, x = time))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```
# Analisar resíduos:
# Opção 1:
resid <- resid(reg)
ggplot(data = WordRecall, aes(y = resid, x = time)) +
         geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg, which = 1:4)
```



Neste modelo linear, 0.57 da variação da VD "proporção de itens lembrados" (prop), deve-se ao efeito da VI "tempo" (time). O ^beta = -5.57 aponta que para o aumento de 1 unidade na VI time, haverá uma diminuição de 5.57 unidades na VD prop. O erro médio quadrártico (RSE) é de 0.15. Ao plotar a regressão é possível perceber que alguns pressupostos foram violados; há uma relação não linear entre as variáveis, é possível identificar que há resíduos com valores destoantes, além de não estarem distribuídos aleatoriamente ao longo da linha da média no gráfico da análise residual. A distribuição dos resíduos parece aproximar-se de uma normal.

```
# Modelo level-log:

# Transformar dados e criar nova base:
```

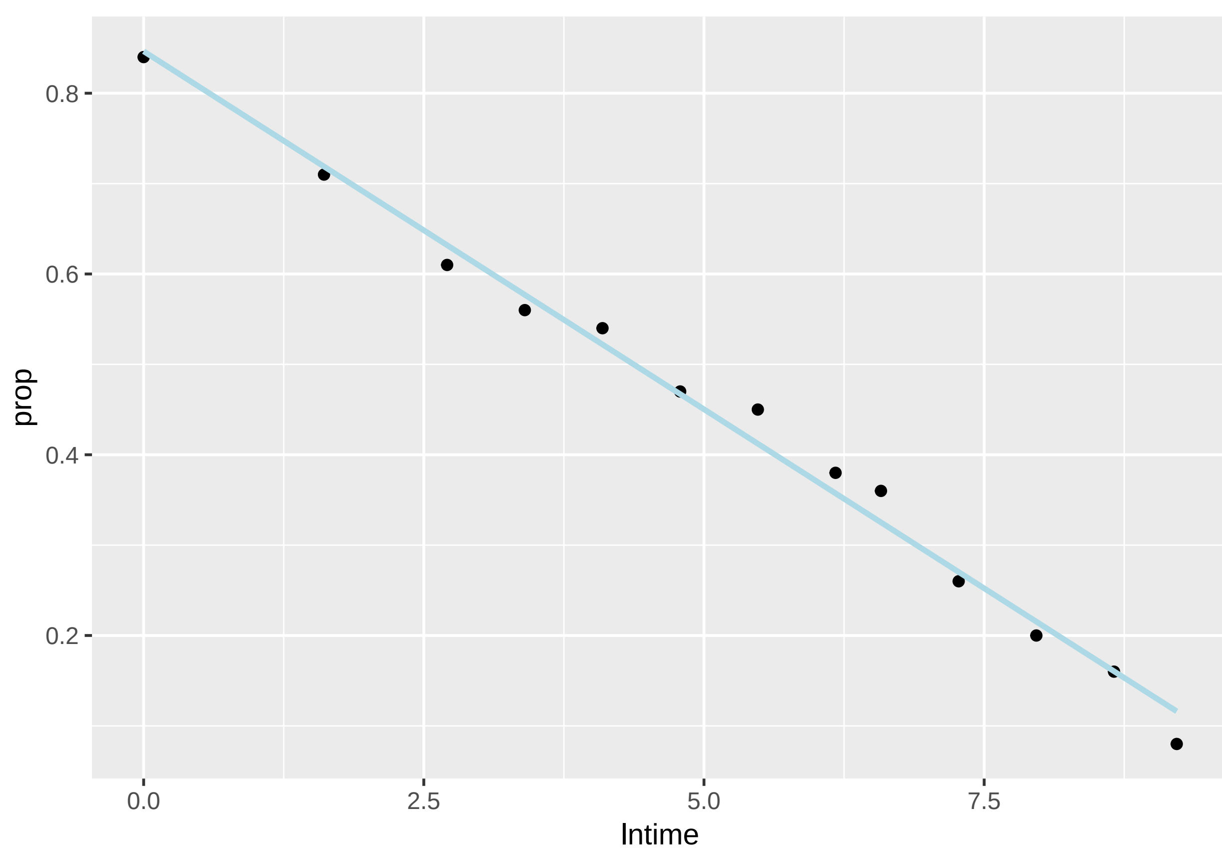
```
require(dplyr)
```

```
lntime <- log(WordRecall$time)
WordRecall2 <- mutate(WordRecall, new = lntime)

# Regressão
reg.log <- lm(prop ~ lntime, data = WordRecall2)
summary(reg.log)
```

```
##
## Call:
```

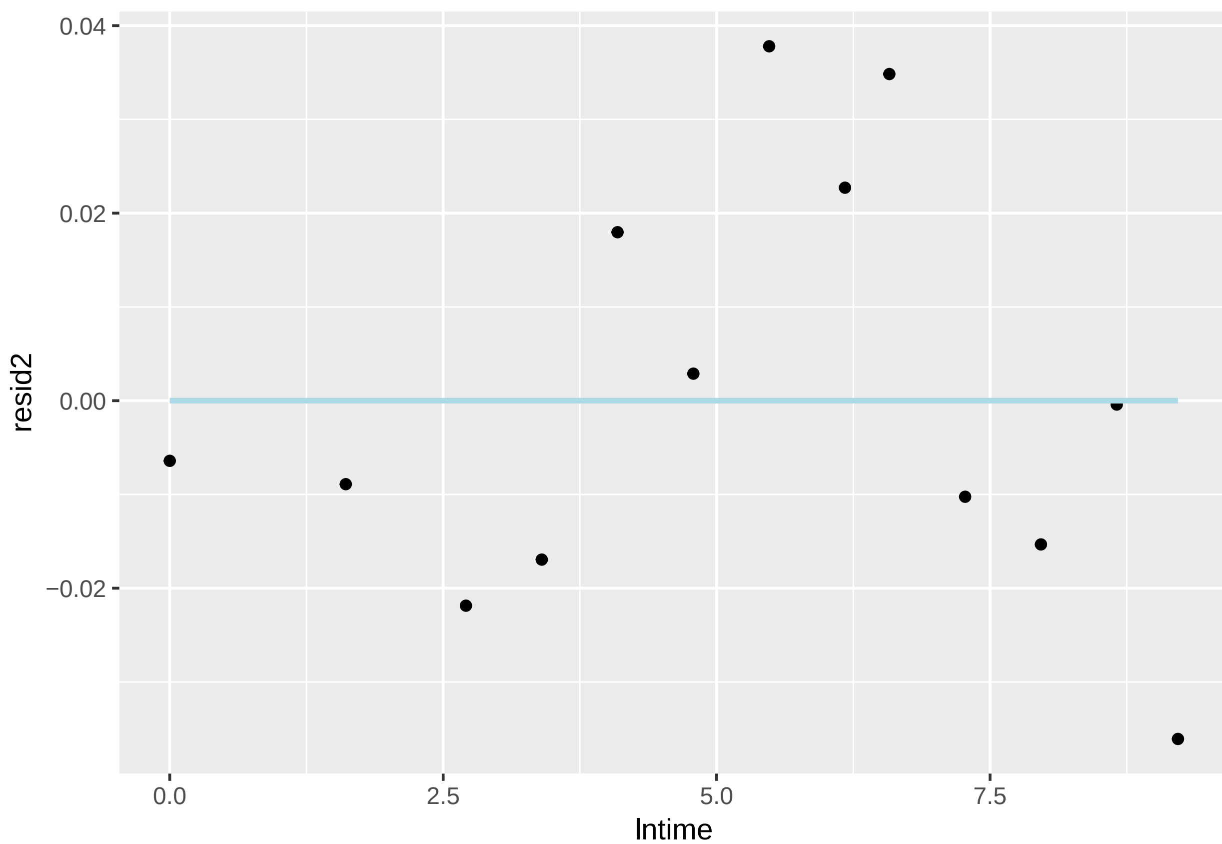
```
## lm(formula = prop ~ lntime, data = WordRecall2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036077 -0.015330 -0.006415  0.017967  0.037799
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
## lntime      -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
```

```
# Plotar:
ggplot(data = WordRecall2, aes(y = prop, x = lntime))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```

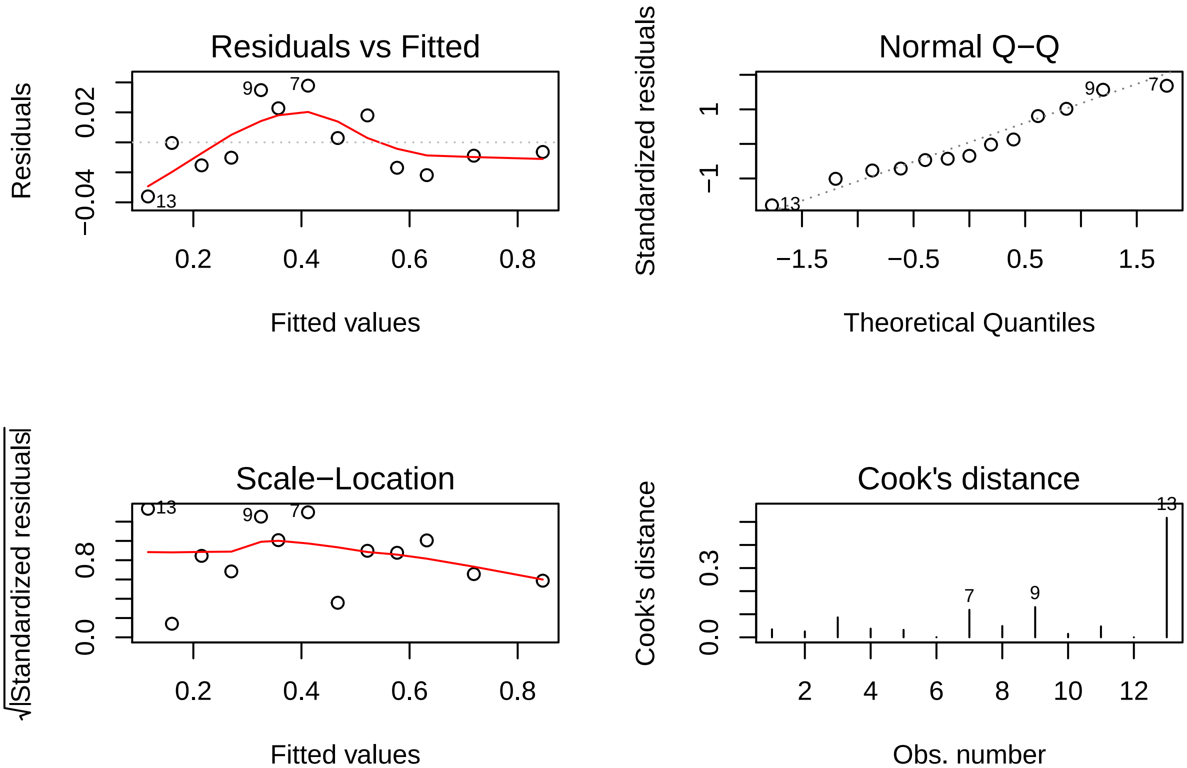



```
# Analisar resíduos:
# Opção 1:
resid2 <- resid(reg.log)
```

```r
ggplot(data = WordRecall2, aes(y = resid2, x = lntime)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg.log, which = 1:4)
```



