Untitled

Ananda Bordignon, Brendha Lima, Giovanna Lazzarin

12 de novembro de 2018

Resumo

Em torno do mundo, câncer de mama é o tipo mais comum de câncer em mulheres e é o segundo maior em termos de taxas de mortalidade. O diagnóstico do câncer de mama é obtido quando um caroço anormal é encontrado (por auto exame ou raio-x) ou quando um minúsculo grão de cálcio é encontrado (raio-x).Depois que o caroço suspeito é encontrado, o doutor vai conduzir um diagnóstico para determinar se é cancerígeno, e se for, se ele se espalhou para outras partes do corpo. Este conjunto de dados foi obtido da University of Wisconsin Hospitals, em Madison através do Dr. William H. Wolberg.

Introdução

O objetivo deste trabalho é apresentar uma análise estatística, por meio de um modelo linear generalizado para dados de contagem, em dados referentes ao número de pacientes diagnosticadas ou não com câncer de mama.

O trabalho contém uma breve análise descritiva (para melhor entender a base de dados), ajuste de um modelo buscando explicar a quantidade de diagnósticos positivos em função das covariáveis disponíveis, diagnóstico para verificação se o modelo proposto nas circunstâncias se ajusta bem aos dados disponíveis, comparativo entre as distribuições propostas e quais os eventuais problemas dos dados e do método utilizado para a análise.

Entre as covariáveis disponíveis para explicar o número de visitas há, por exemplo, área, raio,tamanho e espessura do nódulo, etc.

Material e métodos

conjunto de dados

Os dados utilizados para aplicação do modelo linear generalizado provêm de um estudo da University of Wisconsin Hospitals e contém um total de XX observações.

A base de dados contém uma série de covariáveis; as quais tiveram sua significância testada no que diz respeito a sua influência no diagnóstico etc, são elas:

- mean_radius: raio médio da distância do centro ao perímetro;
- mean_texture: textura média, irregularidades (standard deviation of gray-scale values = ANÁLISE DE TEXTURA ESTATÍSTICA http://www.lcad.icmc.usp.br/~jbatista/procimg/2012/textura);
- mean_perimeter: perímetro médio do tumor;
- mean_area: área média;
- mean_smoothness: regularidade média (mean of local variation in radius lengths)
- diagnosis: 1: nódulo maligno; 0: nódulo benigno

Os dados se dispuseram desta forma:

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Table 1: Primeiras linhas da base

raio	textura	perímetro	área	regularidade	diagnostico
17.99	10.38	122.80	1001.0	0.11840	0
20.57	17.77	132.90	1326.0	0.08474	0
19.69	21.25	130.00	1203.0	0.10960	0
11.42	20.38	77.58	386.1	0.14250	0
20.29	14.34	135.10	1297.0	0.10030	0
12.45	15.70	82.57	477.1	0.12780	0

recurso computacionais

O *software R* foi utilizado para ajustar os modelos lineares generalizados aos dados descritos. Os pacotes utilizados para auxílio deste trabalho foram: o pacote car, effects, statmod entre outros.

