Modelos Lineares Generalizados para Dados de Contagem

Ananda Bordignon, Brendha Lima, Giovanna Lazzarin

28 de novembro de 2018

Introdução

Dados

Os dados a respeito do número de acidentes no estado de Alagoas no ano de 2016 foram extraídos do Departamento de Informática do Sistema Único de Saúde (DATASUS), suas características sociodemográficas do último senso de 2010 foram extraídas no portal do Instituto de Pesquisa Econômica Aplicada (IPEADATA). Segundo o Batalhão de Polícia de Trânsito (BPTRAN), todo evento ocorrido na via pública (incluindo calçadas), decorrente do trânsito de veículos e pessoas, que resulta em danos humanos e/ou materiais é definido como acidente de trânsito . Compreende: colisões entre veículos, choque com objetos fixos, capotamentos, tombamentos, atropelamentos, queda de pedestres e ciclistas, etc. Além disso toda ocorrência fortuita ou danosa, envolvendo veículos em circulação, ou parados, respectivos ocupantes, pedestres e objetos móveis ou fixos.

Os dados a serem trabalhados neste estudo referem-se à acidentes de trânsito ocorridos nas vias municipais, sem incluir as Rodovias Estaduais e Federais.

Cada linha da base corresponde a 1 dos 102 municípios do estado do Alagoas, as características sociodemográficas selecionadas foram estas:

frota - Frota total de veículos.

pib - Pib per capita a preços correntes.

populacao - População residente.

emergencia - Estabelecimentos de saúde com atendimento de emergência total.

o interesse deste trabalho é modelar o número de acidentes de trânsito em funções das demais.

As primeiras 5 linhas da base de dados tem a seguinte forma:

```
##   codigos             municipios obitos pib populacao frota emergencia
## 1  270010            Água Branca     14  24     19377  2370          1
## 2  270020                 Anadia      2  49     17424  1718          1
## 3  270030              Arapiraca      6 101    214006 68913          9
## 4  270040                Atalaia     16  81     44322  4164          1
## 5  270050 Barra de Santo Antônio      2  54     14230  1134         NA
## 6  270060    Barra de São Miguel      2  98      7574   935         NA
```

