Swiss

Grupo: Americo Freitas, Arleks dos Santos e Luciano Ozorio

2022-05-08

Descrição do conjunto de dados :

Medidas de fecundidade padronizadas e indicadores socioeconômicos para cada uma das 47 províncias francófonas da Suíça por volta de 1888.

Formato do conjunto de dados:

** Um quadro de dados com 47 observações em 6 variáveis, cada uma delas em porcentagem, ou seja, em [0, 100]. **

- [,1] Fertility- Ig, 'medida comum de fertilidade padronizada'
- [,2] Agriculture- % de homens envolvidos na agricultura como ocupação
- [,3] Examination- % de recrutas que receberam a nota mais
- [,3] Examination- % de recrutas que receberam a nota mais alta no exame do exército
- [,4] Education- % de educação além da escola primária para recrutas.
- [,5] Catholic- % 'católico' (em oposição a 'protestante').
- [,6] Infant.Mortality- nascidos vivos que vivem menos de 1 ano.

** Todas as variáveis, exceto 'Fertilidade', fornecem proporções da população. **

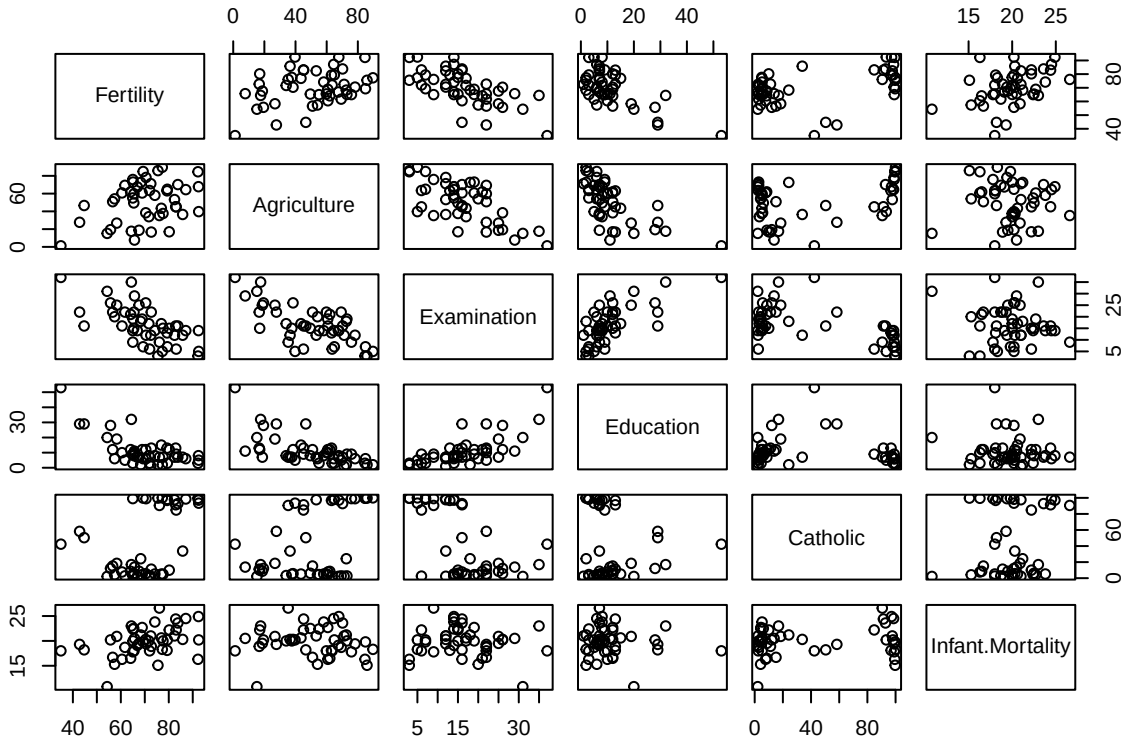
```
data(swiss)
summary(swiss)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

Gráfico de correlação das variáveis

- O gráfico mostra a relação linear entre Agriculture e Examination.
- Além disso, entre Examination e Education.
- A interpretação dos coeficientes será afetada.

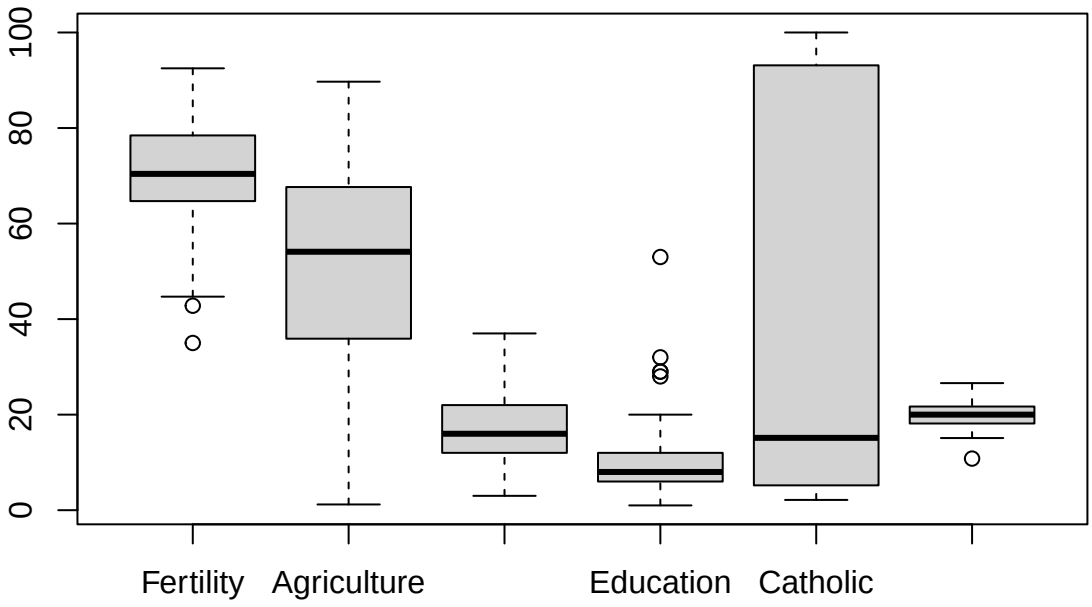
```
pairs(swiss)
```



box-plot analysis

- Catholic variável cobre uma ampla gama de valores
- Infant.Mortality variável é muito condensada
- Education parece ter alguns outliers