métodos

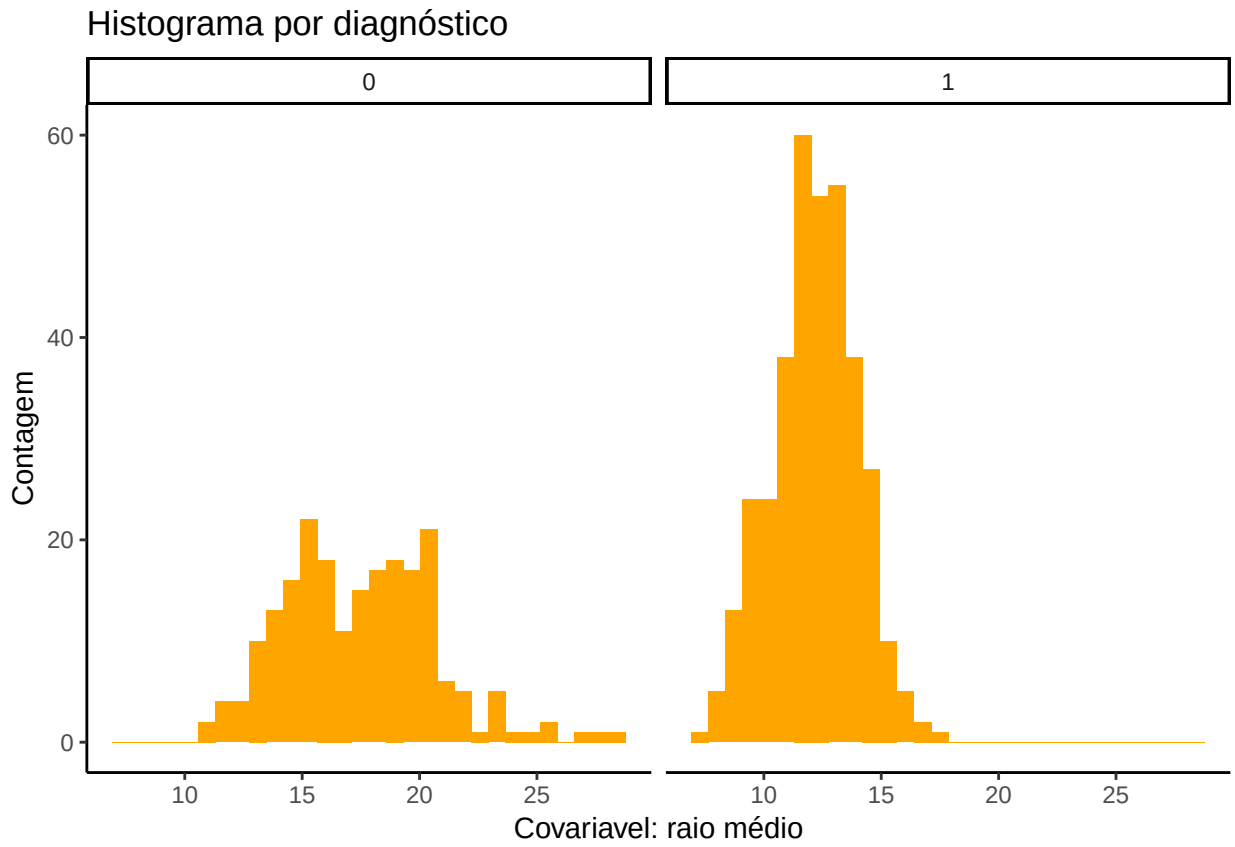
A proposta para modelar o número de diagnósticos foi o modelo linear generalizado aditivo com distribuições Poisson e Binomial negativa, tais modelos são amplamente utilizadas quando a variável de resposta é uma contagem.

Resultados e discussão

análise descritiva

A análise descritiva dos dados, realizada antes dos ajustes dos modelos se dá assim:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Aqui, vê-se que os histogramas são diferentes para as duas classes, aparentemente as pessoas diagnosticada com câncer possuem um raio médio menor do que as que nao tem a doença (e etc).

ajuste dos modelos

Foram testadas 4 funções para modelos lineares generalizados da família Binomial para número o de diagnósticos, são elas: função logit, probit, complemento log-log e Cauchy.

Dizer aqui que a função PROBITO funciona melhor pois têm menor valor AIC

```
##    ajuste      aic    logLik
## 1  logito 174.3758 -81.18792
## 2 probito 173.9882 -80.99411
## 3 cloglog 180.2789 -84.13943
## 4  cauchy 190.1267 -89.06336
```

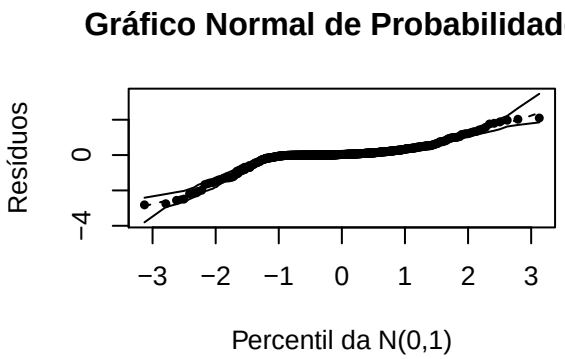


Gráfico Normal de Probabilidades
Resíduos
−4
0
−3
−2
−1
0
1
2
3
Percentil da N(0,1)

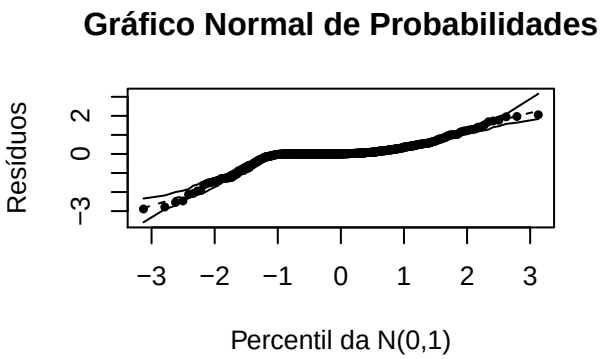


Gráfico Normal de Probabilidades
Resíduos
−3
0
2
−3
−2
−1
0
1
2
3
Percentil da N(0,1)