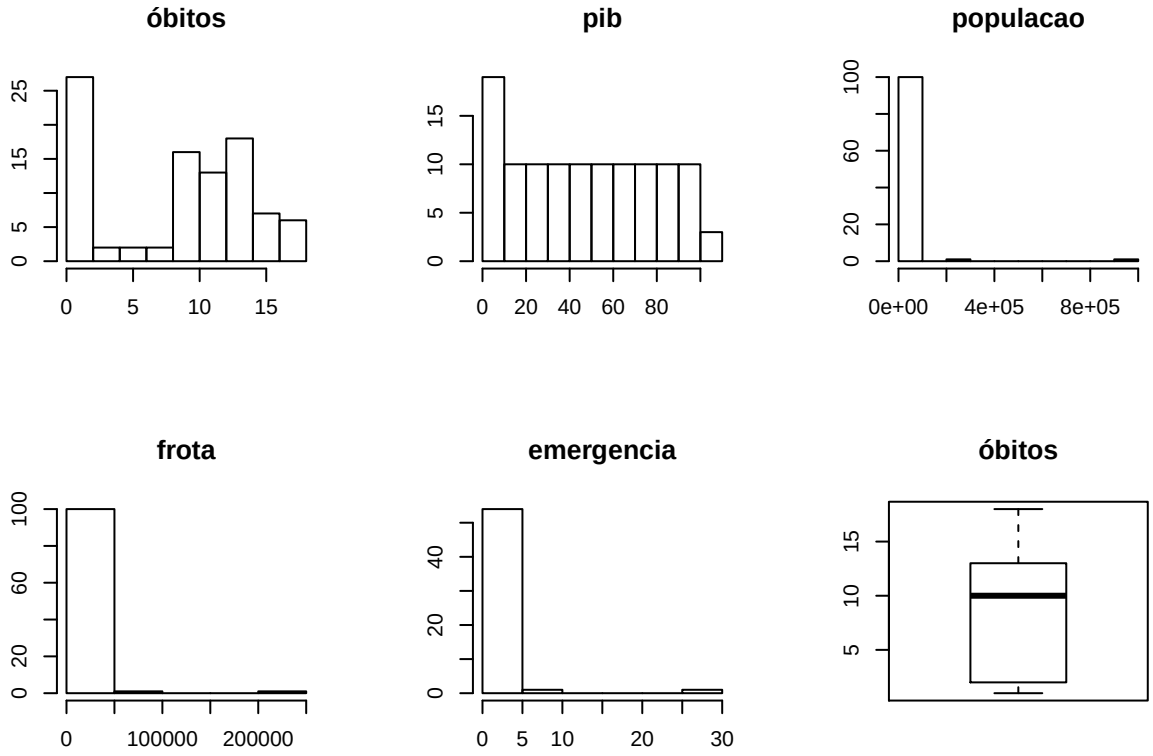
Análise Descritiva

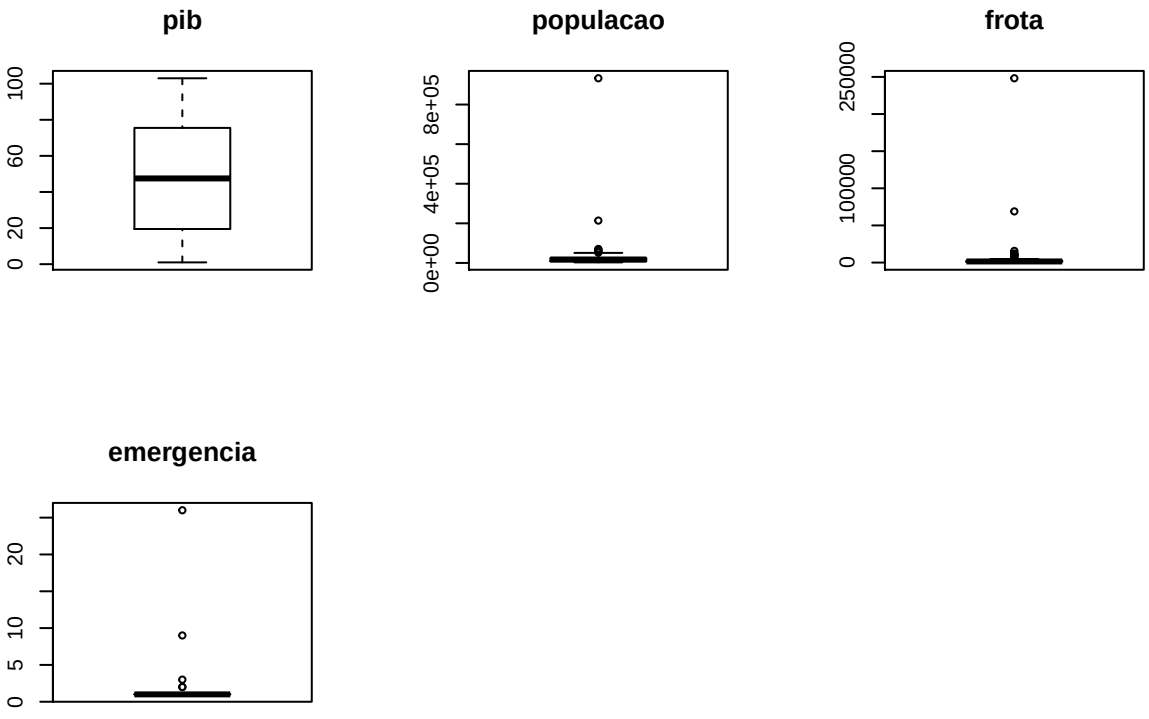
Para iniciar uma breve análise descritiva, podemos usar a função summary para visualizar o mínimo, máximo, mediana e quartis das varíaveis explicativas do estudo.

```
##     codigos             municipios     obitos           pib        
##  Min.   :270010              : 9   Min.   : 1.000   Min.   :  1.00  
##  1st Qu.:270243              : 1   1st Qu.: 2.000   1st Qu.: 19.75  
##  Median :270475   Água Branca: 1   Median :10.000   Median : 47.50  
##  Mean   :270478   Anadia     : 1   Mean   : 9.065   Mean   : 47.90  
##  3rd Qu.:270708   Arapiraca  : 1   3rd Qu.:13.000   3rd Qu.: 75.25  
##  Max.   :270940   Atalaia    : 1   Max.   :18.000   Max.   :103.00  
##  NA's   :10       (Other)    :98   NA's   :19                       
##    populacao          frota            emergencia    
##  Min.   :  2866   Min.   :   253.0   Min.   : 1.000  
##  1st Qu.:  8444   1st Qu.:   789.2   1st Qu.: 1.000  
##  Median : 17077   Median :  1486.0   Median : 1.000  
##  Mean   : 30593   Mean   :  5450.6   Mean   : 1.661  
##  3rd Qu.: 25352   3rd Qu.:  2617.2   3rd Qu.: 1.000  
##  Max.   :932748   Max.   :248178.0   Max.   :26.000  
##  NA's   :10       NA's   :10         NA's   :56
```

Na variável município são elencados todos os municípios do Alagoas. Para a variável resposta, nota-se um total de XX de dados faltantes, ou seja, não foram registrados acidentes de trânsito nestes municípios.

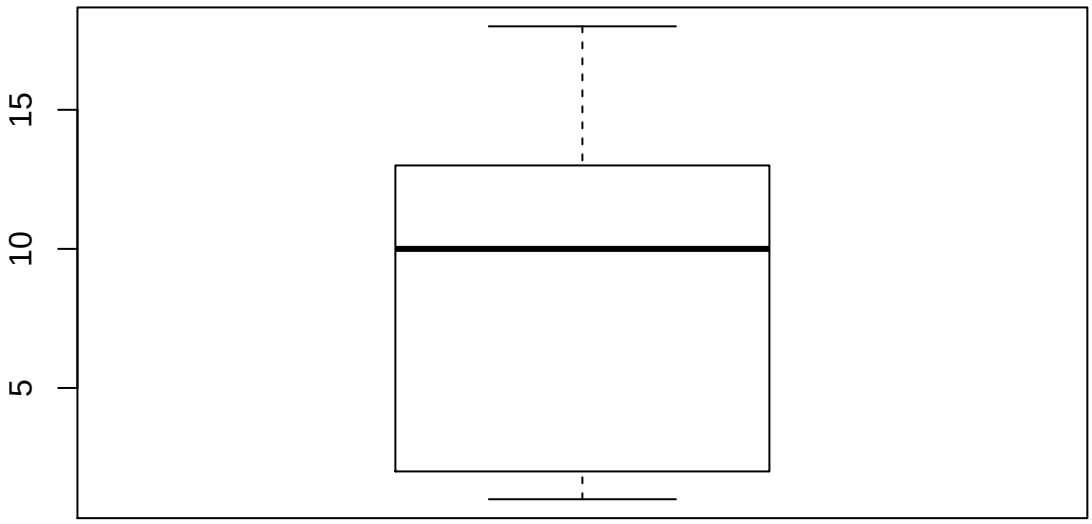
Observamos alguns histogramas e box-plots que facilitam a visualização dos dados. Nota-se que algumas variáveis têm um ou mais pontos discrepantes. Em alguns casos pode ser conveniente trabalhar com o log da variável para obter uma maior simetria, ou, outra alternativa é remover os valores discrepantes.