```
boxplot(swiss)
```



Correlação

Observamos que não há problema de Multicolinearidade, pois não há correlação maior ou igual a 70% entre as covariáveis (variáveis explicativas) * Todas as correlações com Fertility são menores que 0,8, indicando que

não há sinais de forte multicolinearidade. * As correlações estão entre 0,3-0,8, indicando multicolinearidade leve.

```
cor(swiss)
```

```
##                  Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility             0.41655603
## Agriculture          -0.06085861
## Examination          -0.11402160
## Education            -0.09932185
## Catholic              0.17549591
## Infant.Mortality      1.00000000
```

Ajustando um Modelo Linear

O valor de p-value demonstra que ao menos uma variável é significativa. Vemos que exceto a variável Examination, todas as outras são significativas pois o valor de Pr(>|t|) é extremamente grande (0,31546 > 0,05). O valor de R2 ajustado é menor do que 70 % porém bem próximo (0,67).

```
modelo <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, swiss)
summary(modelo)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic + Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Vamos gerar um novo sumário retirando a variável Examination devido ao seu valor de Pr(>|t|)

Agora todas as variáveis são significativas (valor de p abaixo de 5% no teste t) O valor de R2 ou coeficiente de explicação da variável resposta pelas variváveis explicativas ainda é menor do que 70% e permaneceu em

0,6707 , próximo de 70% mas ainda moderado.

```
modelo2 <- lm(Fertility ~ Agriculture +  Education + Catholic + Infant.Mortality, swiss)
summary(modelo2)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Perceba que o resultado é o mesmo (sugere retirar a variável Examination) utilizando.

```
modelo3=step(modelo, direction = "backward")
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

```
summary(modelo3)
```

```
##
```

```
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Transformando a variável resposta com log

O modelo passou para um nível alto de significância ou seja > 70% . O coeficente. de explicação (Adjusted R-squared) é de aproximadamente 71%, o que significa que as variáveis Agriculture, Education, Catholic e Infant.Mortality (retirada a variável Examination) explicam 71% das variações da variável resposta Fertility. Dessa forma, o modelo fica:

Fertility = 4.136 -0.002 *Agriculture - 0.017* Education + 0.001 * Catholic + 0.016 * Infant.Mortality

Houve tentativa de outros ajustes como log log, semi log, inverso e raiz-quadrada mas em todos o R quadrado ajustado permaneceu abaixo de 70 %, indicando um grau moderado de explicação da variável resposta pelas variáveis explicativas.

Permanecemos então com o ajuste log linear.

O teste F é usado para testar se a hipótese nula os verdadeiros coeficientes de inclinação são simultaneamente iguais a zero, indicando se há relação significante da variável dependente com o conjunto de variáveis independentes, ou seja a significância total do modelo i.e. se há algum B i diferente de 0.

O modelo está adequado globalmente pois p-value em 1.484e-11 (menor do que 5%) e o R quadrado ajustado ou coeficiente de explicação está com 71%;

O teste F soma o poder preditivo de todas as variáveis independentes e determina que é improvável que todos os coeficientes sejam iguais a zero. No entanto, é possível que cada variável não seja preditiva o suficiente para ser estatisticamente significativa. Em outras palavras, a amostra fornece evidências suficientes para concluir que seu modelo é significativo, mas não o suficiente para concluir que qualquer variável individual é significativa. Ao analisarmos o summary do novo modelo vemos que todas as variáveis explicativas são significativas com todas as variáveis explicativas com p abaixo de 5%(após a retirada da variável Examination):.

```
modelo2 <- lm(log(Fertility) ~ Agriculture +  Education + Catholic + Infant.Mortality, swiss)
summary(modelo2)
```

```
##
## Call:
## lm(formula = log(Fertility) ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24152 -0.08664  0.01589  0.05670  0.20559
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.1364914  0.1420897  29.112  < 2e-16 ***
## Agriculture      -0.0022781  0.0010088  -2.258 0.029183 *
## Education        -0.0168237  0.0021915  -7.677 1.59e-09 ***
## Catholic          0.0015495  0.0004274   3.625 0.000775 ***
## Infant.Mortality  0.0166930  0.0056491   2.955 0.005108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.106 on 42 degrees of freedom
## Multiple R-squared:  0.733,  Adjusted R-squared:  0.7076
## F-statistic: 28.83 on 4 and 42 DF,  p-value: 1.484e-11
```

Análise de resíduos para variável Agriculture