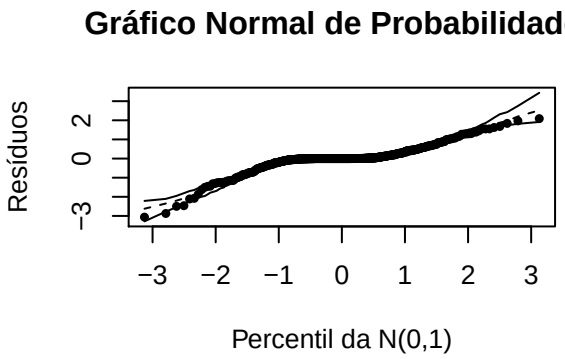


Gráfico Normal de Probabilidades
Resíduos
−3
0
2
−3
−2
−1
0
1
2
3
Percentil da N(0,1)

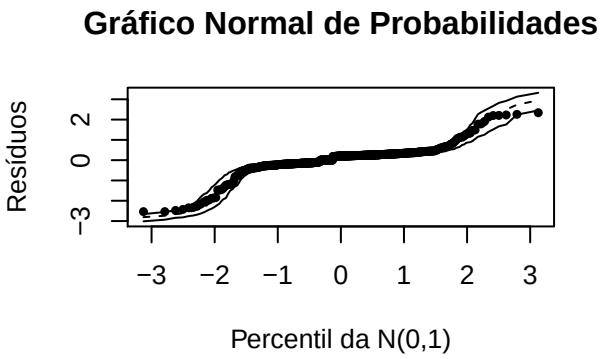
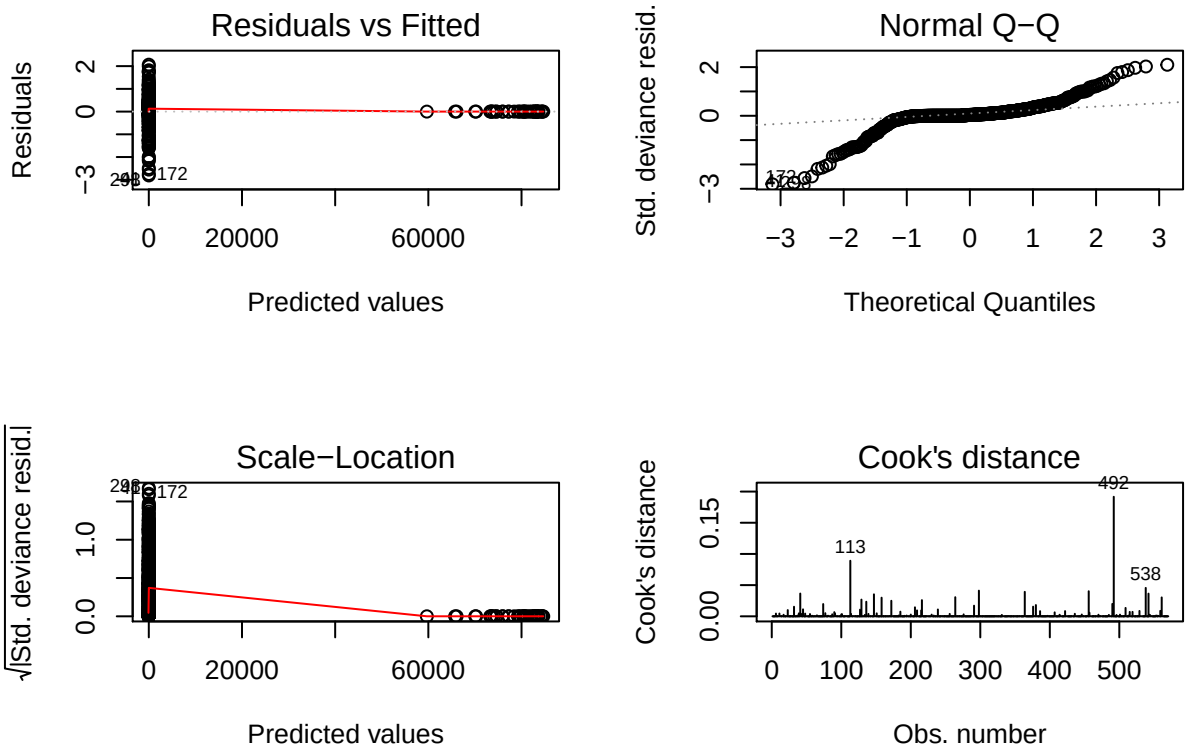
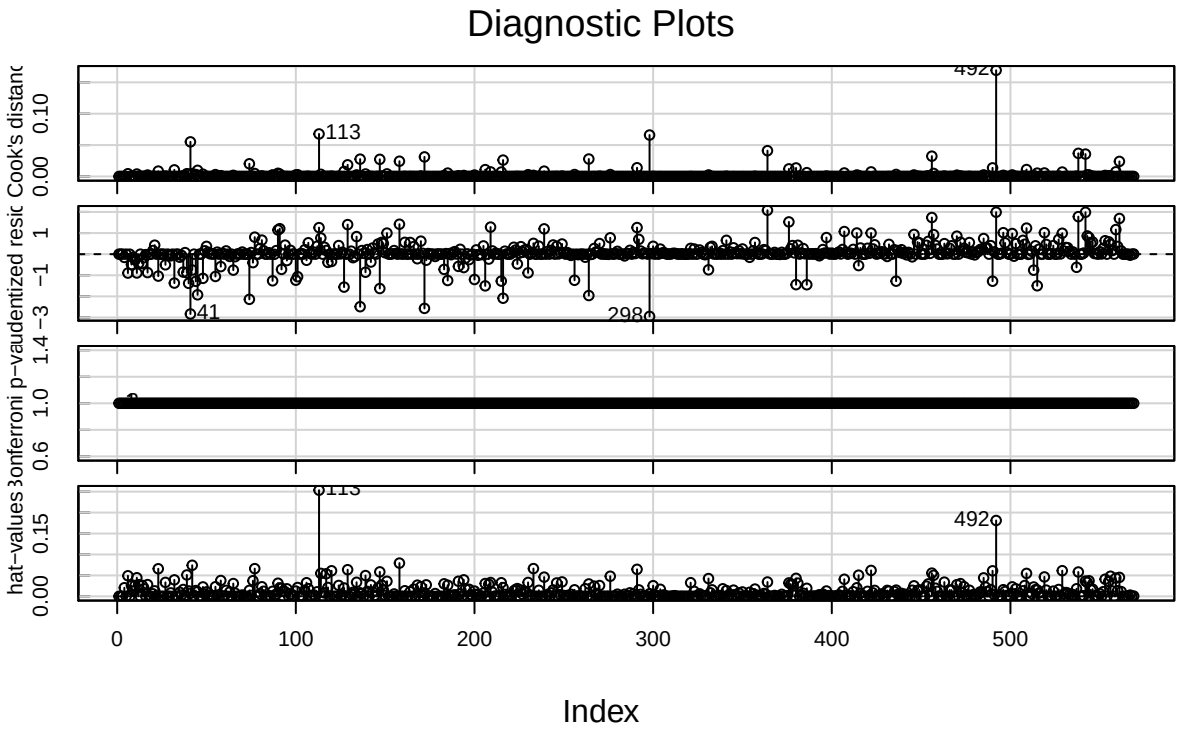