Neste segundo modelo, a capacidade explicativa aumentou; o valor do R^2 aponta que 0.99 da variação na VD deve-se ao efeito da VI em questão. Este modelo também apresentou valores mais significativos para ^alfa e ^beta (menor p-valor). O pressuposto da linearidade passa a ser atendido quando a variável independente é transformada em log.

letra b)

```
# Carregar base:
setwd("C:/Users/Duda/Desktop/PPGCP/Análise de Dados/lista_09")
ShortLeaf <- read_tsv("shortleaf.txt")
```

```
## Parsed with column specification:
## cols(
##   Diam = col_double(),
##   Vol = col_double()
## )
```

```
head(ShortLeaf)
```

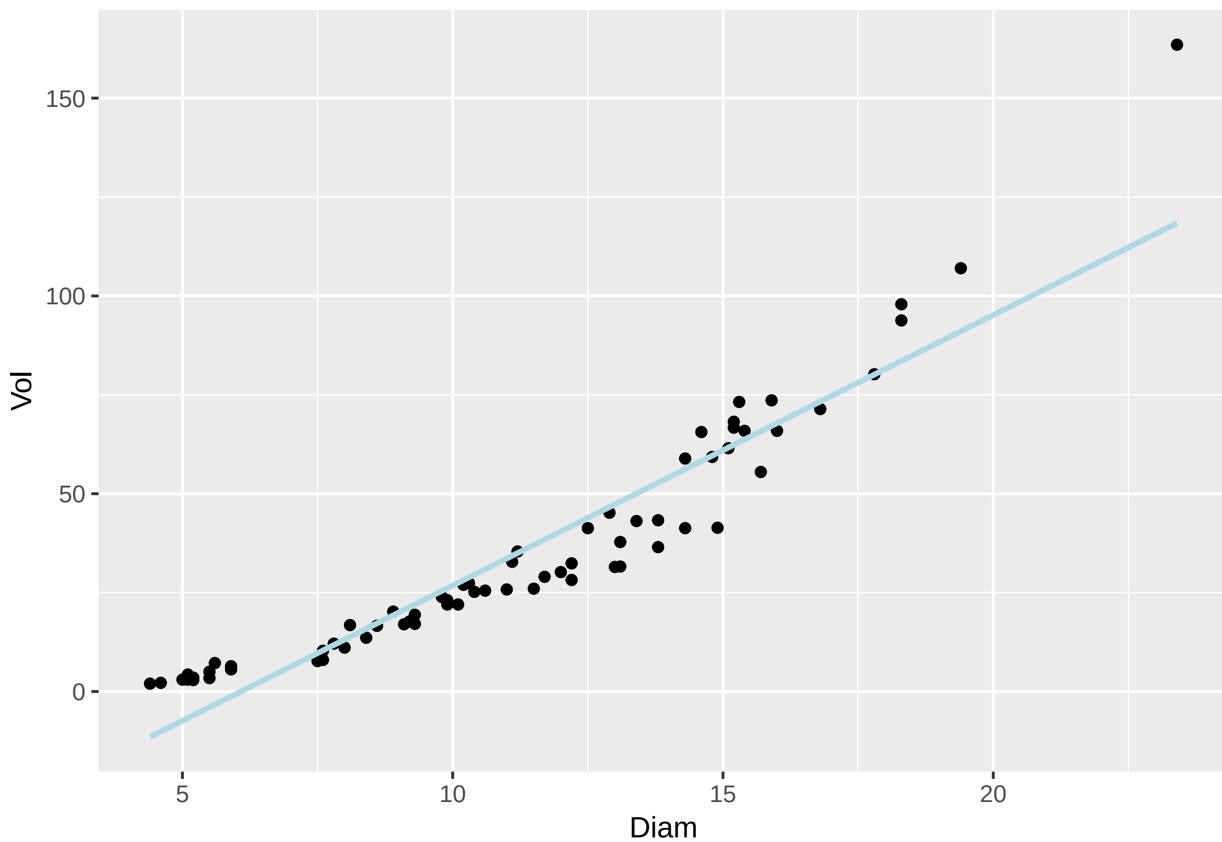
```
## # A tibble: 6 x 2
##    Diam   Vol
##   <dbl> <dbl>
```

```
## 1   4.4   2
## 2   4.6   2.2
## 3   5     3
## 4   5.1   4.3
## 5   5.1   3
## 6   5.2   2.9
```

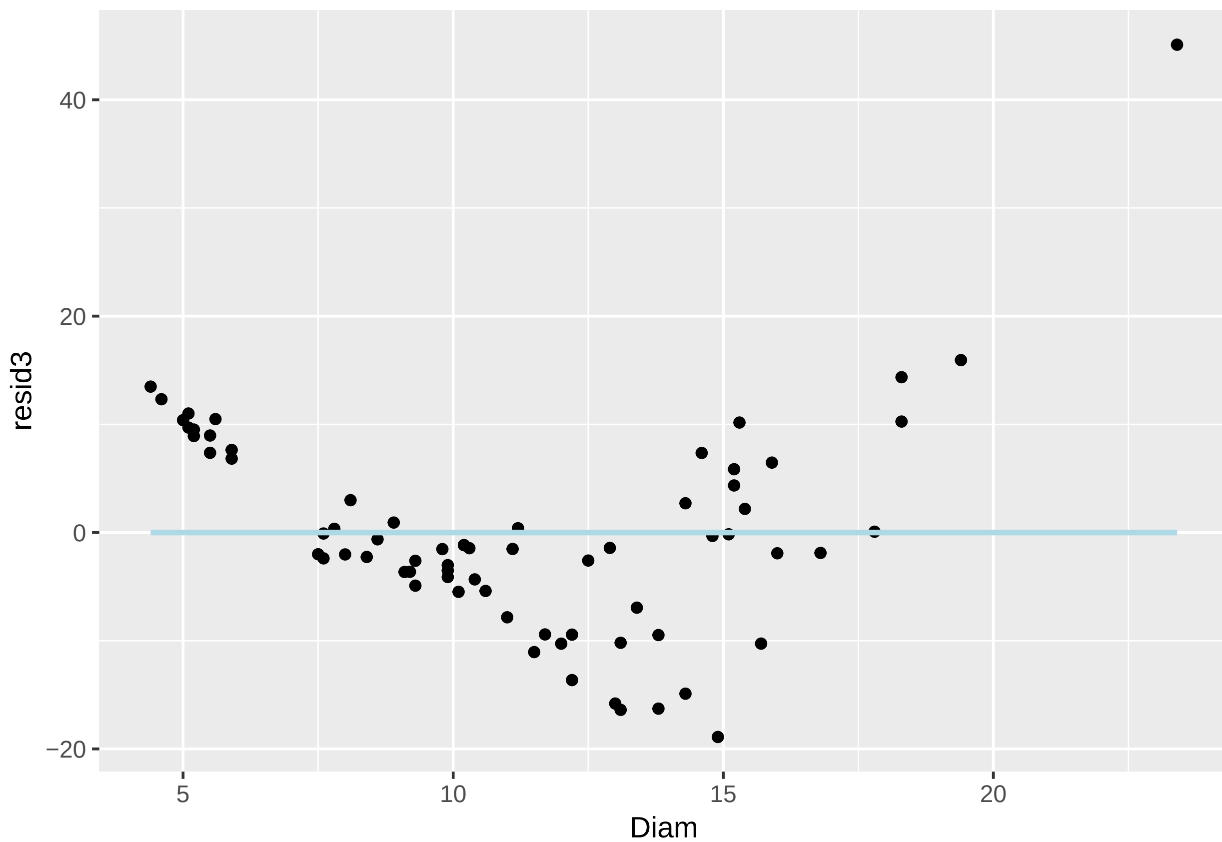
```r
# Modelo Linear:

# Regressão:
reg2 <- lm(Vol ~ Diam, data = ShortLeaf)
summary(reg2)
```

```
##
## 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
```