```
library(palmerpenguins)
library(tidyverse)
```

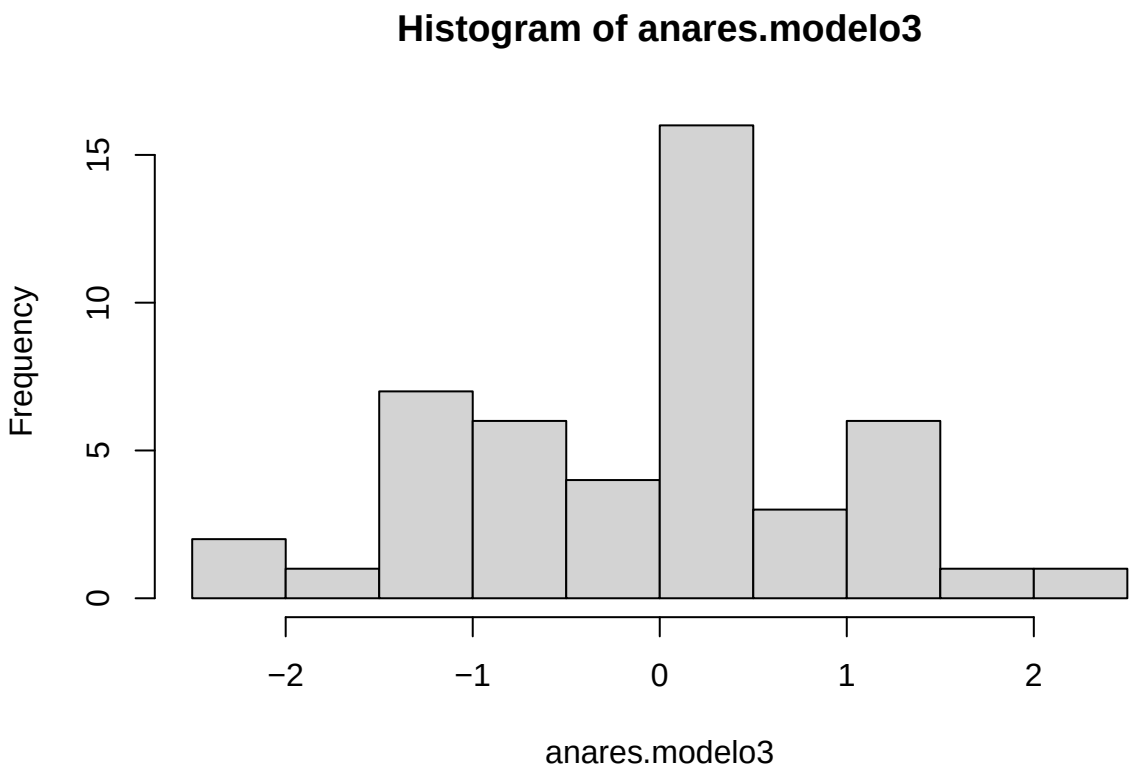
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

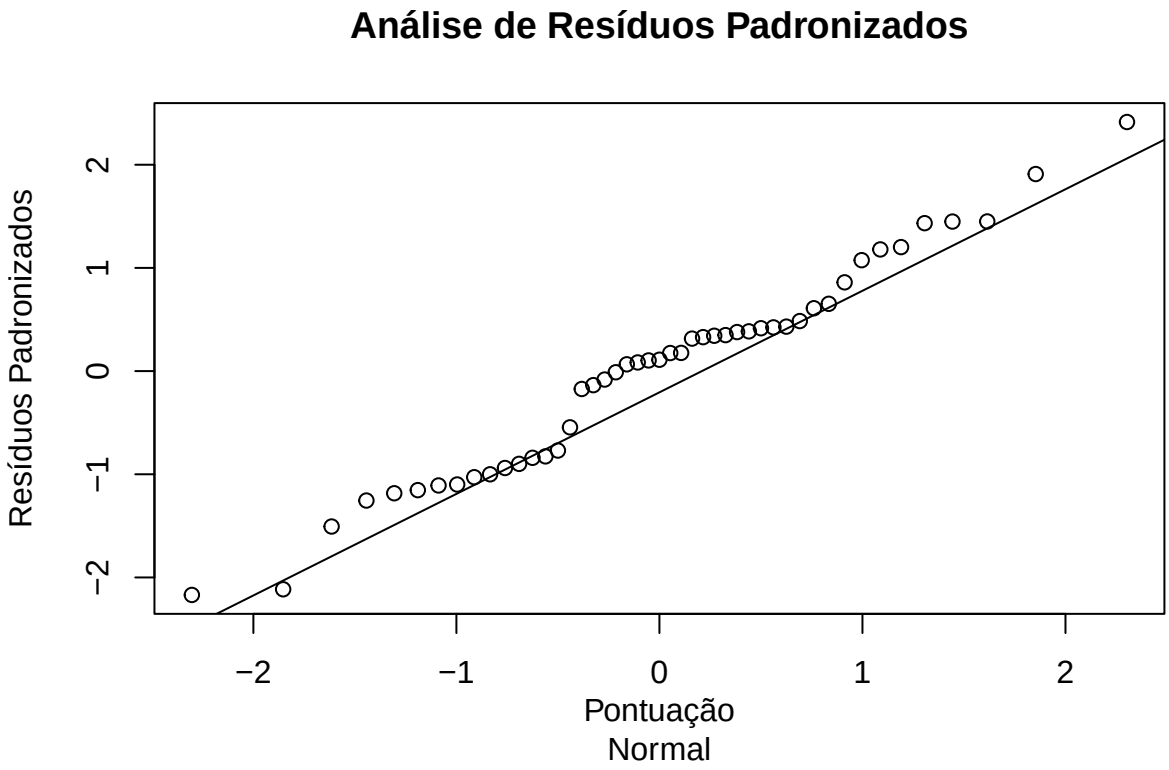
```
theme_set(theme_bw())
library(performance)
```

```
#check_model(modelo3, check = c("linearity", "qq", "homogeneity", "outliers"))
anares.modelo3 = rstandard(modelo3)
hist(anares.modelo3)
```



Com os resíduos em ordem vemos que o plot do qqnorm+qqline é aproximadamente uma reta de 45 graus portanto existe normalidade dos resíduos. No gráfico QQ, não vemos caudas pesadas, então a suposição de normalidade também é válida.