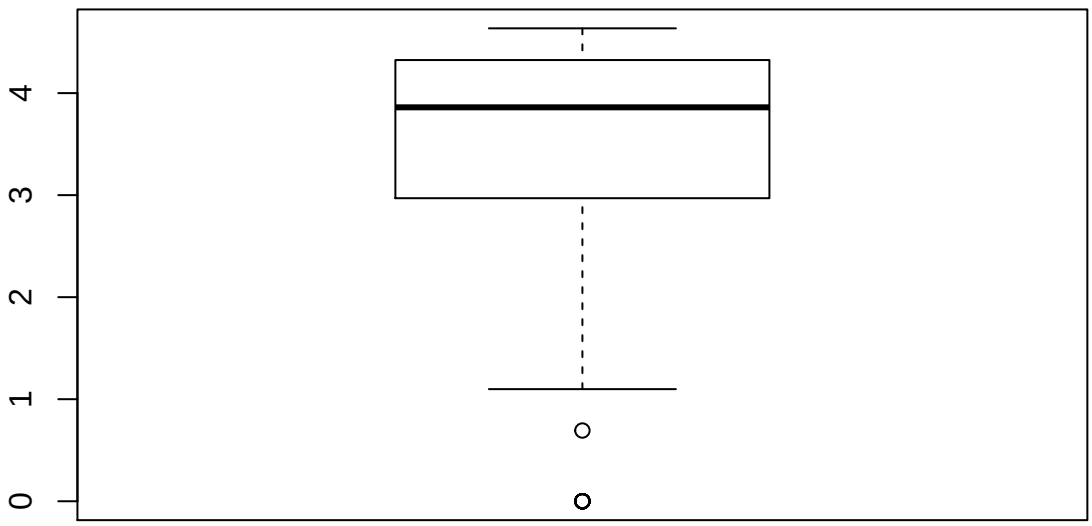


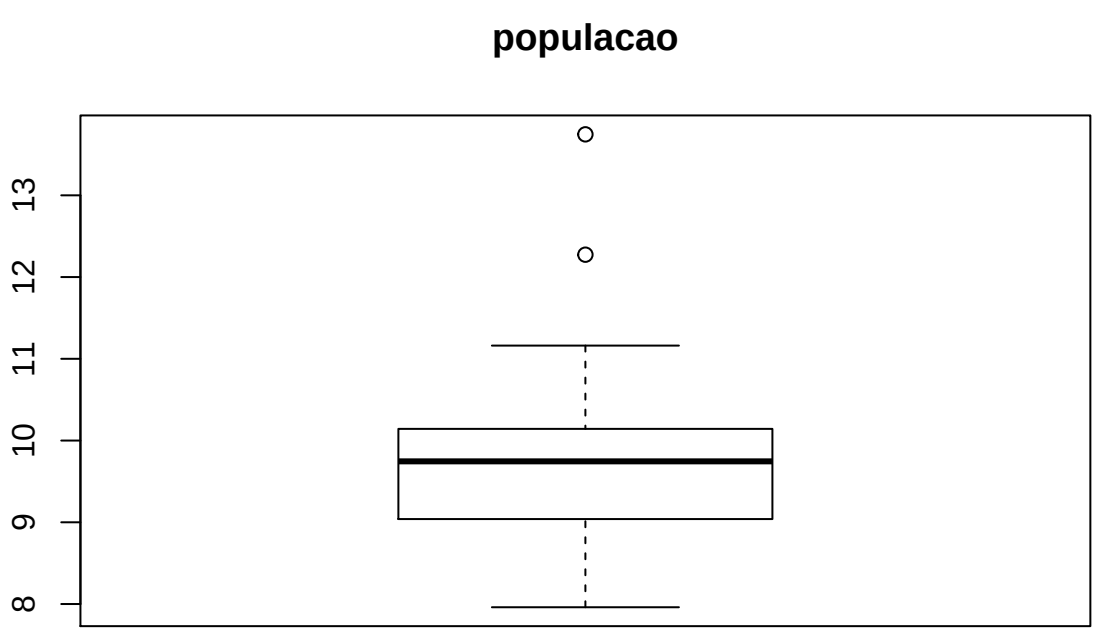


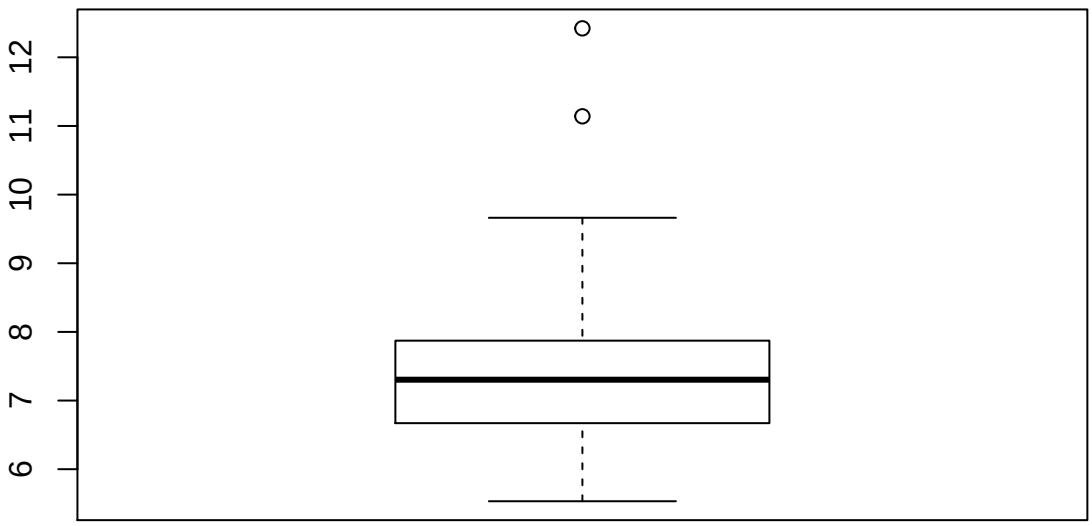
Transformação das covariáveis

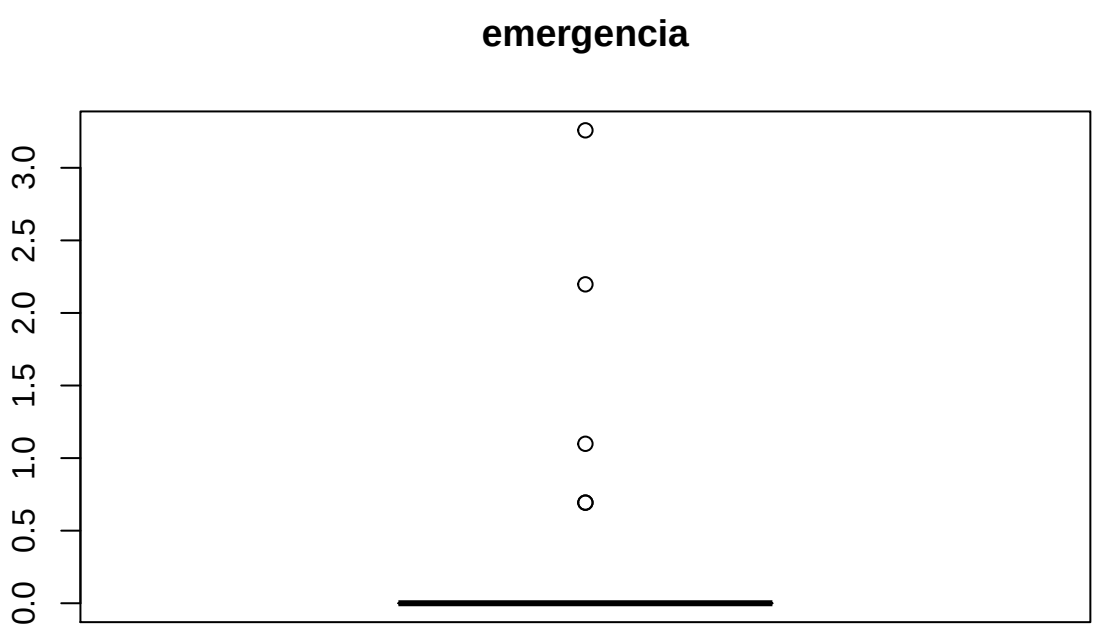
Aplicaremos uma transformação logaritmica nas variáveis observadas como mais assimétricas nas análises descritivas e verificaremos novamente a forma da distribuição das transformações.


