Continuação exercicio 1

Cynthia Tojeiro

2025-02-08

Inicialmente vamos carregar os dados no R, utilizando os códigos a seguir.

```
# Limpa a memória

#rm(list=ls())

# Pacote necessario para leitura dos dados
library(readxl) #para ler no excell

# Dados referentes a diabetes mellitus tipo 1 (DM1)
dados <- read.table("C:\\datascience\\exercicio1.txt")

#Para ver as 6 primeiras linhas do conjunto de dados

head(dados)
```

```
##   V1 V2 V3 V4   V5  V6 V7       V8
## 1  1  1 57 11 25.8  56  0 7.413015
## 2  2  1 67 17 29.6 189  0 6.252037
## 3  3  1 42  9 25.2 122  1 4.335380
## 4  4  0 62  8 24.6 169  1 7.390550
## 5  5  1 50  8 20.2 133  0 5.463982
## 6  6  1 62  8 26.2 172  0 7.458546
```

```
attach(dados)
```

```
sexo <- dados[,2]
idade<-dados[,3]
escolaridade<-dados[,4]
imc<-dados[,5]
tempodiabetes<-dados[,6]
usoinsulina<-dados[,7]
hemoglobina<-dados[,8]
```

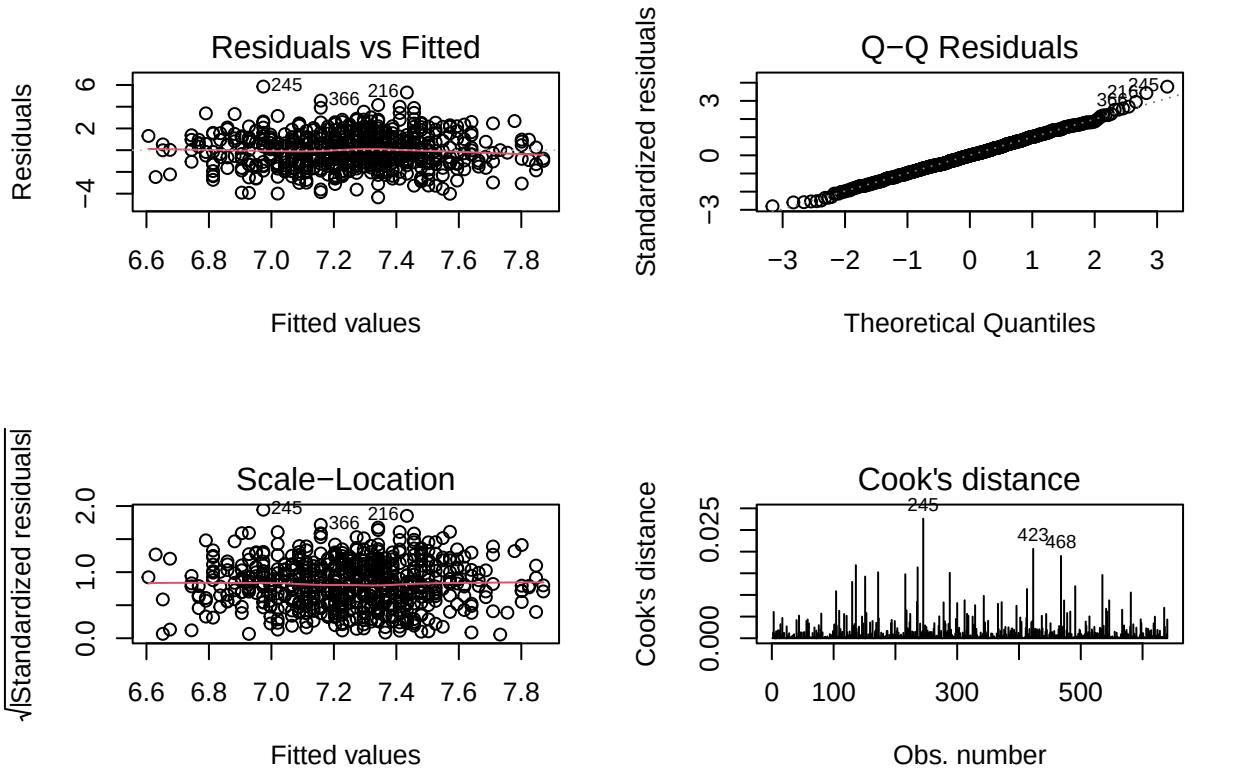
Ajustando o modelo com hemoglobina em função da idade

```
fit.model<-result<-lm(hemoglobina~idade)
summary(result)
```

```
##
## Call:
## lm(formula = hemoglobina ~ idade)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3418 -1.0183 -0.0115  1.0479  5.8510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.986025   0.341365  17.536  < 2e-16 ***
## idade       0.022979   0.006086   3.776 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.55 on 638 degrees of freedom
## Multiple R-squared:  0.02186,    Adjusted R-squared:  0.02032
## F-statistic: 14.26 on 1 and 638 DF,  p-value: 0.0001745
```