```
qqnorm(anares.modelo3, ylab="Resíduos Padronizados", xlab="Pontuação
Normal", main="Análise de Resíduos Padronizados")
qqline(anares.modelo3)
```

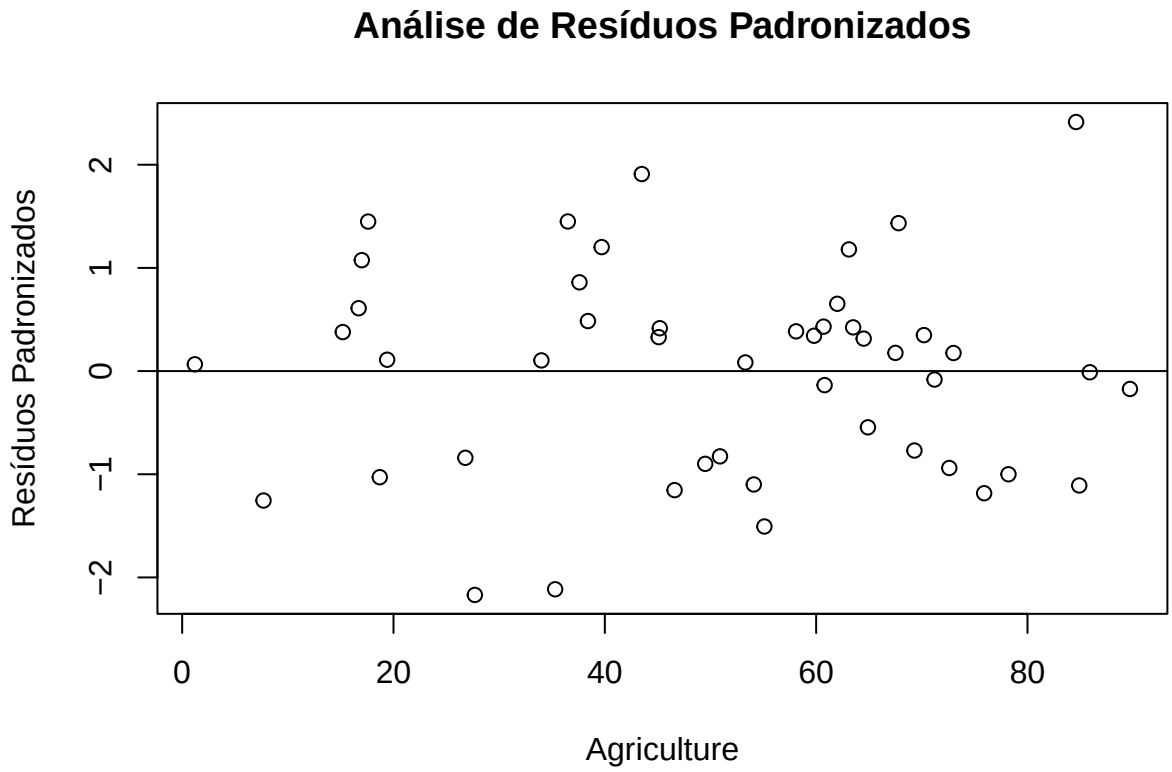


###

Análise de resíduos para variável Agriculture

Não se percebe nenhuma formação de padrão (distribuição aleatória dos resíduos) podendo-se presumir em normalidade dos resíduos.

```
plot(swiss$Agriculture, anares.modelo3, ylab="Resíduos Padronizados",
     xlab="Agriculture", main="Análise de Resíduos Padronizados") +
abline(0,0)
```

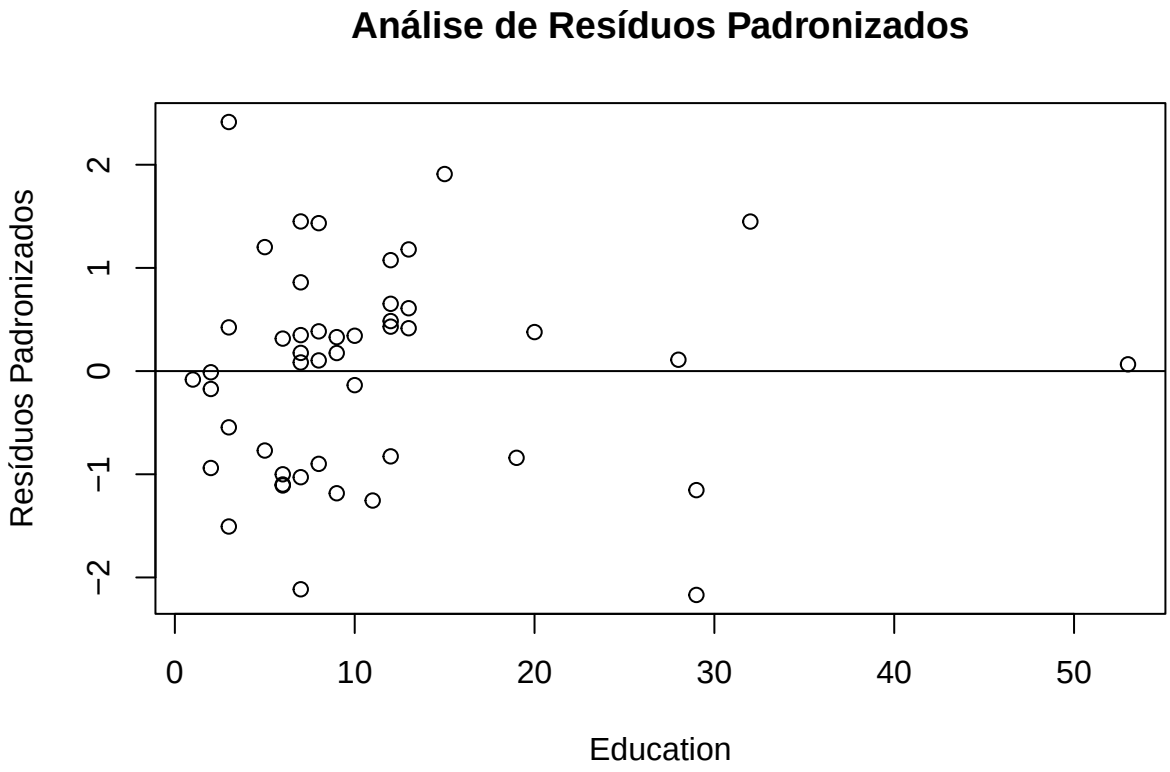