óbitos
15
10
5


pib
0
1
2
3
4


populacao
8
9
10
11
12
13


frota
6
7
8
9
10
11
12



Agora podemos observar uma simetria maior nas distribuições das variáveis transformadas.

Correlação

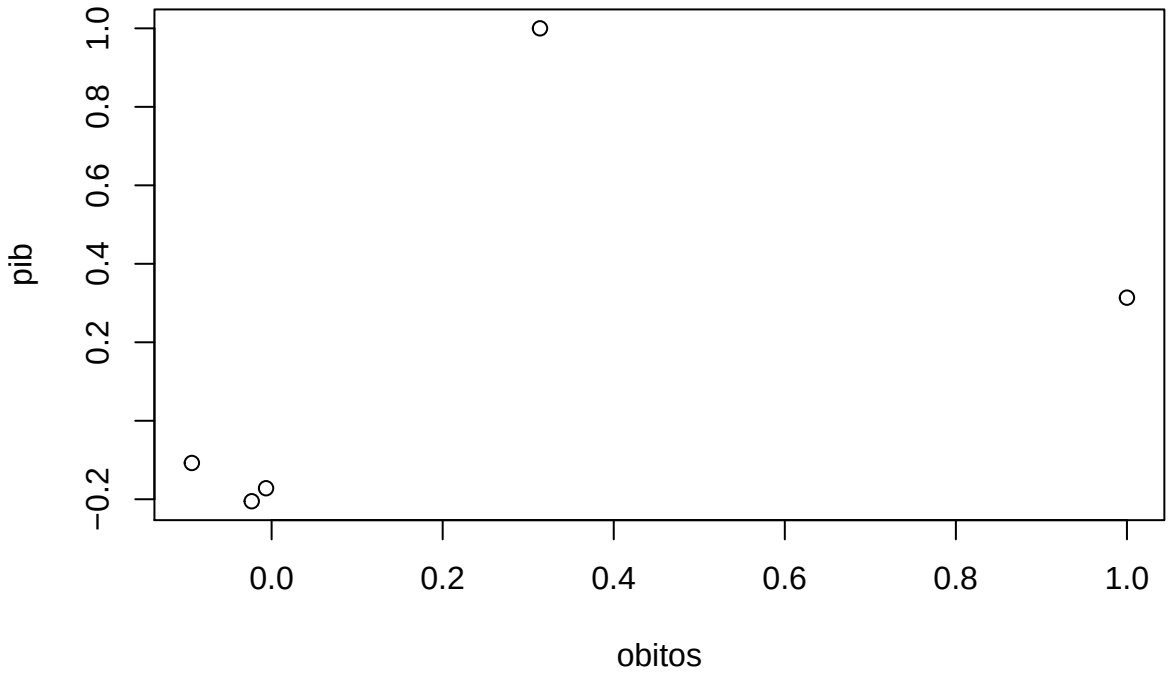
Agora, verificaremos a correlação entre as variáveis em estudo, nesta etapa vamos considerar as variáveis transformadas no tópico anterior.

Como já visto, a variável resposta do estudo possui dados faltantes, para resolver este problema é possível obter a correlação desta com as demais utilizando o argumento use da função cor.

```
#PRECISA FAZER

cor <- cor(da[,c(3,4,5,6,7)], use = "na.or.complete")

plot(cor)
```



Gráficos de Dispersão

OI AMIGAS SE QUISEREM COLOQUEM O GRÁFICO DE DISPERSÃO AQUI!!!!!!!!!

Ajuste dos Modelos de Regressão

Neste trabalho, queremos modelar uma variável de contagem, ou seja, uma variável discreta com suporte no conjunto dos inteiros não negativos. Nossa resposta é o número de acidentes de trânsito, para problemas deste tipo, comumente a primeira alternativa de modelagem via modelo linear generalizado faz uso da distribuição Poisson com função de ligação logaritmica.

(ESCREVER ESTE PARÁGRAFO COM OUTRAS PALAVRAS) Caso o GLM Poisson não se ajuste bem há outras opções de distribuições para a resposta. Uma das alternativas mais utilizadas é a Binomial Negativa. O principal diferencial dessa distribuição em relação à Poisson é que a Binomial Negativa comporta casos em que há superdispersão, enquanto na Poisson o parâmetro de dispersão deve ser fixo e igual a 1 (equidispersão, média igual à variância).