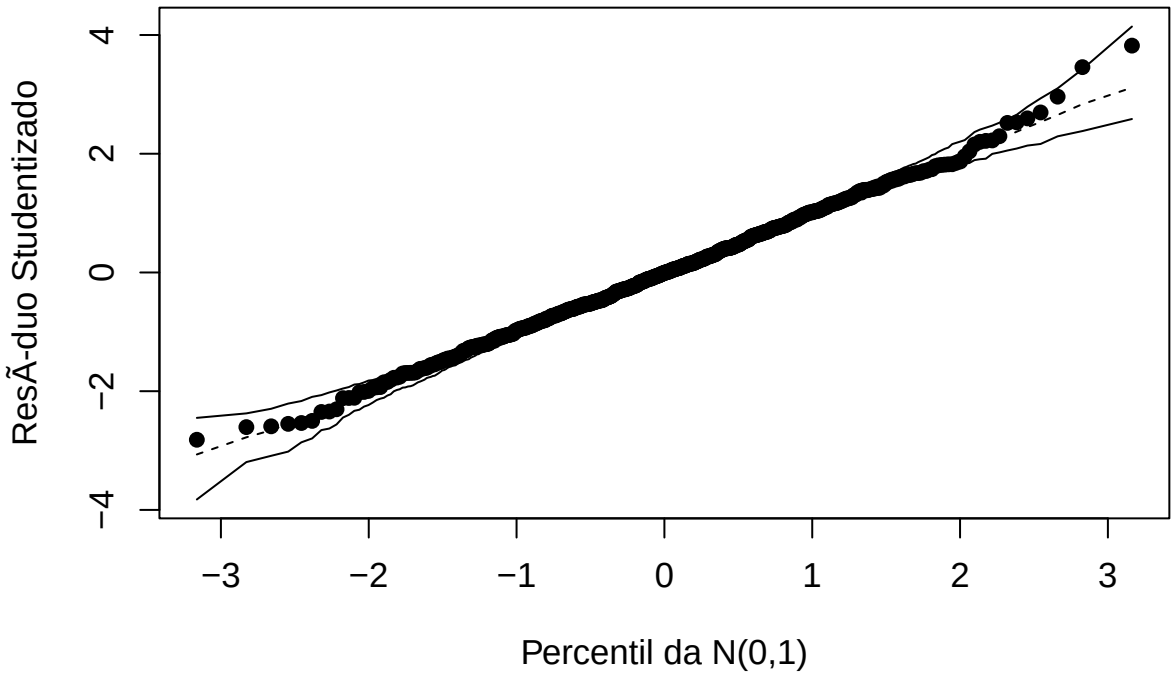
```r
#Análise de Resíduos
par(mfrow=c(2,2))
plot(fit.model, which = 1:4)
```