```
## integer(0)
```

Análise de resíduos para variável Education

Não se percebe nenhuma formação de padrão (distribuição aleatória dos resíduos) podendo-se presumir em normalidade dos resíduos.

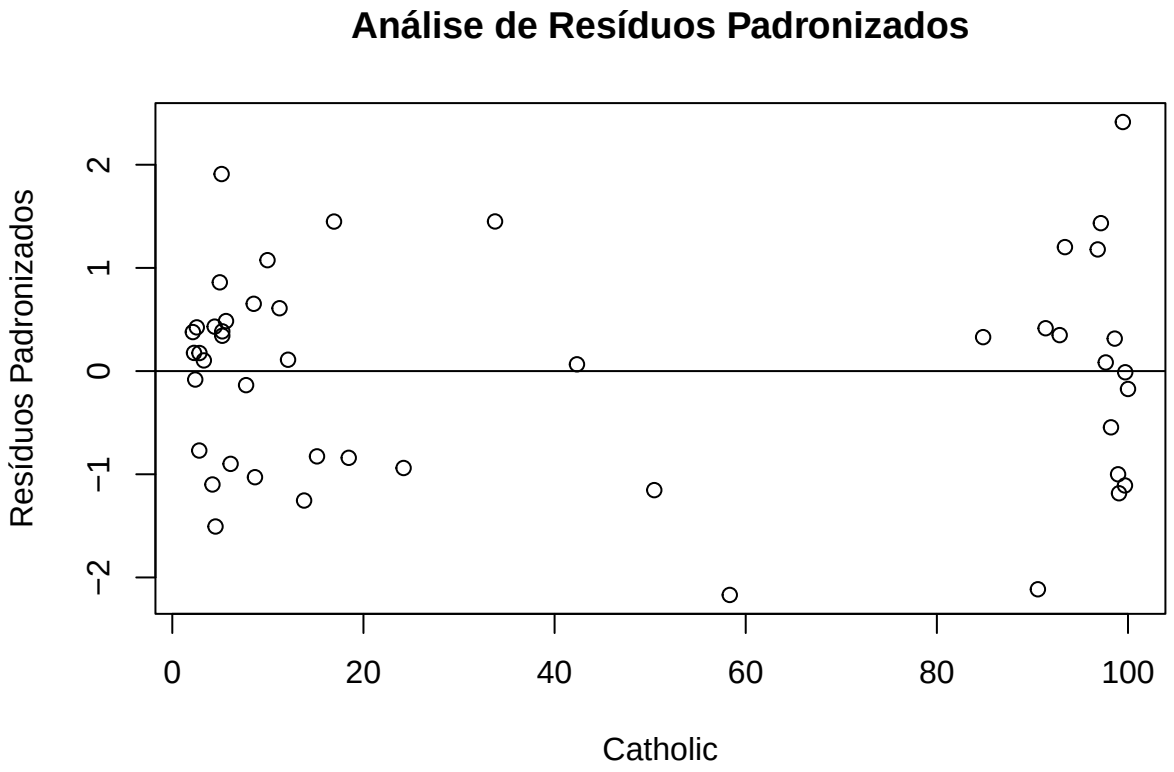
```
plot(swiss$Education, anares.modelo3, ylab="Resíduos Padronizados",
     xlab="Education", main="Análise de Resíduos Padronizados") +
abline(0,0)
```




```
## integer(0)
```

Não se percebe nenhuma formação de padrão (distribuição aleatória dos resíduos) podendo-se presumir em normalidade dos resíduos.

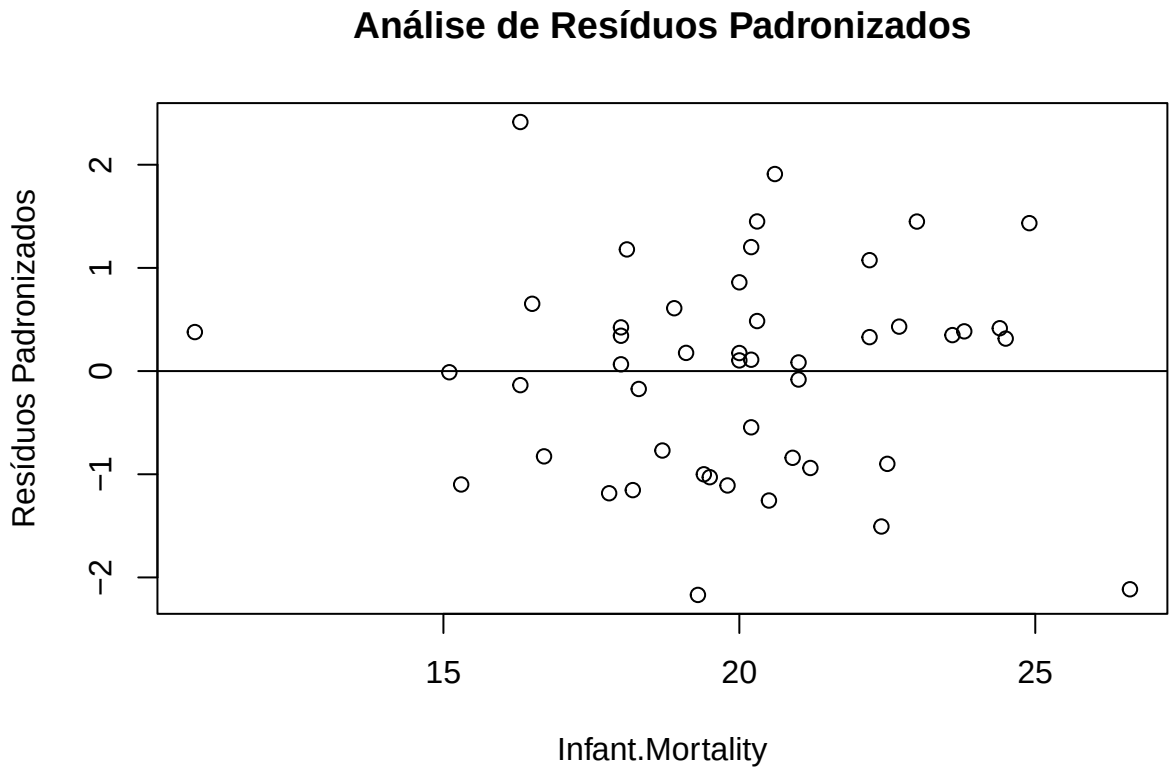
```
plot(swiss$Catholic, anares.modelo3, ylab="Resíduos Padronizados",
     xlab="Catholic", main="Análise de Resíduos Padronizados") +
abline(0,0)
```