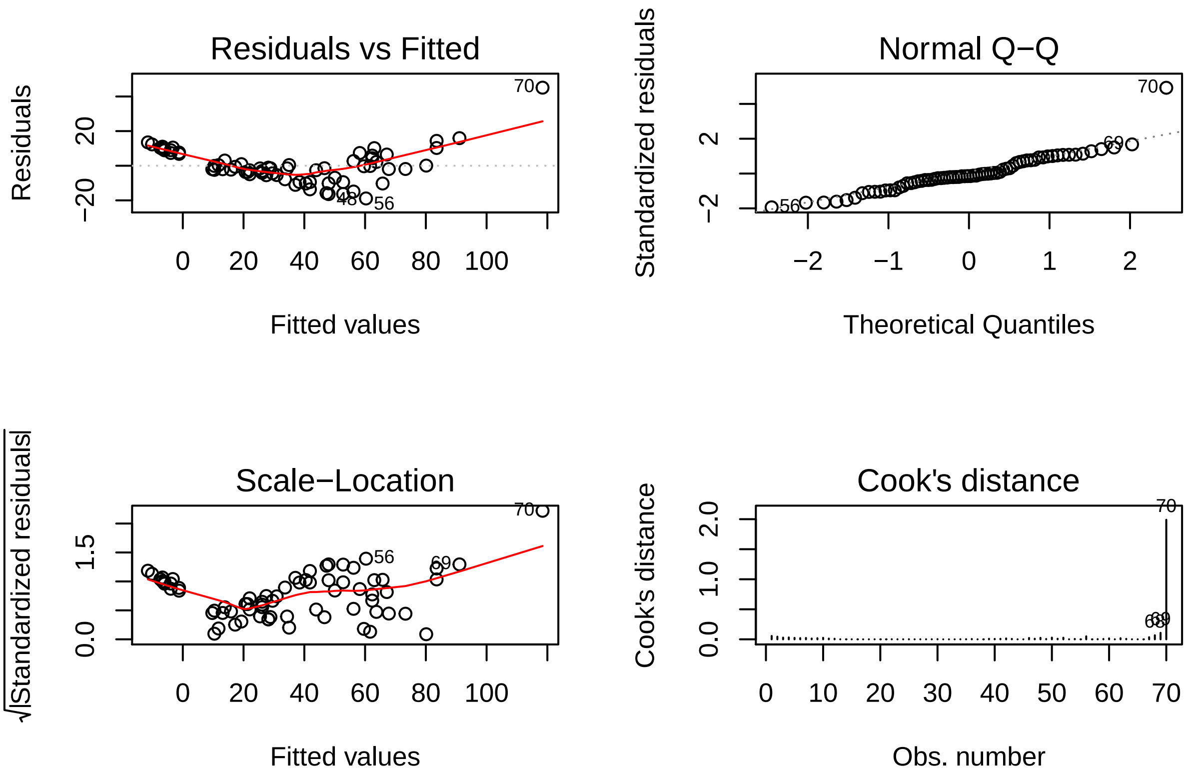
```r
# Plotar:
ggplot(data = ShortLeaf, aes(y = Vol, x = Diam))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```
# Analisar resíduos:
# Opção 1:
resid3 <- resid(reg2)
ggplot(data = ShortLeaf, aes(y = resid3, x = Diam)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg2, which = 1:4)
```



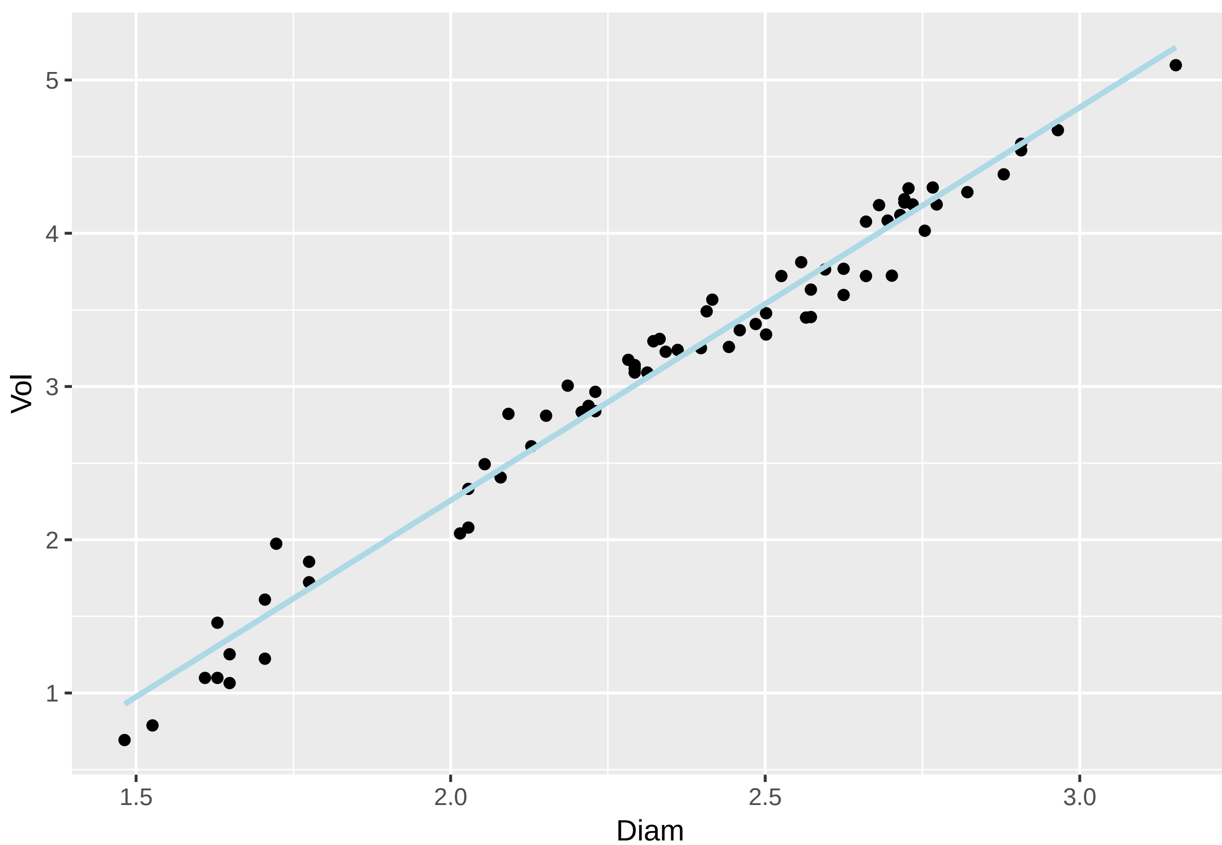
A VI é responsável por 0.89 da variação na VD. Neste caso, a distribuição não aproxima-se de uma normal, há uma relação não linear entre as variáveis e o erro padrão dos resíduos não apresentam a mesma variância (o RSE apresenta um valor relativamente alto, de 9.87). O modelo adequado para este caso é o log-log, que transforma em log tanto X, quanto Y.

```
# Modelo log-log:

# Transformar dados:
ShortLeaf2 <- log(ShortLeaf)

# Regressão:
reg.log2 <- lm(Vol ~ Diam, data = ShortLeaf2)
summary(reg.log2)
```

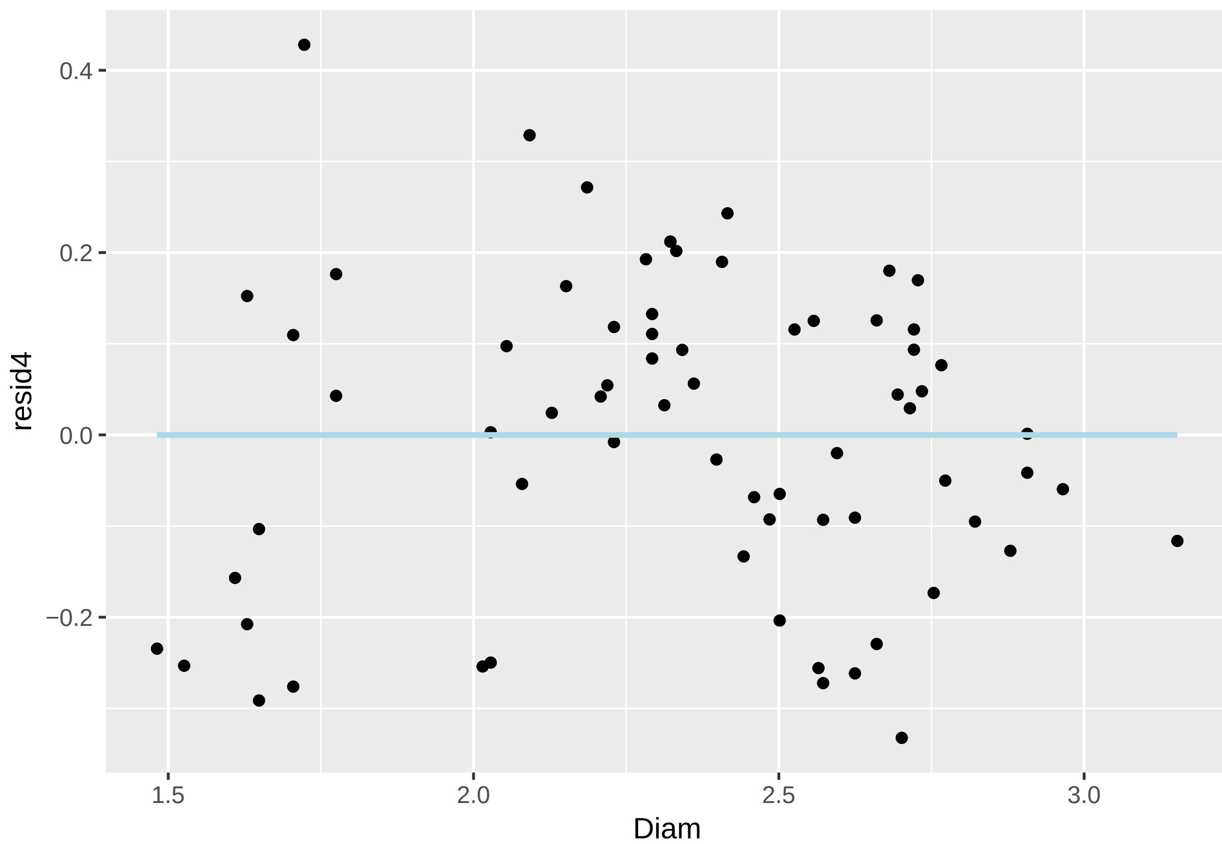
```
##
## Call:
## lm(formula = Vol ~ Diam, data = ShortLeaf2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3323 -0.1131  0.0267  0.1177  0.4280
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.8718     0.1216  -23.63   <2e-16 ***
```

```
## Diam          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
```