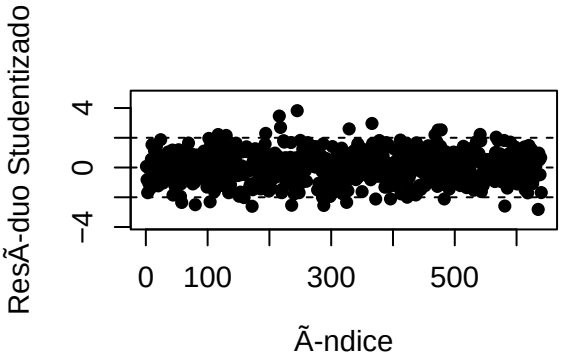
- O gráfico do canto esquerdo superior trás os resíduos ordinários versus os valores ajustados. Esse gráfico é importante para verificação de possíveis padrões não aleatórios, heterocedasticidade, presença de outliers e pontos influentes; o padrão que indica bom ajuste é o de pontos dispersos aleatóriamente e a linha de tendência aproximadamente constante em torno de zero.
- O gráfico do canto direito superior mostra os quantis teóricos da distribuição normal padrão contra os resíduos padronizados. Esse gráfico permite avaliar a pressuposição de normalidade e, caso não haja normalidade, permite avaliar a forma da distribuição, além de indicar possíveis outliers. Pontos dispersos aleatoriamente, nas proximidades da linha identidade (pontilhada), indicam normalidade.

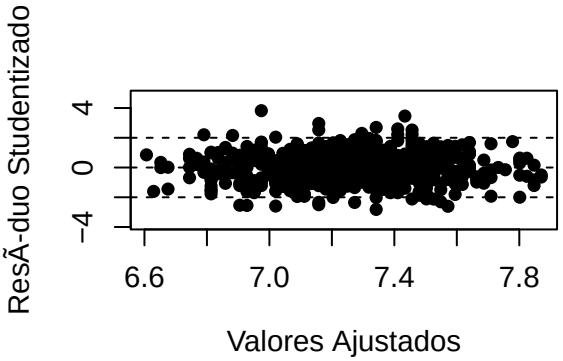
- O gráfico do canto inferior esquerdo apresenta a raiz quadrada dos resíduos padronizados versus os valores ajustados. Esse gráfico é uma alternativa ao primeiro, baseado nos resíduos padronizados. Tendências nesse gráfico são indicativos de variância não constante.
- E por fim, o gráfico do canto inferior direito apresenta os valores da distância de Cook para cada observação. A distância de Cook é uma medida de diferença das estimativas dos parâmetros do modelo ao considerar e ao desconsiderar uma particular observação no ajuste. Observações com valores elevados para essa medida devem ser verificadas. Os gráficos de resíduos indicam aparente heterocedasticidade, além de observações atípicas.

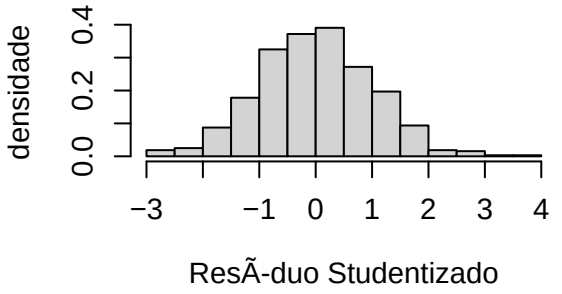
```
source("C:\\datascience\\Programas\\Diag2.norm.r")
source("C:\\datascience\\Programas\\Envel_norm.r")
source("C:\\datascience\\Programas\\anainflu_norm.r")

par(mfrow=c(1,1))
envelnorm(fit.model)
```