(ESCREVER ESTE PARÁGRAFO COM OUTRAS PALAVRAS)

A seguir, mostraremos os ajustes dos Modelos Lineares Generalizados log-linear de Poisson e com distribuição Binomial Negativa para a resposta. N

Escolhendo o Modelo

(ESCREVER ESTE PARÁGRAFO COM OUTRAS PALAVRAS) Para seleção de modelos diversas medidas podem ser utilizadas, em especial vamos utilizar a verossimilhança e o AIC dos modelos. vale lembrar que o modelo Binomial Negativo estima um parâmetro de dispersão que não existe no modelo Poisson. *(ESCREVER ESTE PARÁGRAFO COM OUTRAS PALAVRAS)*

```
#O modelo que apresentou menor AIC e maior verossimilhança foi o modelo Binomial Negativo (m2)
##obs: Ambos modelos são muito proximo e muito bem ajustado, se gente quise poderia trabalhar com a Pois

ajuste = c('m1', 'm2')
aic    = c(AIC(m1), AIC(m2))
verossimilhanca = c(logLik(m1),logLik(m2))
data.frame(ajuste, aic, verossimilhanca)
```

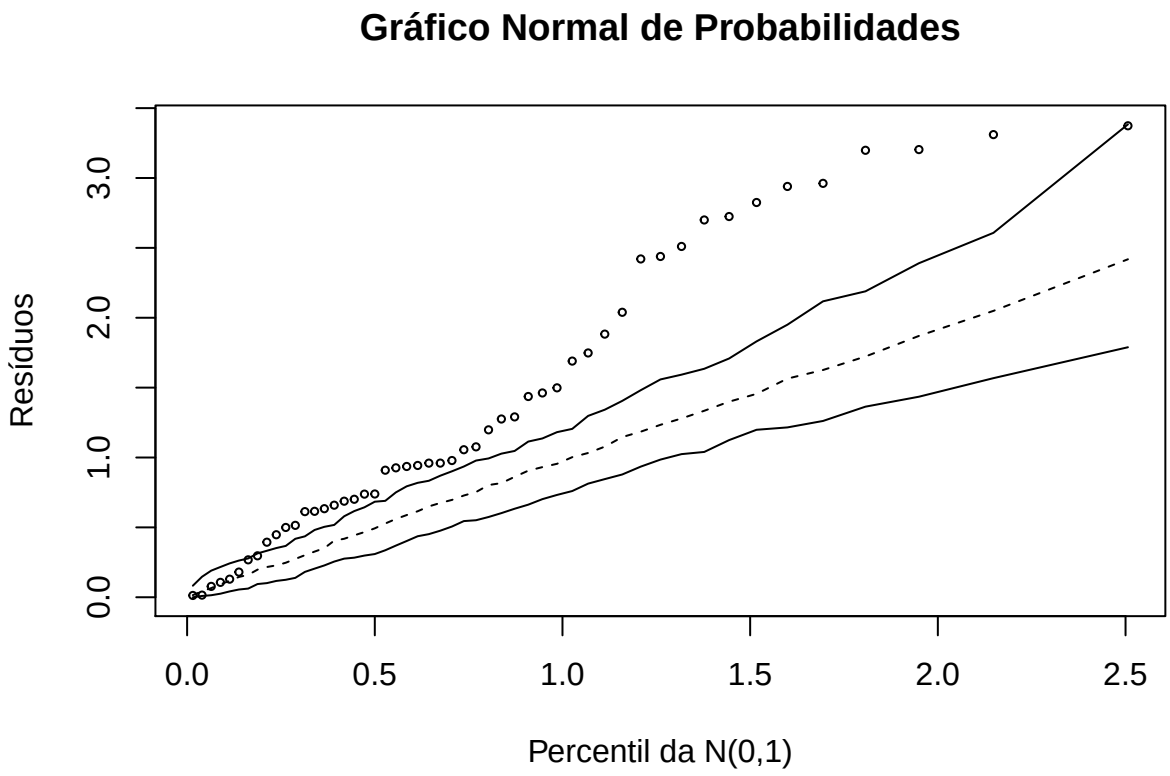
```
##   ajuste      aic verossimilhanca
## 1     m1 352.3540       -171.1770
## 2     m2 324.8689       -156.4345
```

ESCREVER INTERPRETAÇÕES SOBRE OS AICs E VEROSSIMILHANÇAS E PQ ESCOLHEU QUAL MODELO ETC

A seguir, obtivemos os gráficos envelopes dos dois ajustes, com eles podemos observar qual modelo está obtendo melhor ajuste, através de seus resíduos. Deve-se observar se há presença de pontos fora dos limites ou se há pontos dentro dos limites porém apresentando padrões sistemáticos.

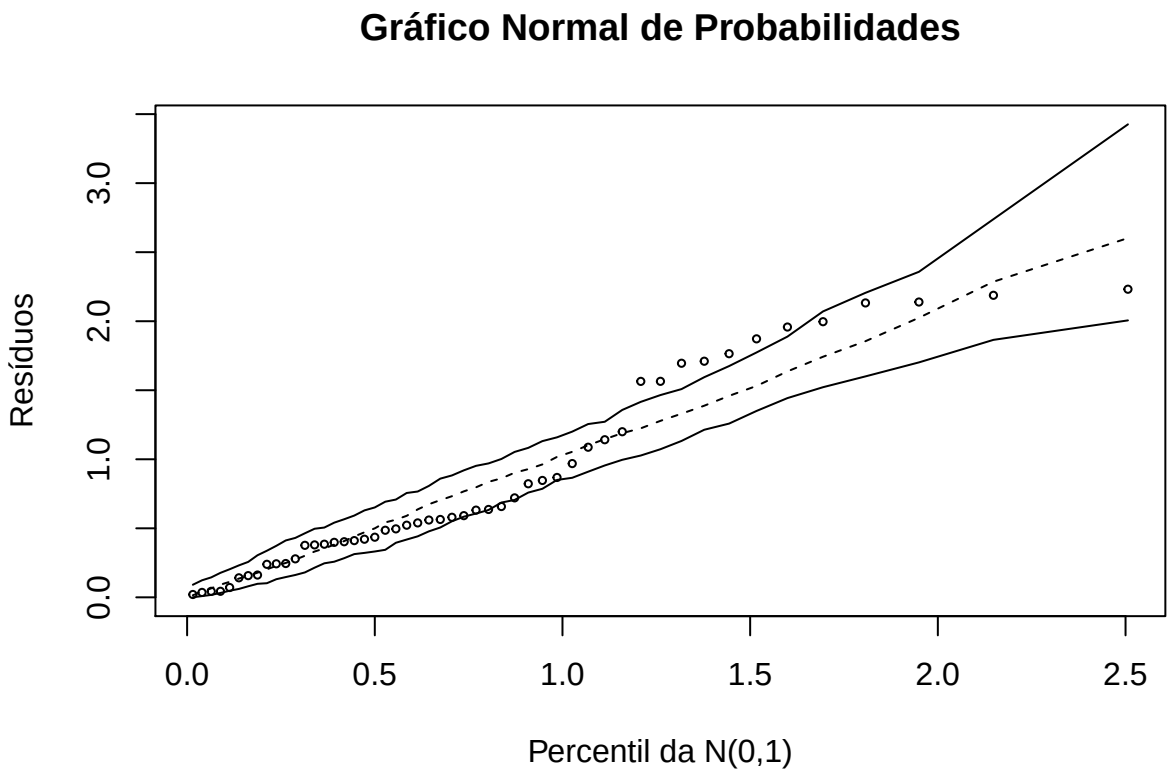
```
#GRÁFICOS ENVELOPES PARA VER SE AJUSTES ESTÃO BONS.
##NÃO CONSEGUI RODAR ESTES GRÁFICOS!!!!!!!!!!
hnp(m1, xlab = 'Percentil da N(0,1)', ylab = 'Resíduos', main = 'Gráfico Normal de Probabilidades')
```

```
## Poisson model
```




```
hnp(m2, xlab = 'Percentil da N(0,1)', ylab = 'Resíduos', main = 'Gráfico Normal de Probabilidades')
```

```
## Negative binomial model (using MASS package)
```