```
## integer(0)
```

Não se percebe nenhuma formação de padrão (distribuição aleatória dos resíduos) podendo-se presumir em normalidade dos resíduos.

```
plot(swiss$Infant.Mortality, anares.modelo3, ylab="Resíduos Padronizados",
     xlab="Infant.Mortality", main="Análise de Resíduos Padronizados") +
abline(0,0)
```




```
## integer(0)
```

A partir do gráfico residual, não há padrão observado, portanto, a suposição de variância constante se mantém.

No gráfico QQ, não vemos caudas pesadas, então a suposição de normalidade também é válida.

partimos para os testes formais

Teste Anderson-Darling Normalidade dos Resíduos

O teste de Anderson-Darling indica que não podemos rejeitar a hipótese de normalidade dos resíduos pois o p-value é maior que 5%

```
library(nortest)
ad.test(anares.modelo3)
```

```
##
##  Anderson-Darling normality test
##
## data:  anares.modelo3
## A = 0.54251, p-value = 0.1552
```

Teste Shapiro-Wilk Normalidade dos Resíduos

O teste de Shapiro-Wilk, com p-value acima de 5%, também indica que não podemos rejeitar a hipótese de normalidade dos resíduos.

```
shapiro.test(anares.modelo3)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anares.modelo3
```

```
## W = 0.97663, p-value = 0.461
```

O teste de Breusch-Pagan, com p-value acima de 5%, indica que não podemos rejeitar a hipótese de homocedasticidade dos resíduos.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(modelo3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo3
## BP = 3.1629, df = 4, p-value = 0.5309
```

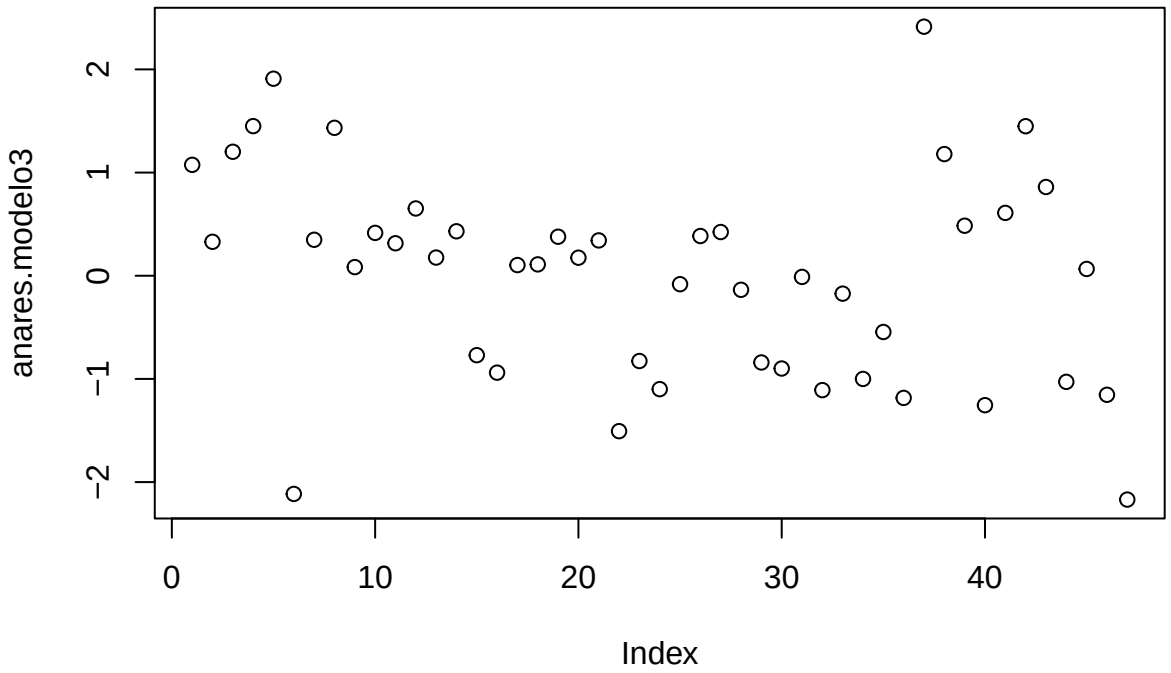
Teste Durbin-Watson de Autocorrelação H0 - Os resíduos não são autocorrelacionados

O teste de Durbin-Watson indica que devemos rejeitar a hipótese de Independência (Não Autocorrelação) dos resíduos. Neste caso o modelo não é satisfatório para uma regressão linear

```
dwtest(modelo3)
```

```
##
##  Durbin-Watson test
##
## data:  modelo3
## DW = 1.465, p-value = 0.01571
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(anares.modelo3)
```



Dessa forma, como os pressupostos do modelo não foram satisfeitos, concluímos que esse primeiro modelo ajustado não é adequado. Assim, podemos fazer alguma transformação na variável resposta ou nas variáveis explicativas para tentar atender os pressupostos do modelo de regressão. Faremos então a aplicação da transformada de Box-Cox.