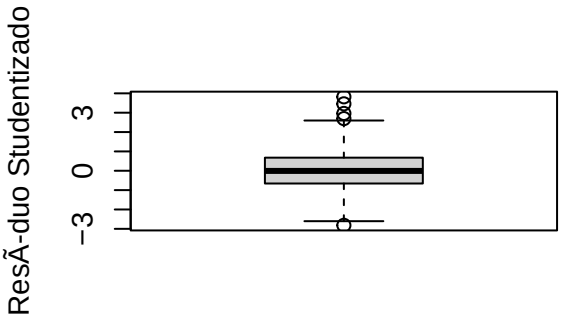



```
diag2norm(fit.model)
```

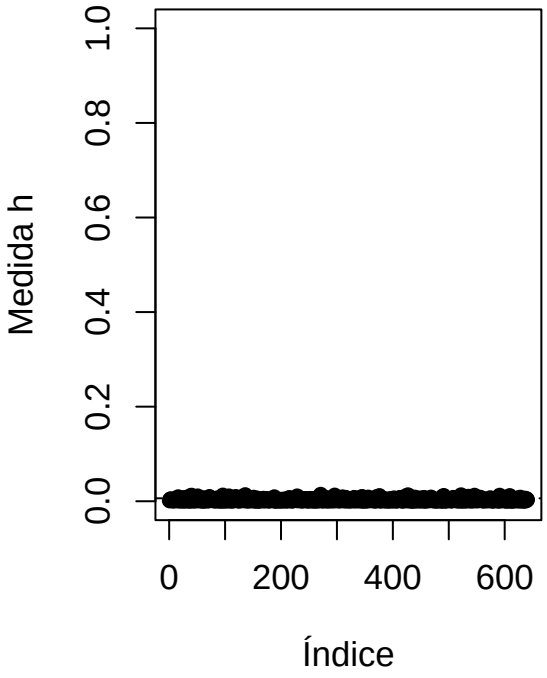


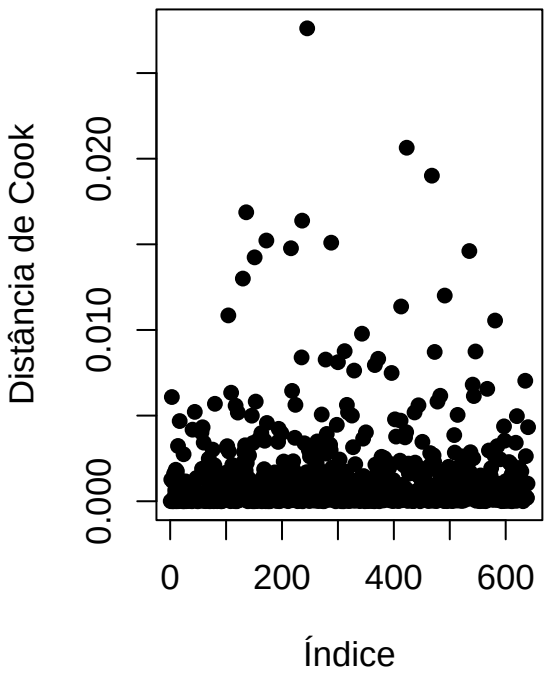







```
anainflu_norm(fit.model)
```






```
# R2

summary(result)$r.squared
```

```
## [1] 0.02185618
```

```
summary(result)$adj.r.squared
```

```
## [1] 0.02032304
```

#Pressupostos do Modelo

1. Normalidade dos Resíduos Teste de Shapiro-Wilk: O Teste de Shapiro-Wilk tem como objetivo avaliar se uma distribuição é semelhante a uma distribuição normal. A distribuição normal também pode ser chamada de gaussiana e tem a forma de sino. Esse tipo de distribuição é muito importante, por ser frequentemente usado para modelar fenômenos naturais. Quando o p-value for maior que 0,05 ($p > 0.05$) a hipótese nula (dos dados seguirem uma distribuição normal) é aceita. Chamamos a função shapiro.test() indicando o vetor "fit.model", selecionando a opção residuals.

```
shapiro.test(fit.model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.model$residuals
## W = 0.99769, p-value = 0.5289
```

Temos um p-value maior que 0,05. Consideramos então a hipótese nula, indicando uma distribuição é normal.

2. Teste de Homocedasticidade: Em análise de variância (ANOVA), há um pressuposto que deve ser

atendido que é de os erros terem variância comum, ou seja, homocedasticidade. Isso implica que cada tratamento que se está sendo comparado pelo teste F, deve ter aproximadamente a mesma variância para que a ANOVA tenha validade. (Esse teste não funciona em caso de resíduos não normais) Para fazer o teste de homocedasticidade chamamos a função bptest(), inserindo nosso modelo (fit.model). Para vusar essa função devemos instalar o pacote lmtest

```
library(lmtest)
bptest(fit.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit.model
## BP = 0.020181, df = 1, p-value = 0.887
```

Assim como no teste de shapiro, quando o p-value for maior que 0,05 não rejeitamos a hipótese nula e consideramos que existe homocedasticidade.