Modelo Escolhido

((((((escrever sobre o modelo escolhido etc etc etc))))))

```
summary(m1)
```

```
##
## Call:
## glm(formula = obitos ~ pib + populacao + frota + emergencia,
##     family = "poisson", data = da)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3737  -1.1000   0.3940   0.9598   2.9616
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.72732    1.18758   0.612 0.540249
## pib          0.18433    0.05096   3.617 0.000298 ***
## populacao    0.06813    0.19723   0.345 0.729752
## frota        0.03130    0.12975   0.241 0.809390
## emergencia  -0.19492    0.12619  -1.545 0.122429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 155.43  on 50  degrees of freedom
## Residual deviance: 139.09  on 46  degrees of freedom
##   (61 observations deleted due to missingness)
## AIC: 352.35
##
## Number of Fisher Scoring iterations: 5
```

Reajuste do modelo

```
m2.1  <- glm.nb(obitos ~ populacao + frota, data = da)
summary(m2.1)
```

```
##
## Call:
## glm.nb(formula = obitos ~ populacao + frota, data = da, init.theta = 4.424949916,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0854  -1.0430   0.3019   0.5831   1.1129
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.62026    1.15365   1.404    0.160
## populacao    0.03718    0.24433   0.152    0.879
## frota        0.04152    0.18587   0.223    0.823
##
## (Dispersion parameter for Negative Binomial(4.4249) family taken to be 1)
##
##     Null deviance: 92.969  on 82  degrees of freedom
## Residual deviance: 91.760  on 80  degrees of freedom
##   (29 observations deleted due to missingness)
## AIC: 522.09
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  4.42
##           Std. Err.:  1.09
##
##  2 x log-likelihood:  -514.095
```

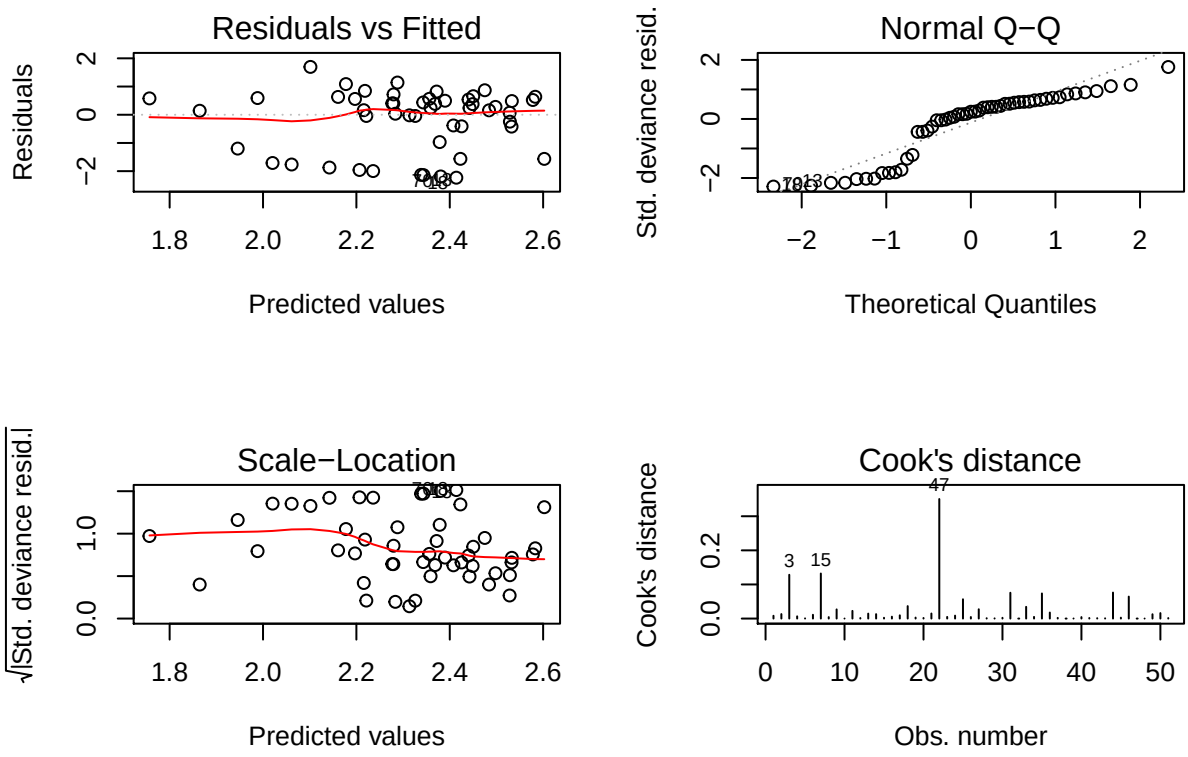
Verificando o modelo ajustado:

```
ajuste = c('m2', 'm2.1')
aic    = c(AIC(m2), AIC(m2.1))
verossimilhanca = c(logLik(m2),logLik(m2.1))
data.frame(ajuste, aic, verossimilhanca)
```

```
##   ajuste      aic verossimilhanca
## 1     m2 324.8689       -156.4345
## 2   m2.1 522.0946       -257.0473
```