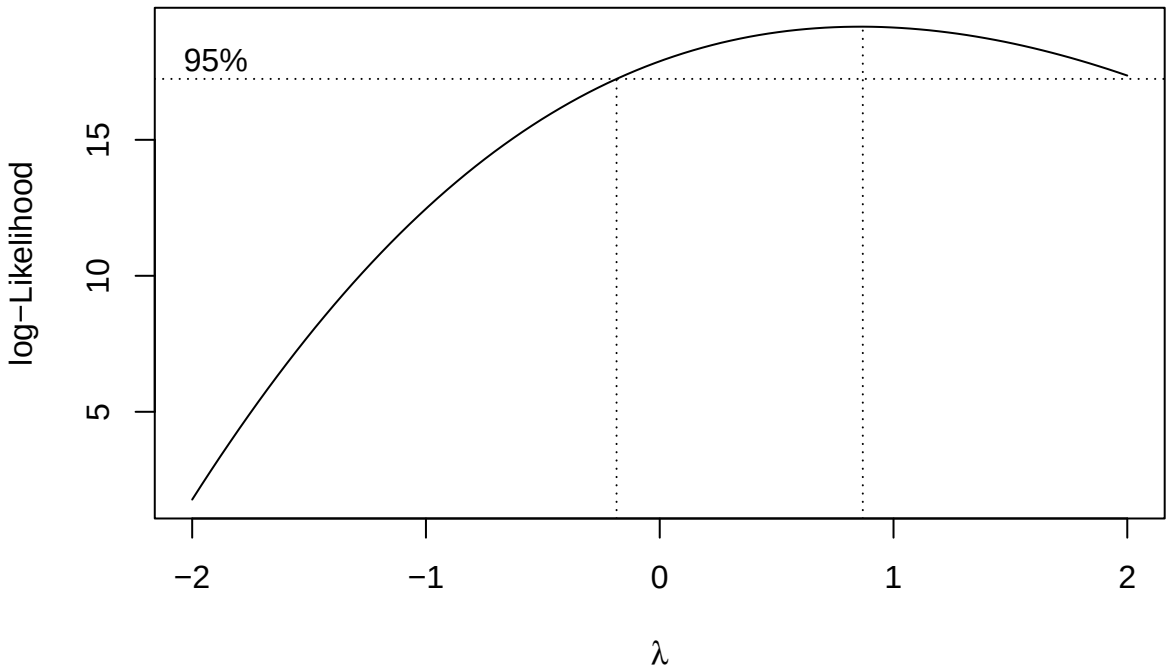
Aplicado a Transformada de Box-Cox

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
boxcox(modelo, lambda = seq(-2,2, 1/10), plotit = TRUE, data=swiss)
```



Lambda aproximadamente em 0,8. Elevaremos a variável resposta a 0,8.