3. Independência dos Resíduos (Durbin Watson): Teste de Durbin Watson: É o modelo mais popular para estudar a relação entre duas variáveis, no qual os parâmetros de interesse são estimados a partir da minimização da soma dos quadrados dos resíduos. Estes estimadores são conhecidos como estimadores de mínimos quadrados ordinários (MQO). Uma autocorrelação positiva é identificada por um agrupamento de resíduos com o mesmo sinal. Uma autocorrelação negativa é identificada por rápidas mudanças nos sinais de resíduos consecutivos. Use a estatística Durbin-Watson para testar a presença de autocorrelação. Para fazer o teste chamamos a função durbinWatsonTest(), inserindo nosso modelo fit.model. Para usar esse teste devemos instalar o pacote car.

```
library(car)
durbinWatsonTest(fit.model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01439274      2.024357   0.728
##  Alternative hypothesis: rho != 0
```

Analisamos aqui a estatística de Durbin Watson (D-W Statistic), esse valor deve estar próximo de 2. Para que exista independência dos resíduos aceita-se valores com intervalo entre 1 a 3

4. Outliers nos Resíduos: Para obtermos os resíduos padronizados utilizamos a função summary(), inserimos nela outra função chamada rstandart() e indicamos nosso modelo (fit.model).

```
summary(rstandard(fit.model))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.802846 -0.657469 -0.007418 -0.000052  0.676986  3.781029
```

Observando os valores Min e Max, percebe-se que os resíduos não estão fora do intervalo -3 e 3. Sendo assim, não há outliers.

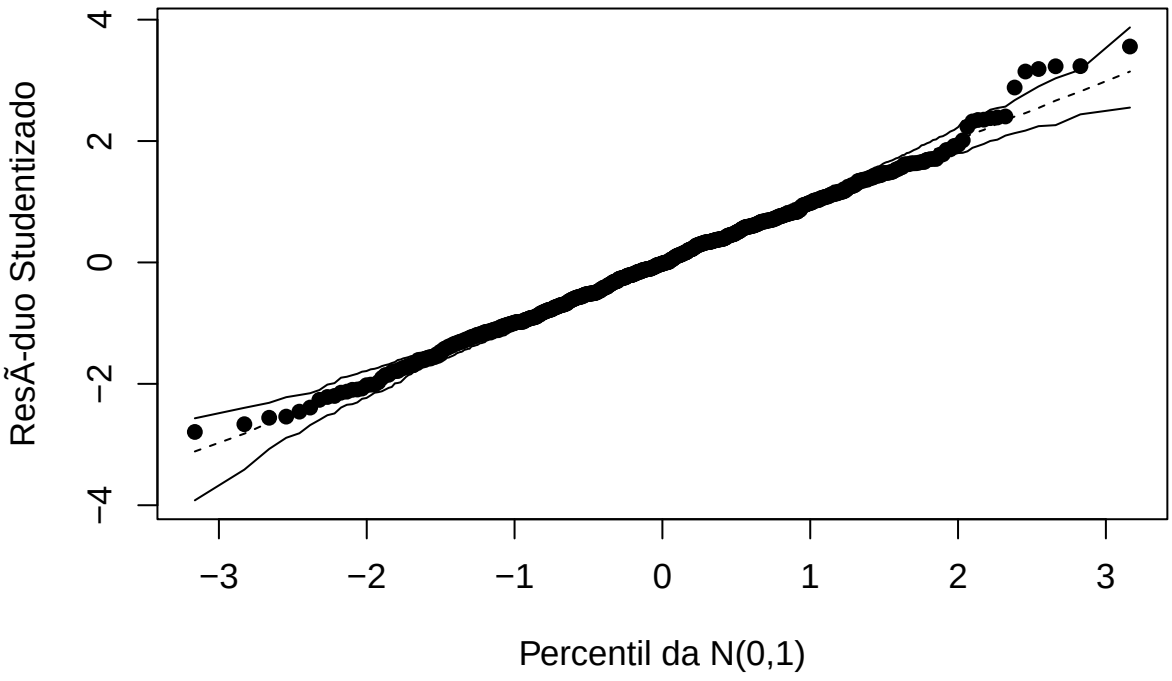
#Modelos de Regressão Múltipla Vamos continuar com os mesmos dados, entretanto agora para fazer o ajuste do modelo de regressão múltipla usamos a função lm da seguinta forma:

```
fit.model<-result<-lm(hemoglobina~sexo+idade+escolaridade+imc+tempodiabetes+usoinsulina)
summary(result)
```