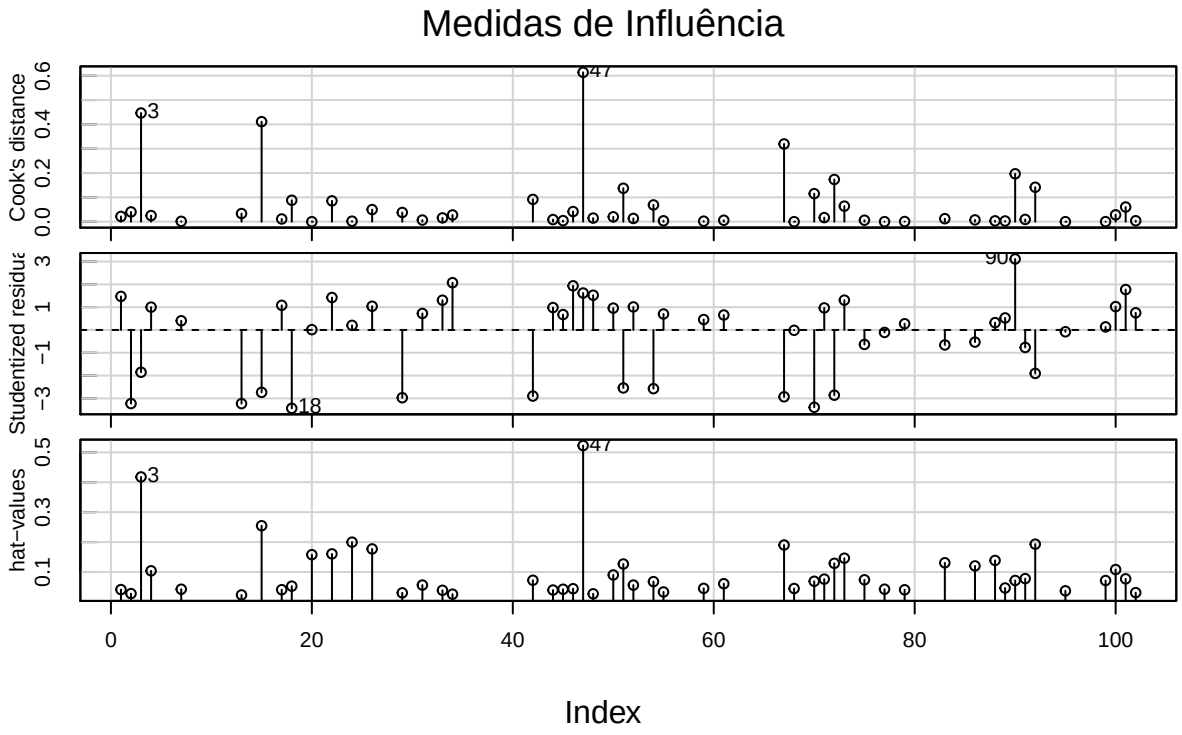
Gráficos de resíduos

```r
par(mfrow=c(2,2))
plot(m2, 1:4)
```




```r
par(mfrow = c(1,1))
```

Medidas de influência



Resíduos Quantílicos Aleatorizados

```
par(mfrow=c(1,2))

res <- qresiduals(m1)

plot(res)

residuos <- qresiduals(m1)
qqnorm(residuos)
qqline(residuos, col = 2)
```

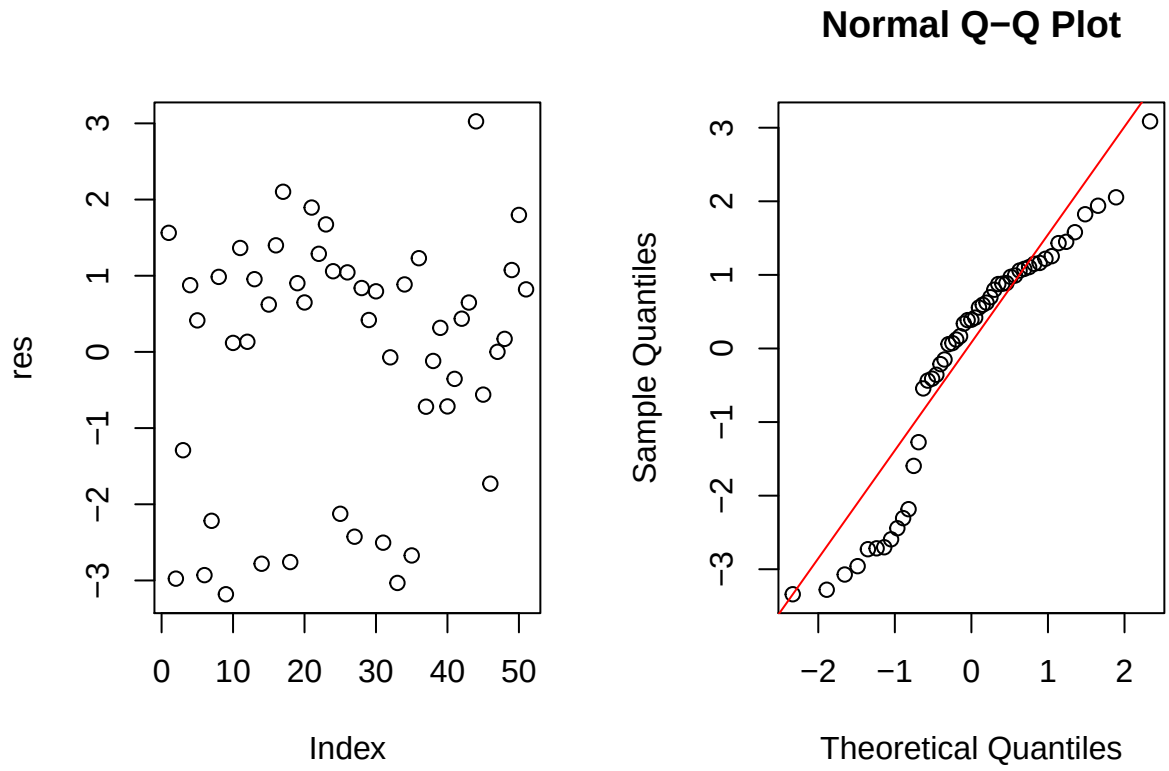
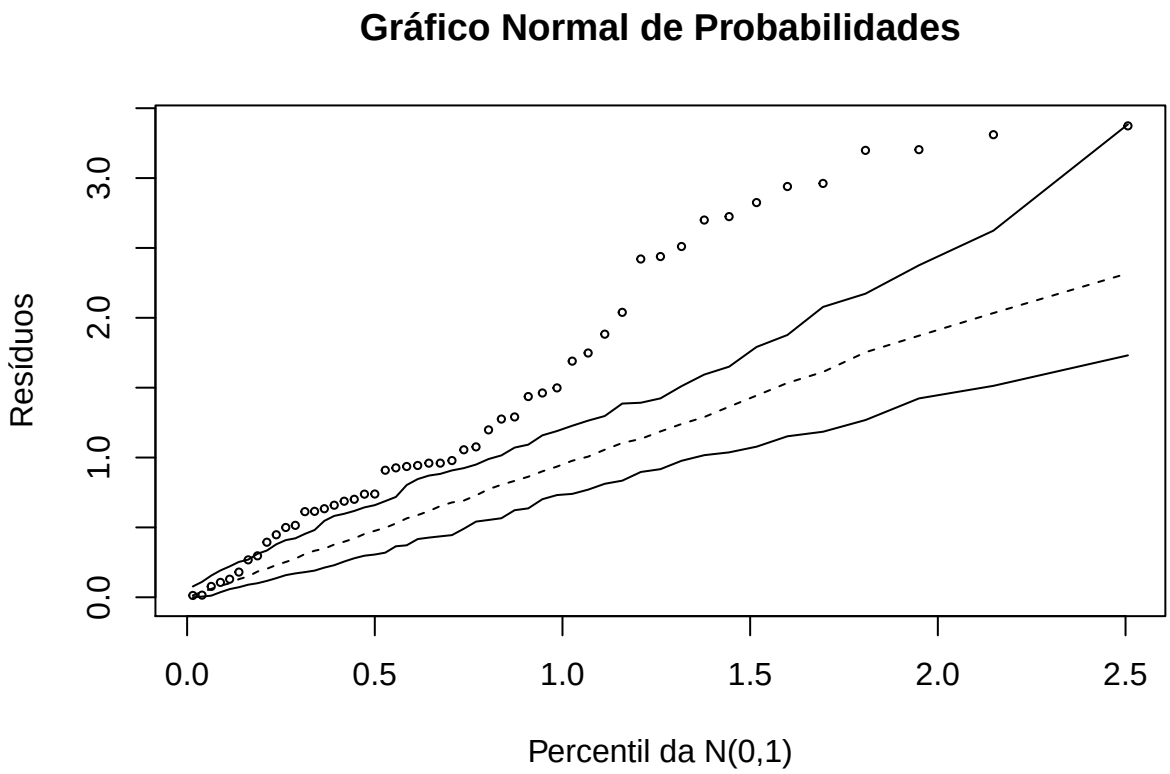