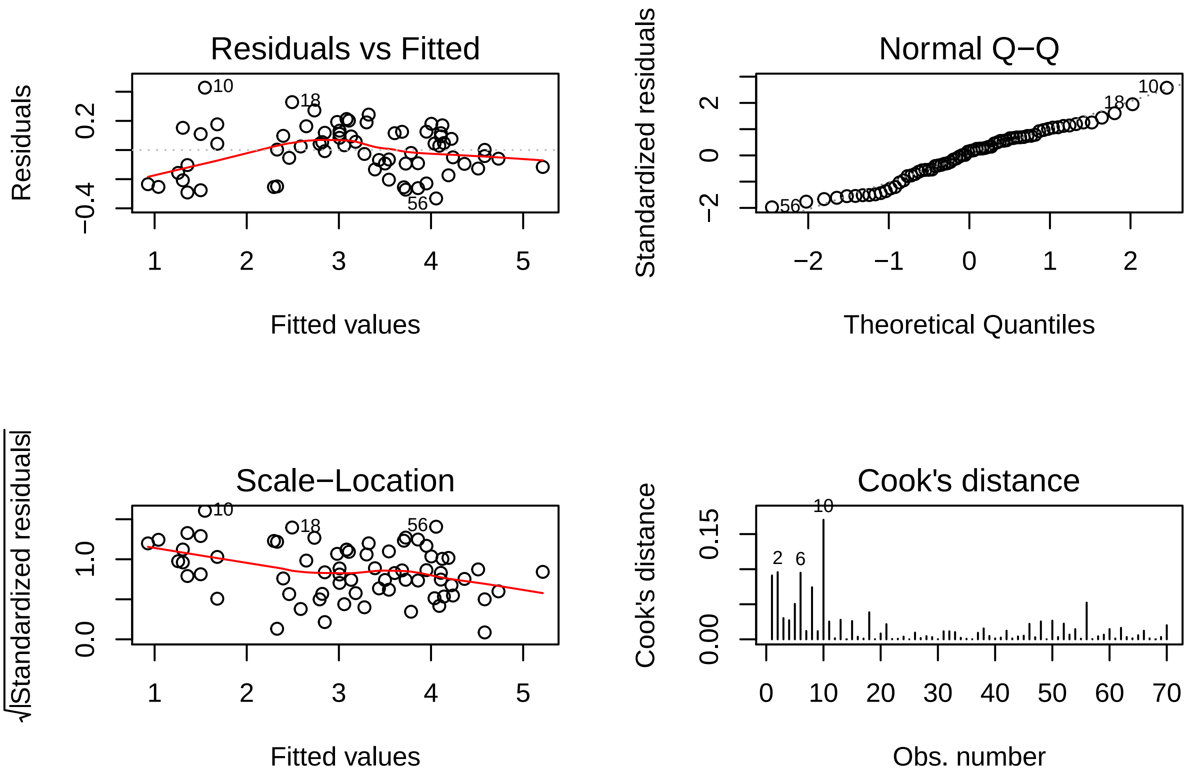
```r
# Plotar:
ggplot(data = ShortLeaf2, aes(y = Vol, x = Diam))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Analisar resíduos:
# Opção 1:
resid4 <- resid(reg.log2)
ggplot(data = ShortLeaf2, aes(y = resid4, x = Diam)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg.log2, which = 1:4)
```



Neste modelo log-log, a VI é reponsável por 0.97 da variação na VD. Ao plotar a regressão e a análise dos resíduos é possível observar uma relação linear entre as variáveis e os resíduos parecem estar distribuídos aleatoriamente (se não existe um padrão, é possível assumir que as variâncias serão aproximadamente iguais). Também, a distribuição dos resíduos parece aproximar-se mais de uma normal, quando comparada à distribuição do modelo anterior.

letra c)

```
# Modelo linear:

# Carregar base:
setwd("C:/Users/Duda/Desktop/PPGCP/Análise de Dados/lista_09")
MammGest <- read_table("mammgest.txt")
```

```
## Parsed with column specification:
## cols(
##   Row = col_double(),
##   Mammal = col_character(),
##   Birthwgt = col_double(),
##   Gestation = col_double()
## )
```

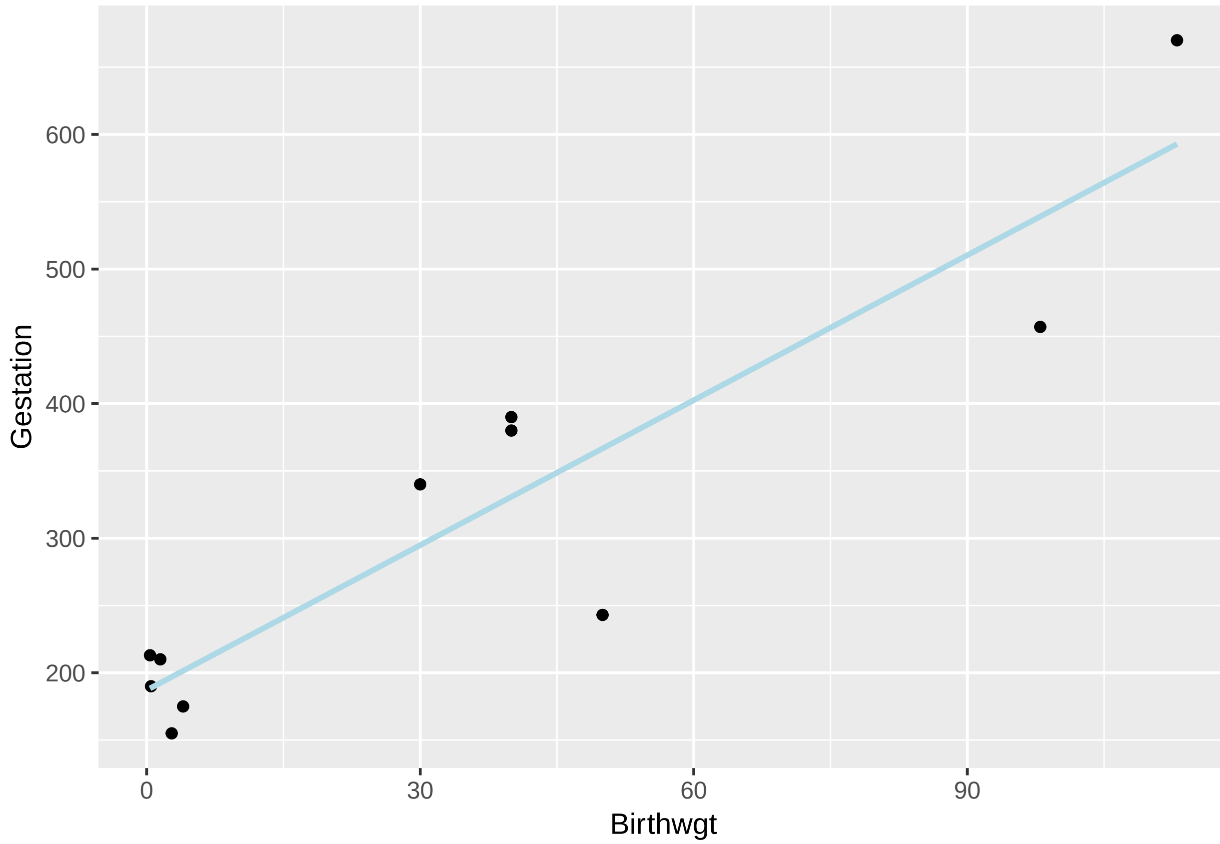
```r
head(MammGest)
```

```
## # A tibble: 6 x 4
##     Row Mammal    Birthwgt Gestation
##   <dbl> <chr>        <dbl>     <dbl>
## 1     1 Goat          2.75       155
## 2     2 Sheep         4          175
## 3     3 Deer          0.48       190
## 4     4 Porcupine     1.5        210
## 5     5 Bear          0.37       213
## 6     6 Hippo        50          243
```