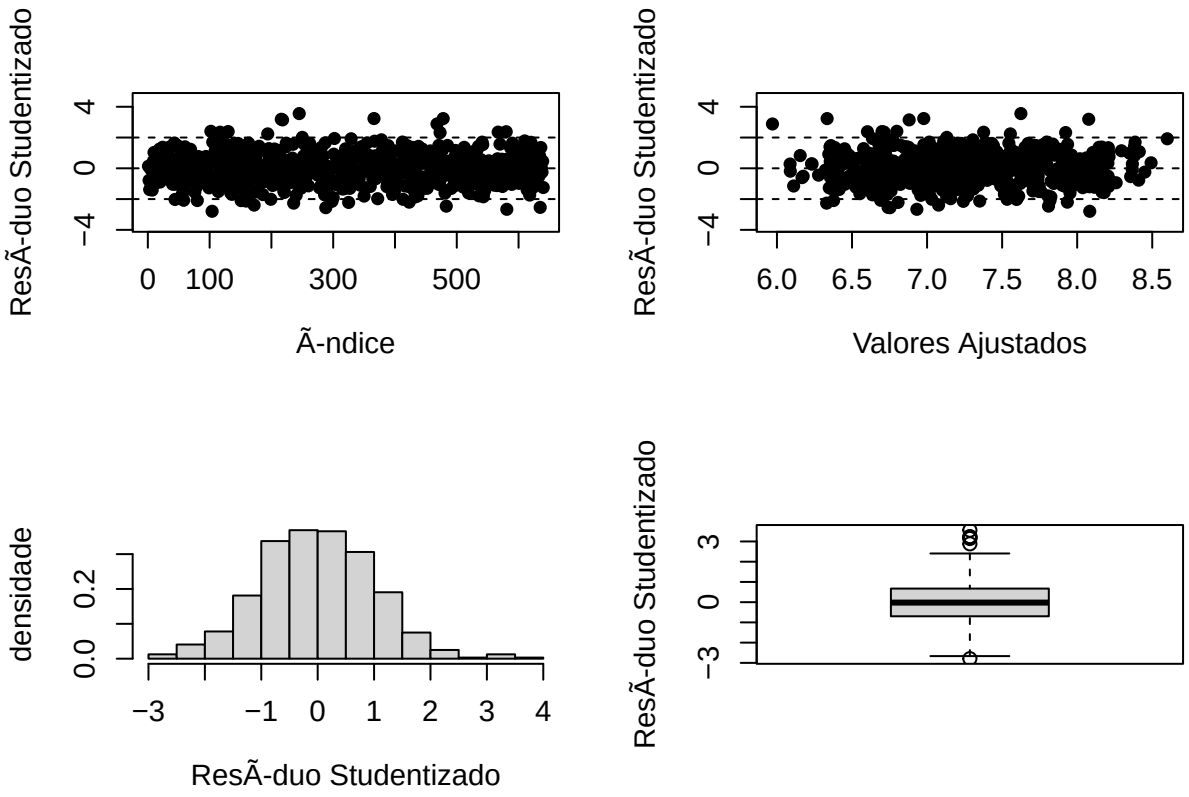
```
##
## Call:
## lm(formula = hemoglobina ~ sexo + idade + escolaridade + imc +
##     tempodiabetes + usoinsulina)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.0892 -1.0246 -0.0327  0.9884  5.2002
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.692291   0.545648  12.265  < 2e-16 ***
## sexo          -0.501599   0.120042  -4.179 3.35e-05 ***
## idade          0.020101   0.005839   3.443 0.000614 ***
## escolaridade  -0.044630   0.014989  -2.978 0.003017 **
## imc            0.010711   0.014669   0.730 0.465559
## tempodiabetes  0.001615   0.000977   1.653 0.098765 .
## usoinsulina   -0.720681   0.117574  -6.130 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 633 degrees of freedom
## Multiple R-squared:  0.1125, Adjusted R-squared:  0.1041
## F-statistic: 13.37 on 6 and 633 DF,  p-value: 2.661e-14
```