Gráfico Normal de Probabilidades com Envelope Simulado

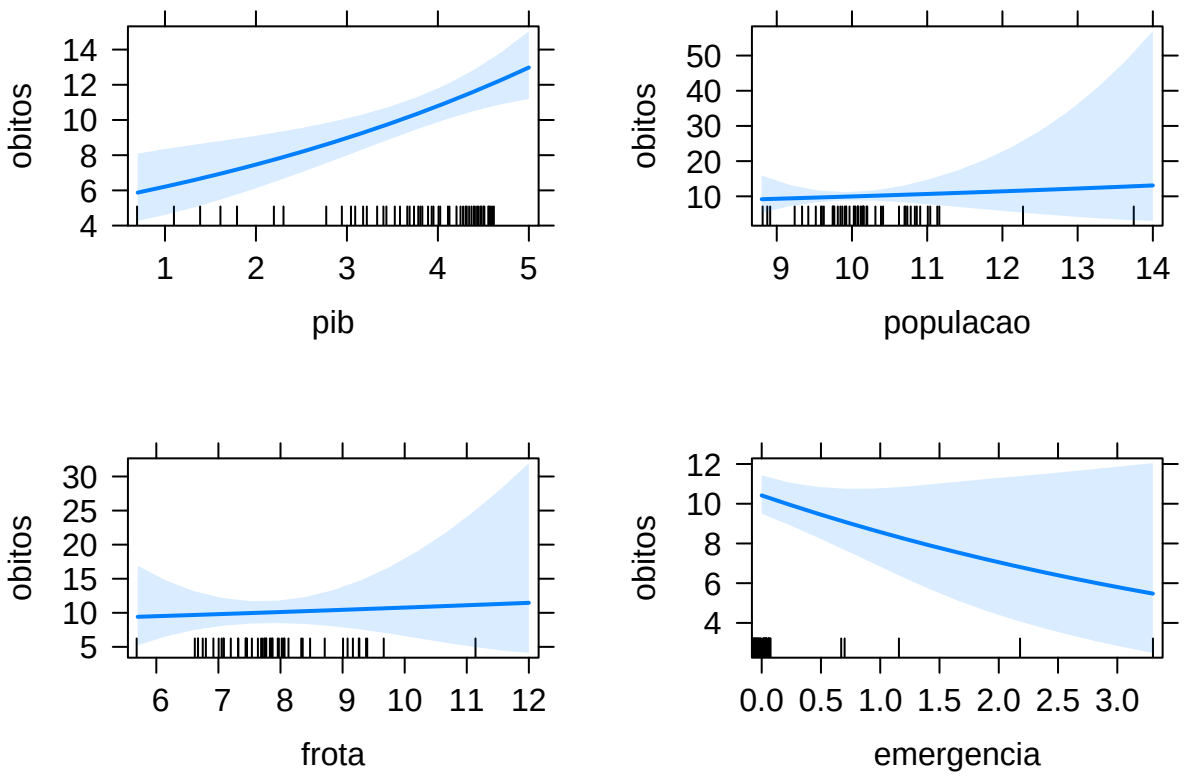
```
par(mfrow=c(1,1))
hnp(m1, xlab = 'Percentil da N(0,1)', ylab = 'Resíduos', main = 'Gráfico Normal de Probabilidades')
```

```
## Poisson model
```



Gráficos de Efeitos

```
plot(allEffects(m1), type = 'response', main = '')
```



PREDIÇÃO ((NÃO SEI FAZE))

AJUSTE E DIAGNOSTICO TBM NAO

Predição

Outra abordagem?? Quase-Verossimilhaça