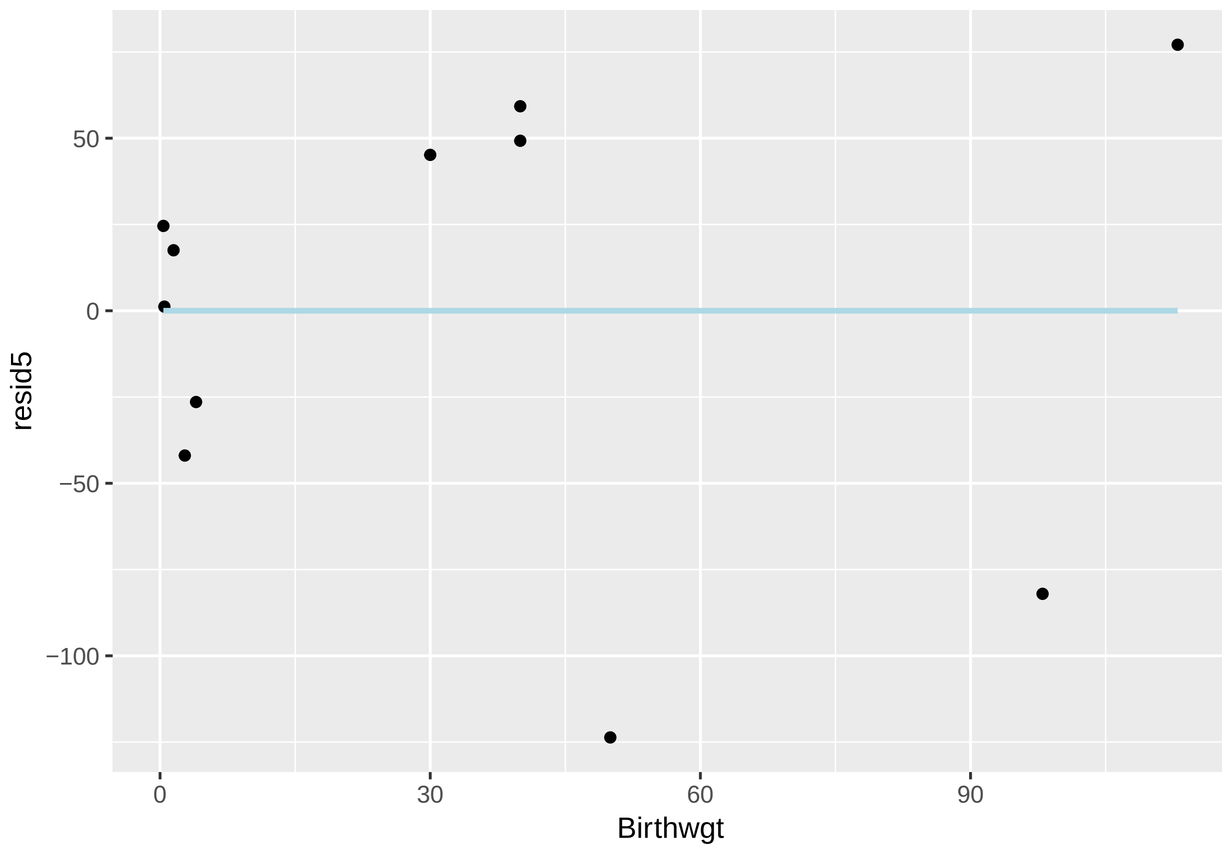
```r
# Regressão:
reg3 <- lm(Gestation ~ Birthwgt, data = MammGest)
summary(reg3)
```

```
##
## Call:
## lm(formula = Gestation ~ Birthwgt, data = MammGest)
##
## 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 ***
## Birthwgt      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
```

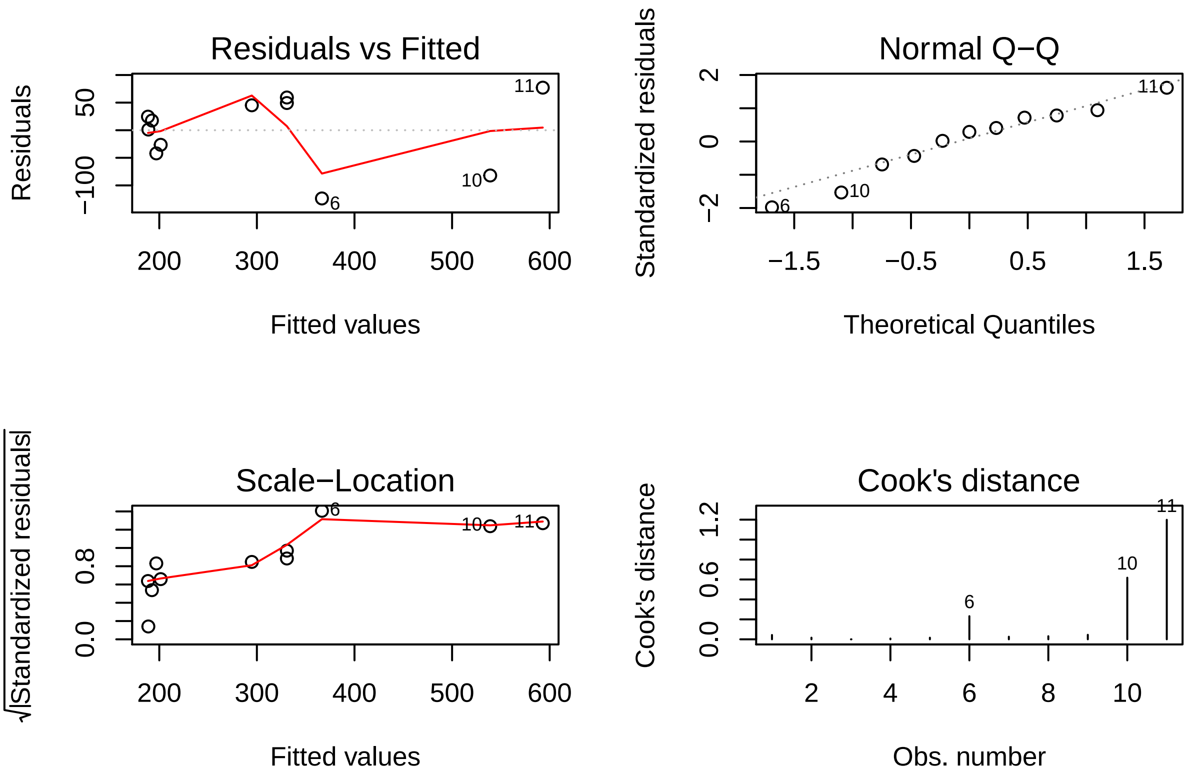
```r
# Plotar:
ggplot(data = MammGest, aes(y = Gestation, x = Birthwgt))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Analisar resíduos:
# Opção 1:
resid5 <- resid(reg3)
ggplot(data = MammGest, aes(y = resid5, x = Birthwgt)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg3, which = 1:4)
```



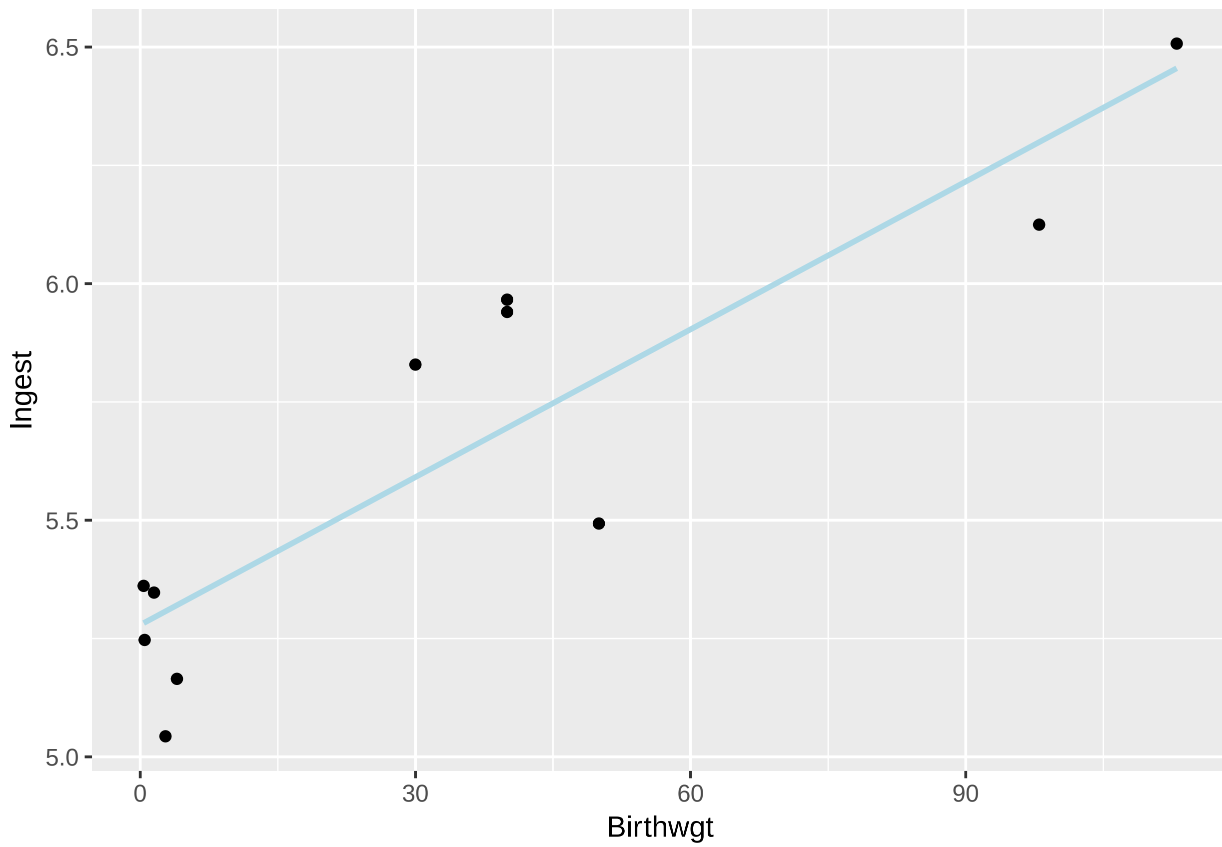
Ao analisar os gráficos, é possível perceber que neste modelo há uma relação linear entre as variáveis e que 0.84 da variação da VD, deve-se ao efeito da VI. Também é possível observar que a distribuição dos resíduos aproxima-se de uma distribuição normal; no entanto, a variância dos resíduos não é igual.Para desenvolver um modelo melhor ajustado, é necessário transformar Y em log.

```r
# Modelo log-level:

# Transformar dados e criar nova base:
lngest <- log(MammGest$Gestation)
MammGest2 <- mutate(MammGest, lngest = lngest)

# Regressão
reg.log3 <- lm(lngest ~ Birthwgt, data = MammGest2)
summary(reg.log3)
```