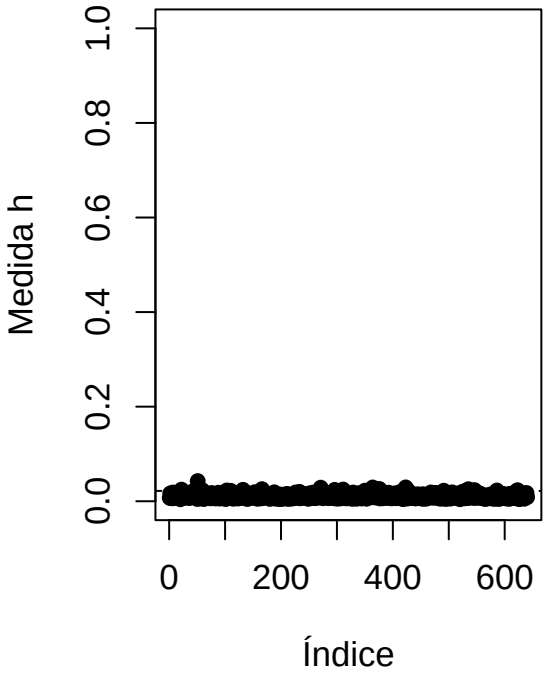
```r
source("C:\\datascience\\Programas\\Diag2.norm.r")
source("C:\\datascience\\Programas\\Envel_norm.r")
source("C:\\datascience\\Programas\\anainflu_norm.r")

par(mfrow=c(1,1))
envelnorm(fit.model)
```

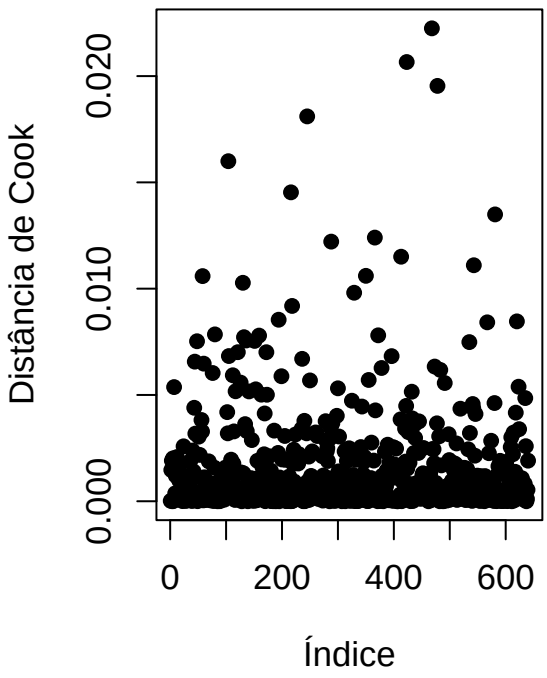


```
diag2norm(fit.model)
```




```
anainflu_norm(fit.model)
```






```
# R2

summary(result)$r.squared
```

```
## [1] 0.1124862
```

```
summary(result)$adj.r.squared
```

```
## [1] 0.1040737
```

#Pressupostos do Modelo

```
shapiro.test(fit.model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.model$residuals
## W = 0.99571, p-value = 0.07553
```

```
bptest(fit.model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit.model
## BP = 3.5972, df = 6, p-value = 0.731
```

```r
durbinWatsonTest(fit.model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04272059      2.082977   0.304
##  Alternative hypothesis: rho != 0
```

```r
  summary(rstandard(fit.model))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.777883 -0.696319 -0.022118 -0.000069  0.670777  3.525077
```