```
modelo4 <- lm((Fertility)^0.8 ~ Agriculture  + Education + Catholic + Infant.Mortality, swiss)
summary(modelo4)
```

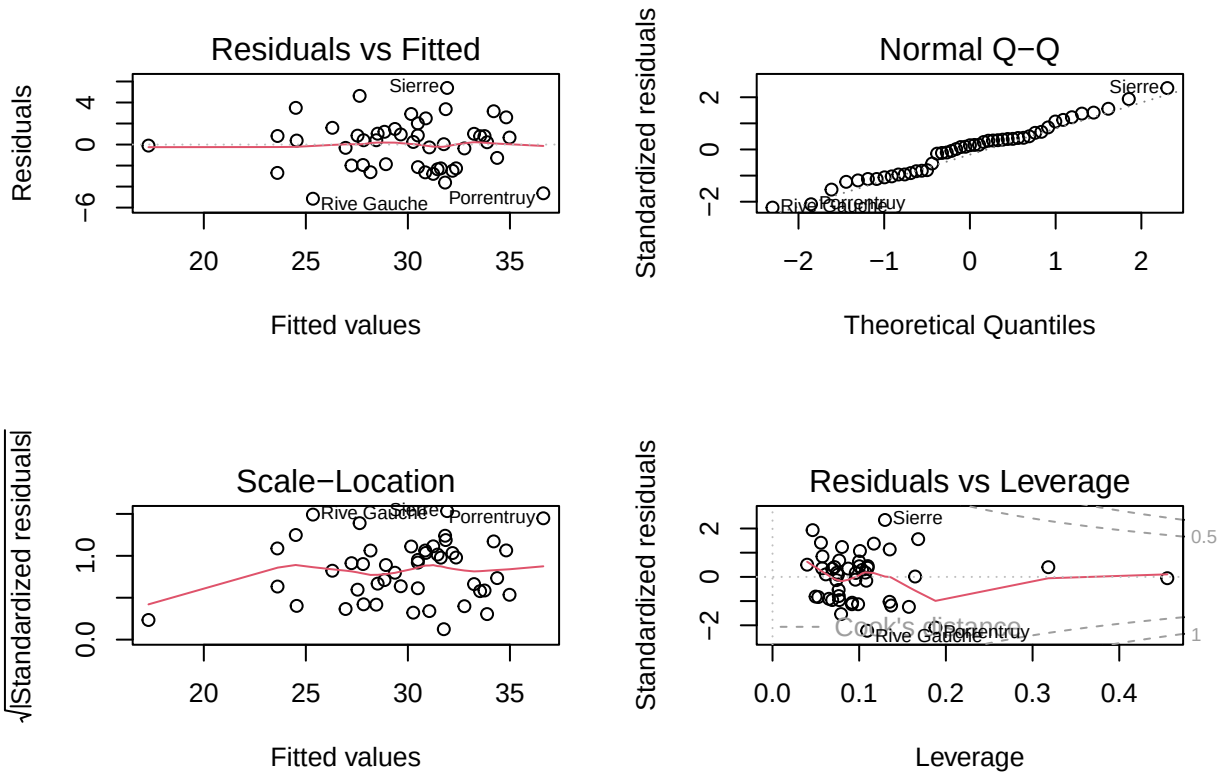
```
##
## Call:
## lm(formula = (Fertility)^0.8 ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1559 -2.0680  0.3781  1.1230  5.3851
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      27.215087   3.289023   8.275 2.32e-10 ***
## Agriculture      -0.053025   0.023350  -2.271 0.028345 *
## Education        -0.346367   0.050727  -6.828 2.56e-08 ***
## Catholic          0.041546   0.009894   4.199 0.000136 ***
## Infant.Mortality  0.374183   0.130763   2.862 0.006546 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 42 degrees of freedom
## Multiple R-squared:  0.7066, Adjusted R-squared:  0.6786
## F-statistic: 25.29 on 4 and 42 DF,  p-value: 1.039e-10
```

Analisando o summary houve uma piora do R-quadrado ajustado (diminuiu um pouco)

Análise dos Resíduos

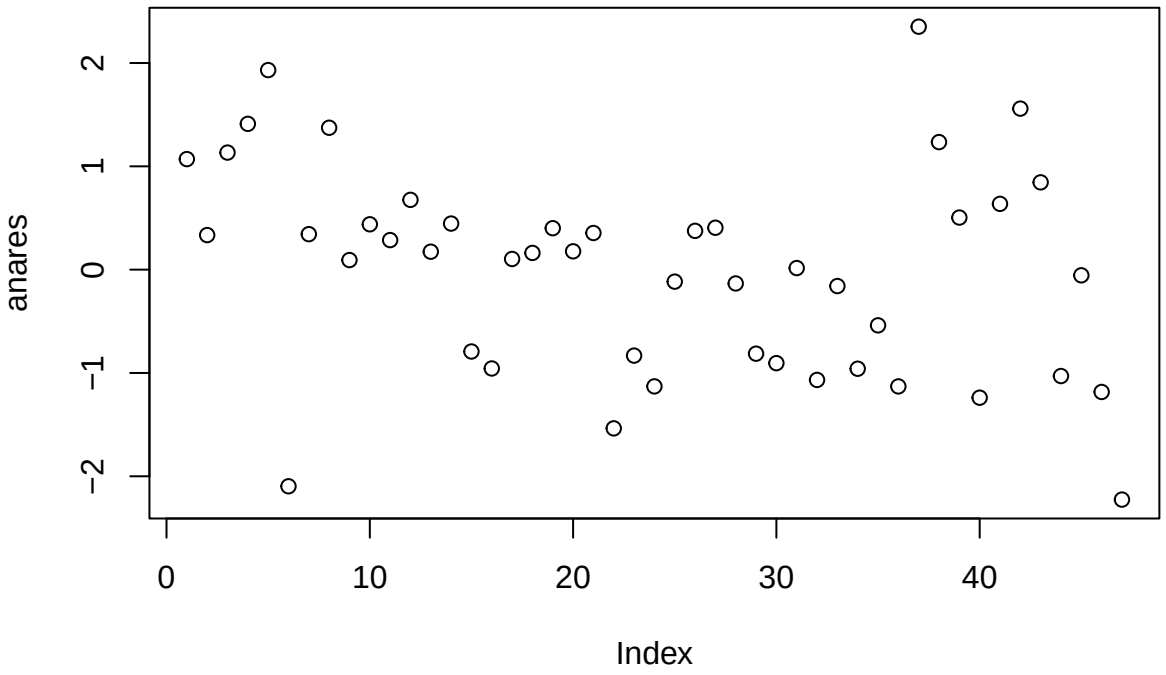
Passemos para a Análise dos resíduos após a transformação box-cox.

```
anares <- rstandard(modelo4)
par(mfrow=c(2,2))
plot(modelo4)
```



Pelo Normal Q-Q podemos verificar que o pressuposto da normalidade parece ter sido atendido para o modelo com a transformação de Box-Cox, assim como a suposição de homocedasticidade, visto que o pelos gráficos 1 e 3 os pontos parecem estar aleatórios. No entanto, o gráfico para os resíduos padronizados abaixo indica que a questão da autocorrelação persiste.

```
plot(anares)
```



Procedendo os testes formais, temos os seguintes resultado:

```
library(nortest)
ad.test(anares)
```

```
##
##  Anderson-Darling normality test
##
## data:  anares
## A = 0.49314, p-value = 0.2069
```

Como p-valor acima de 5%, Verificamos que a transformação de Box-Cox continuamos com a normalidade dos resíduos.

```
shapiro.test(anares)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anares
## W = 0.97929, p-value = 0.5634
```

Como p-valor acima de 5%, verificamos que a transformação de Box-Cox continuamos com a normalidade dos resíduos.

```
library(lmtest)
bptest(modelo4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo4
## BP = 2.9308, df = 4, p-value = 0.5695
```

Também com o valor acima de 5% o p-valor para o teste de homocedasticidade, atende o pressuposto de variância constante dos resíduos, implicando no atendimento dessa suposição.

Teste Durbin-Watson de Autocorrelação

```
dwtest(modelo4)
```

```
##
##  Durbin-Watson test
##
## data:  modelo4
## DW = 1.4395, p-value = 0.01232
## alternative hypothesis: true autocorrelation is greater than 0
```

No entanto, como vimos em sala de aula, a transformação de Box-Cox não conseguiu resolver a questão da autocorrelação (p-valor $< 0{,}05$). Dessa forma, o modelo não é válido.