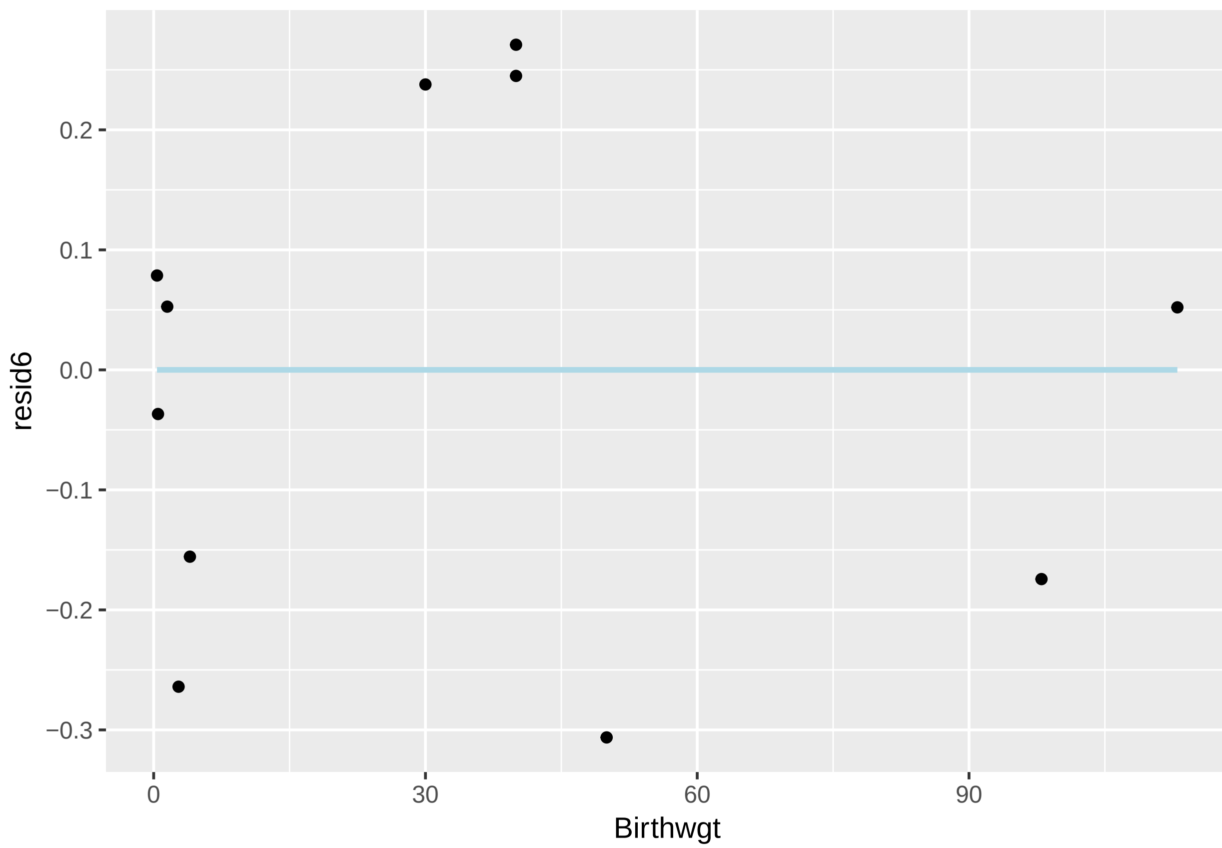
```
##
## Call:
## lm(formula = lngest ~ Birthwgt, data = MammGest2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3063 -0.1650  0.0521  0.1582  0.2709
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.278817   0.088177  59.866  5.1e-13 ***
## Birthwgt    0.010410   0.001717   6.062 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2163 on 9 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.7814
## F-statistic: 36.75 on 1 and 9 DF,  p-value: 0.0001878
```

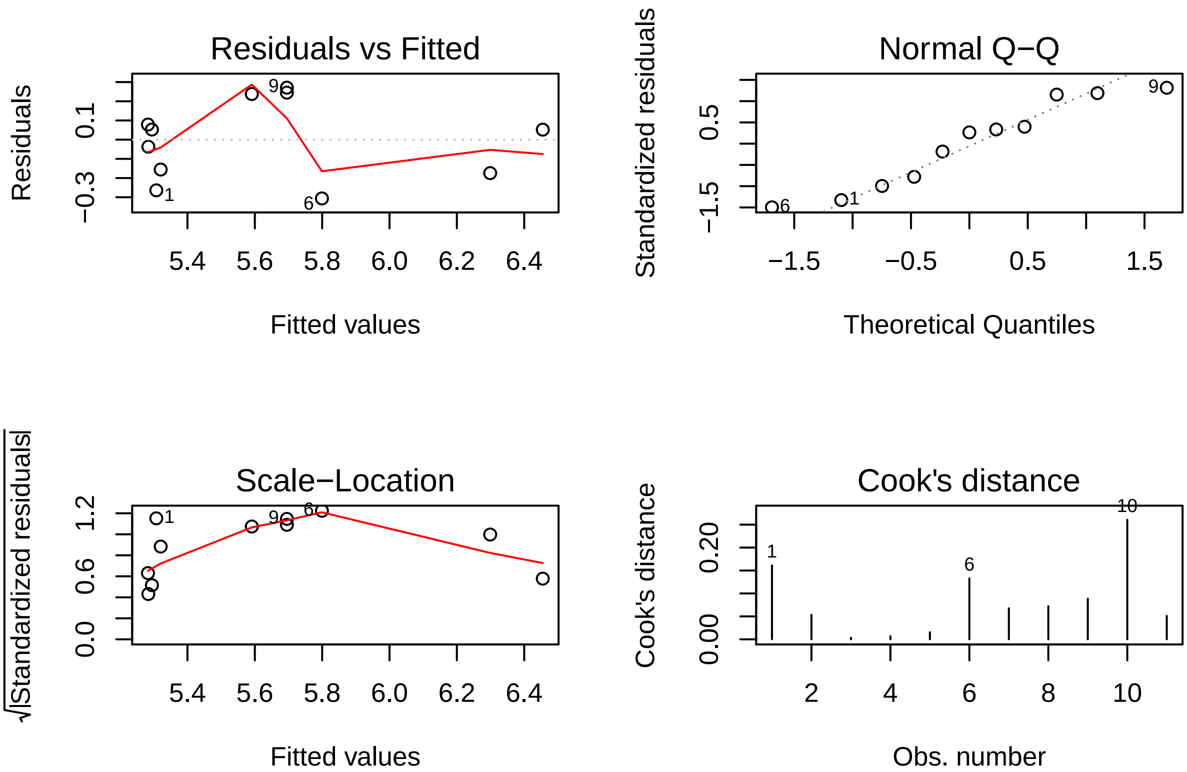
```r
# Plotar:
ggplot(data = MammGest2, aes(y = lngest, x = Birthwgt))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Analisar resíduos:
# Opção 1:
resid6 <- resid(reg.log3)
ggplot(data = MammGest2, aes(y = resid6, x = Birthwgt)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg.log3, which = 1:4)
```



Neste modelo, apesar de o R^2 apresentar menor valor do aquele encontrado no modelo linear, o pressuposto da igual variância dos resíduos é atendido, portanto, para este caso, o modelo log-level se mostra mais adequado.

4.2

letra a)

Os modelos polinomiais buscam representar relações não-lineares entre variáveis. Para estes modelos, são adicionadas novas variáveis "dummys" que adicionam uma "nova dimensão" (ou seja, são necessariamente modelos multivariados) à representação da regressão e faz com que exista um melhor ajuste da linha aos dados. No modelo polinomial a variável adicionada é um termo polinomial da própria variável independente.

letra b)

```
# Carregar base:
setwd("C:/Users/Duda/Desktop/PPGCP/Análise de Dados/lista_09")
BlueGills <- read_tsv("bluegills.txt")
```

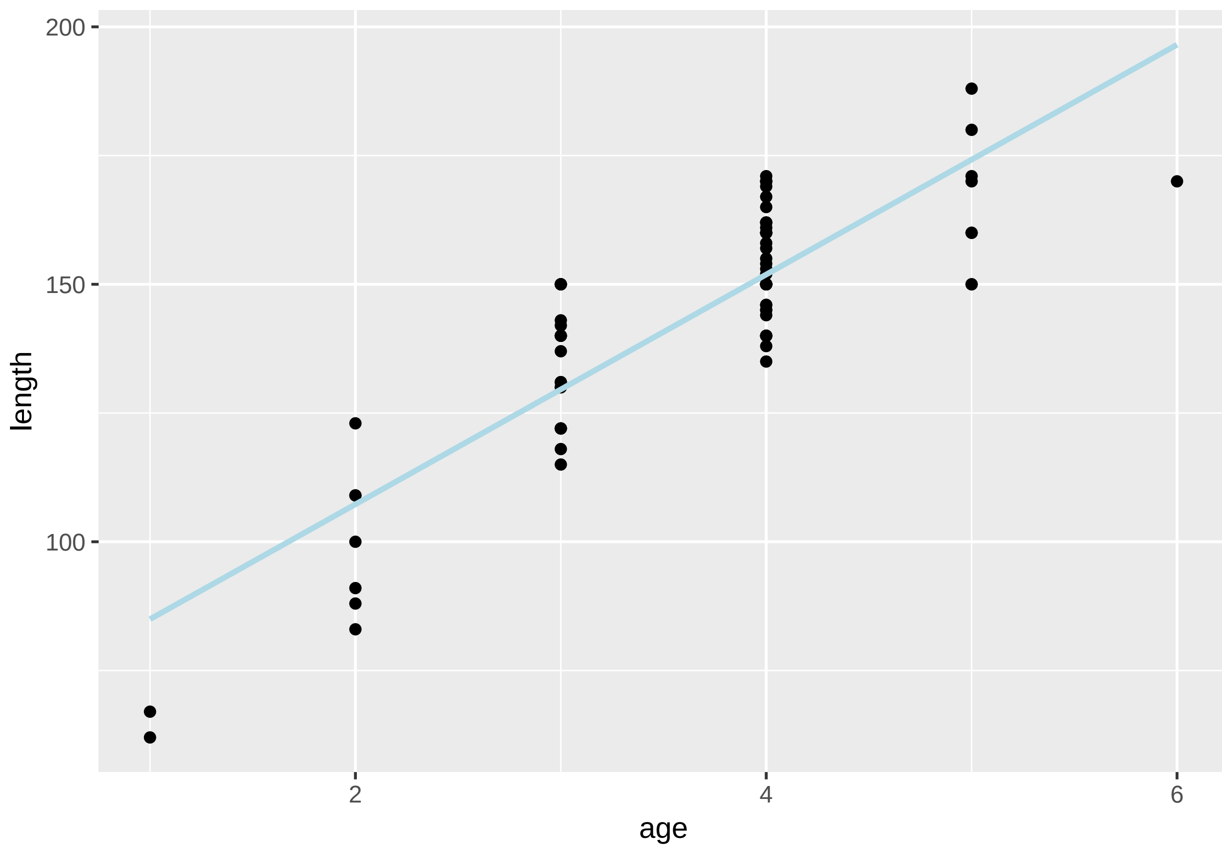
```r
head(BlueGills)
```

```
## # A tibble: 6 x 2
##     age length
##   <dbl>  <dbl>
## 1     1     67
## 2     1     62
## 3     2    109
## 4     2     83
## 5     2     91
## 6     2     88
```

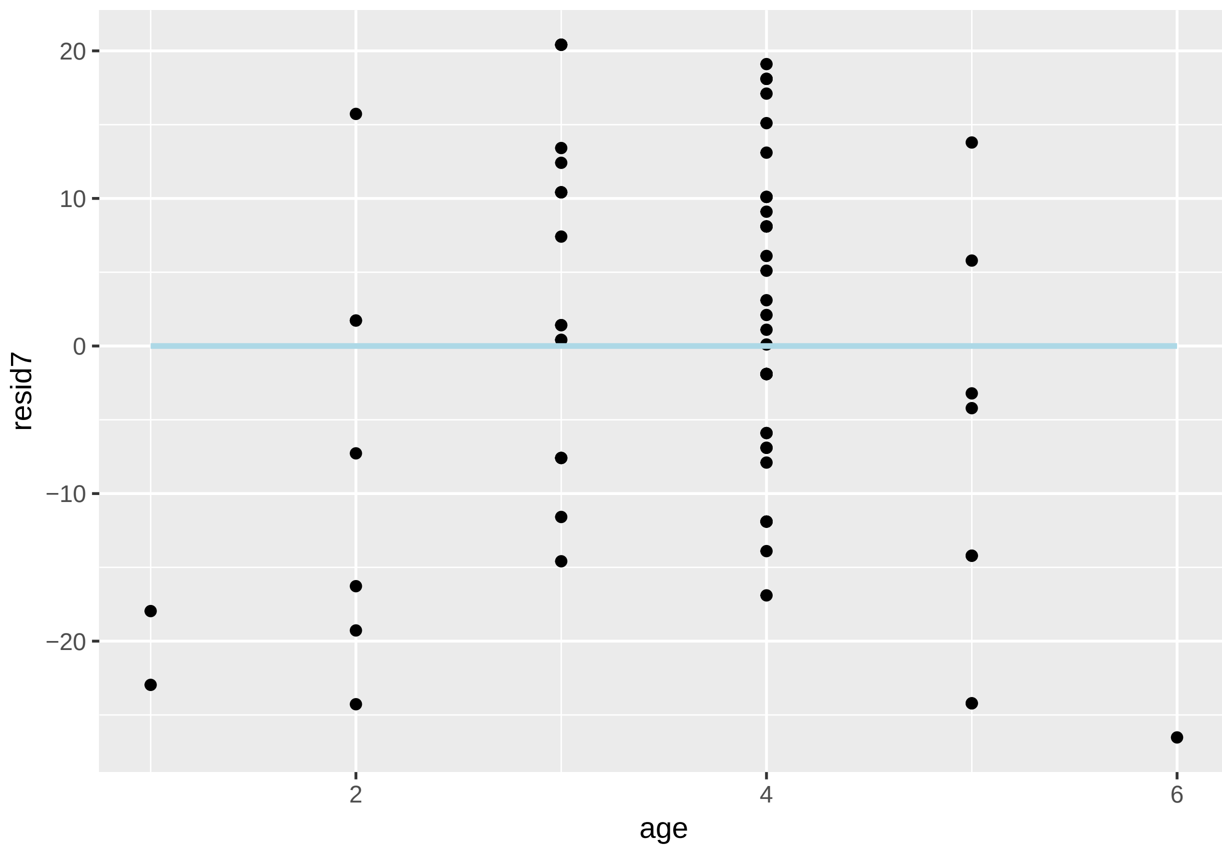
```r
# Modelo Linear:
# Regressão:
reg4 <- lm(length ~ age, data = BlueGills)
summary(reg4)
```

```
##
## 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
```

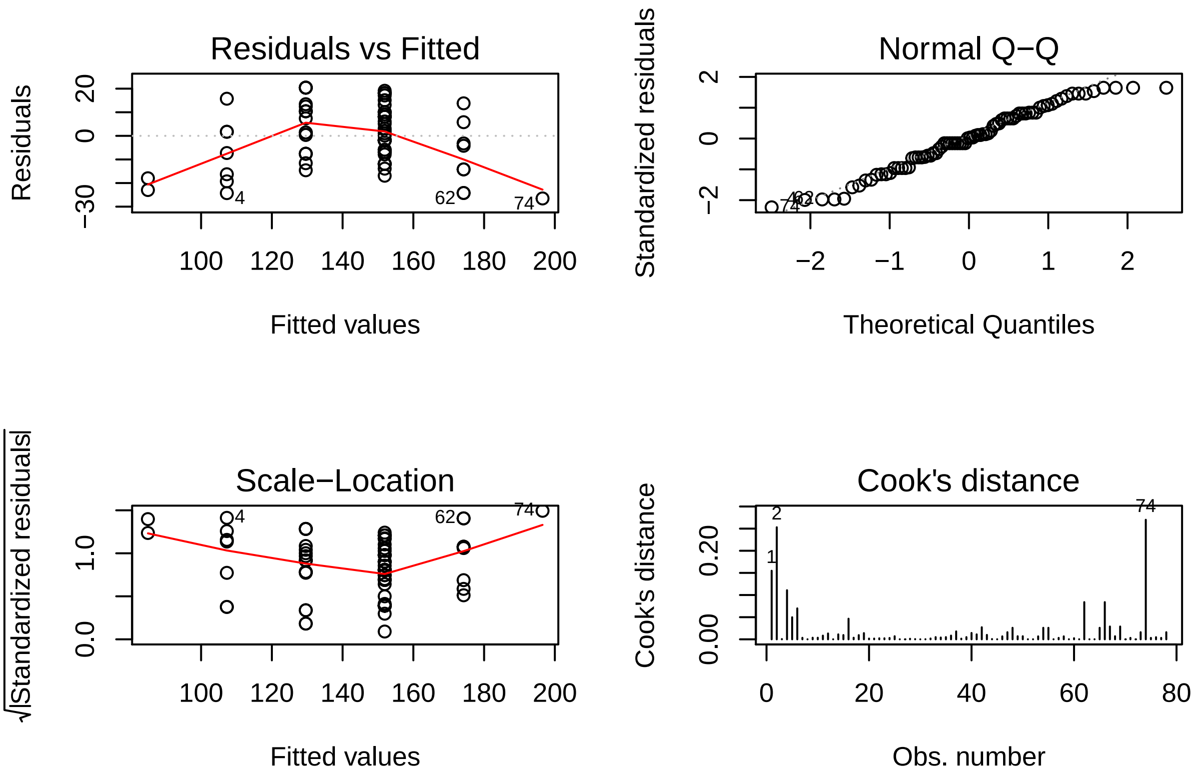
```r
# Plotar:
ggplot(data = BlueGills, aes(y = length, x = age))+
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Analisar resíduos:
# Opção 1:
resid7 <- resid(reg4)
ggplot(data = BlueGills, aes(y = resid7, x = age)) +
  geom_point()+
  geom_smooth(method = "lm", color = "lightblue", se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg4, which = 1:4)
```



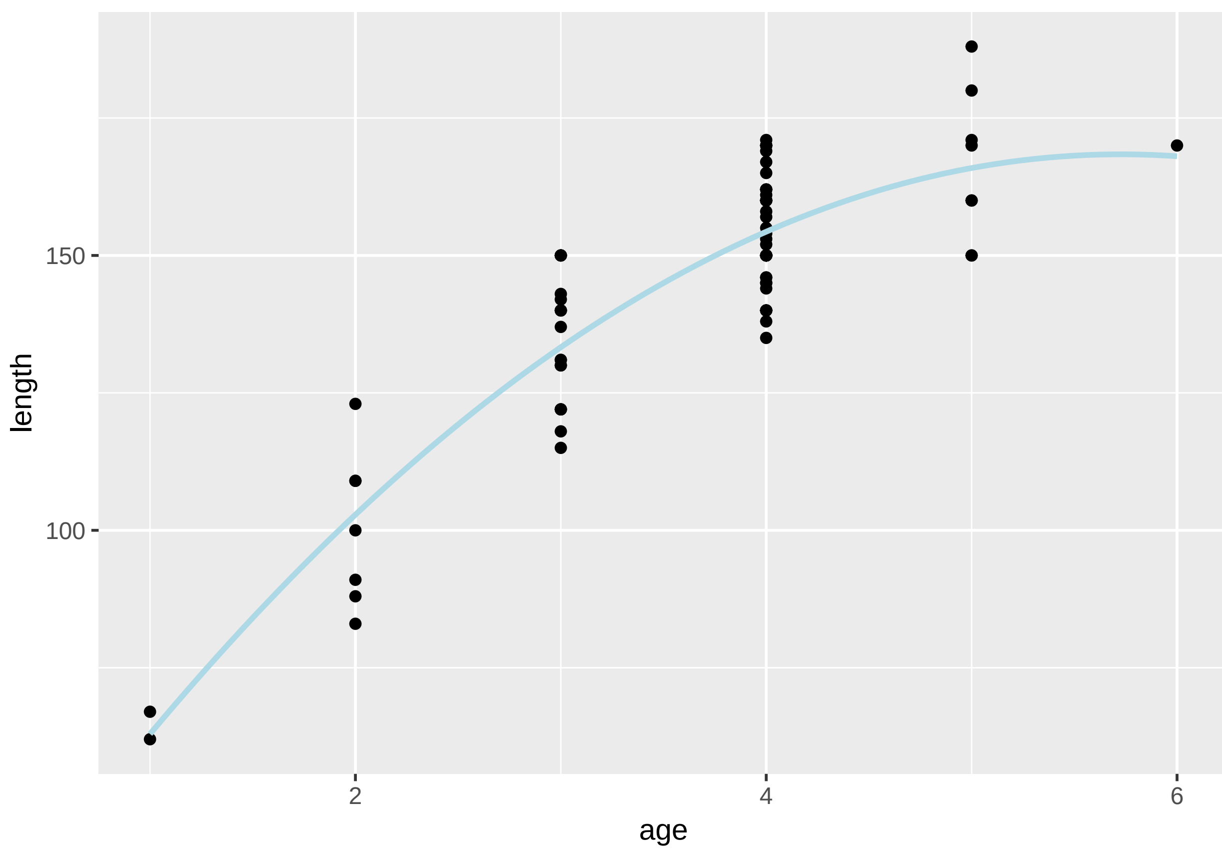
Este modelo, apesar de apresentar um R^2 de 0.73, não parece ser o mais adequado, já que os os resíduos estão muito dispersos e sugerem uma relação não linear entre as variáveis, como é possível verificar pelo valor de 12.51 do RSE e pela análise dos gráficos residuais.

```r
# Modelo quadrático:

# Transformar dados (adicionar termo quadrático):
BlueGills2 <- mutate(BlueGills, age2 = BlueGills$age^2)

# Regressão quadrática:
reg5 <- lm(length ~ age + age2 , data = BlueGills2)
summary(reg5)
```

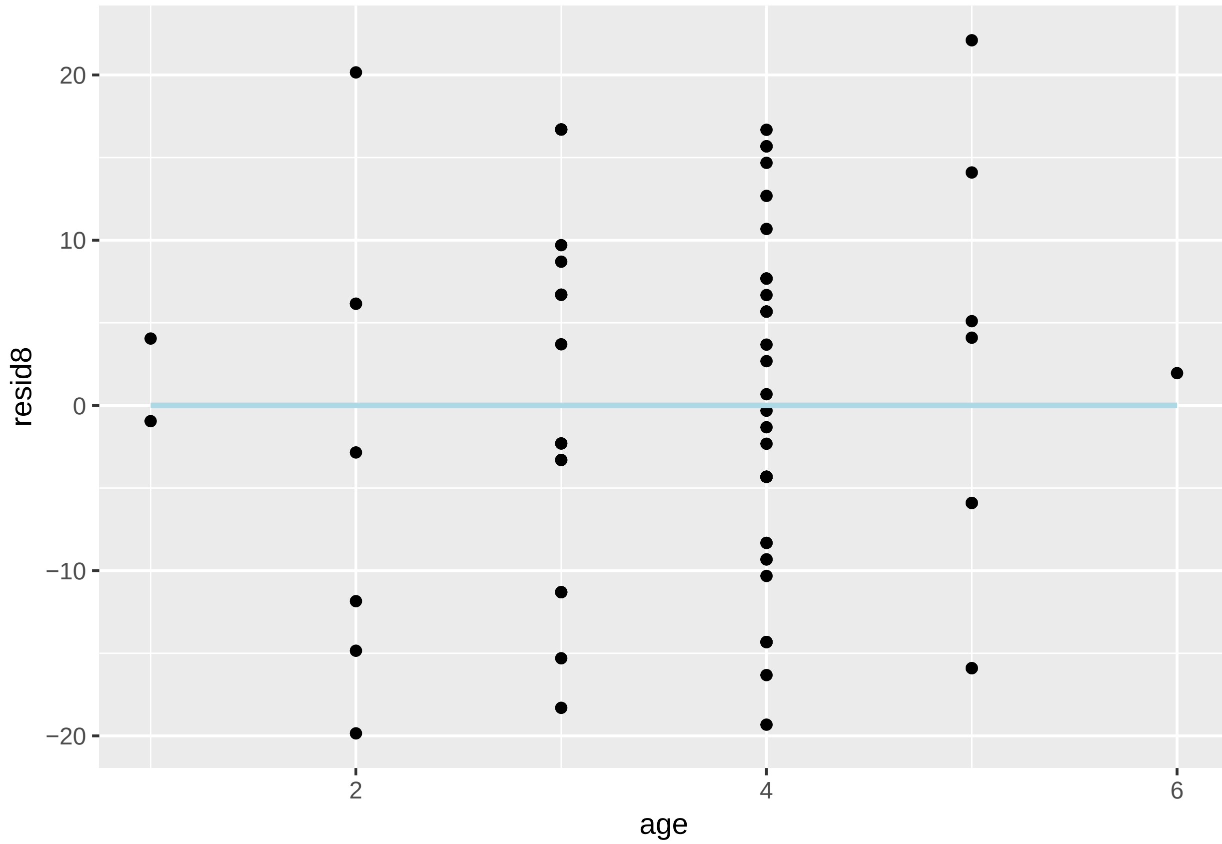
```
##
## Call:
## lm(formula = length ~ age + age2, data = BlueGills2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.846  -8.321  -1.137   6.698  22.098
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.622     11.016   1.237     0.22
## age           54.049      6.489   8.330 2.81e-12 ***
```

```
## age2          -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
```

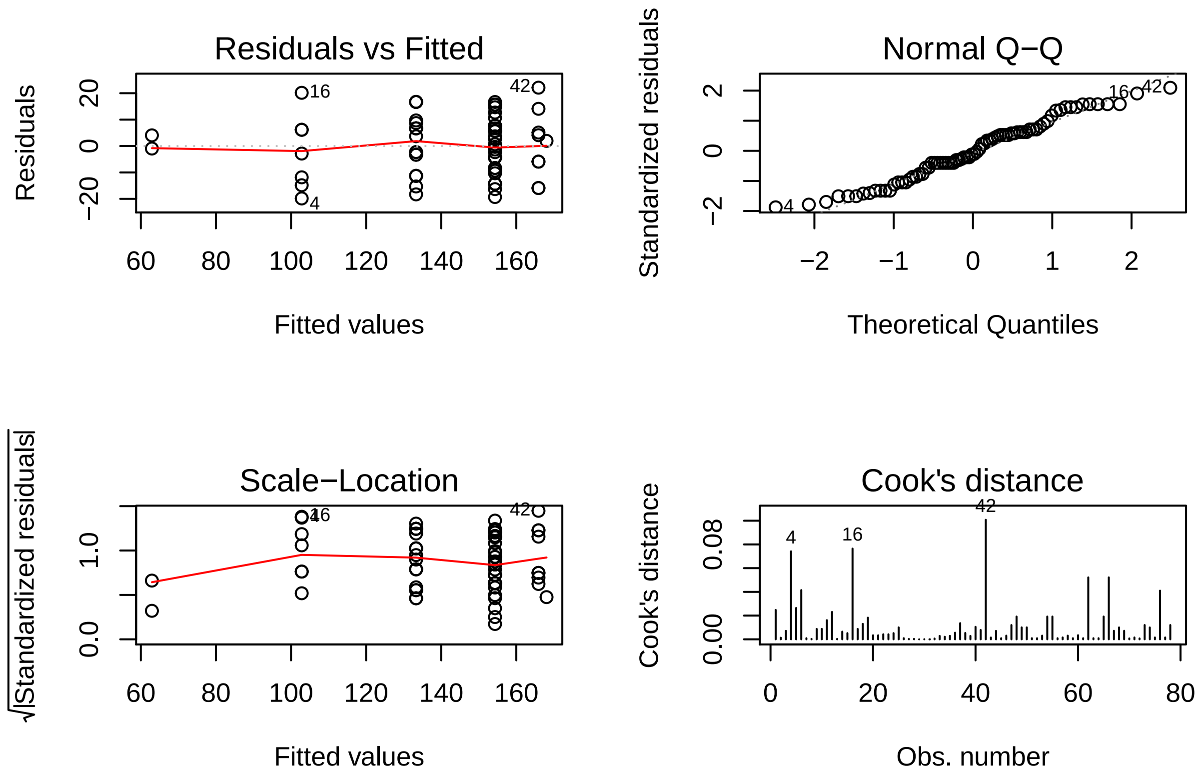
```
# Plotar:
ggplot(data = BlueGills2, aes(y = length, x = age))+
  geom_point()+
  stat_smooth(method = "lm", formula = y ~ poly(x,2), color = "lightblue",
              se = F)
```




```
# Analisar resíduos:
# Opção 1:
resid8 <- resid(reg5)
ggplot(data = BlueGills2, aes(y = resid8, x = age)) +
  geom_point()+
  stat_smooth(method = "lm", formula = y ~ poly(x,2), color = "lightblue",
              se = F)
```




```r
# Opção 2:
par(mfrow=c(2,2))
plot(reg5, which = 1:4)
```



Este modelo apresenta um melhor ajuste, com o R^2 = 0.80 e a análise dos gráficos dos resíduos permite observar que estes estão melhor distribuídos em relação à sua média.

letra c)

É possível observar que a relação entre a idade e o comprimento dos peixes da espécie Bluegill é positiva e não linear. Ao analisar o valor do R^2, é possível observar que aproximadamente 80% da variação no comprimentos destes peixes é explicada pela idade.