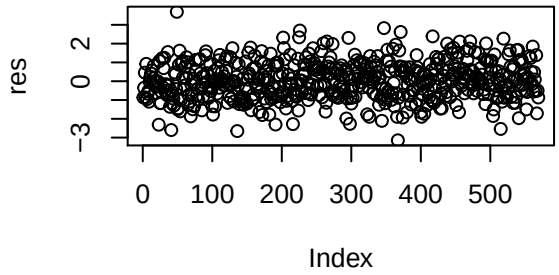
Gráfico Normal de Probabilidades
Resíduos
−3
0
2
−3
−2
−1
0
1
2
3
Percentil da N(0,1)

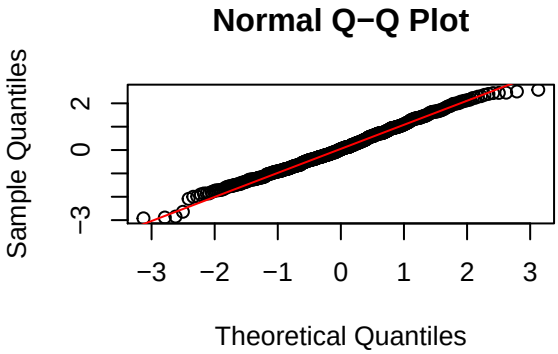



```
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
```




```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.99493, p-value = 0.05809
```


res
−3
0
2
0
100
200
300
400
500
Index


Normal Q−Q Plot
Sample Quantiles
−3
0
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles

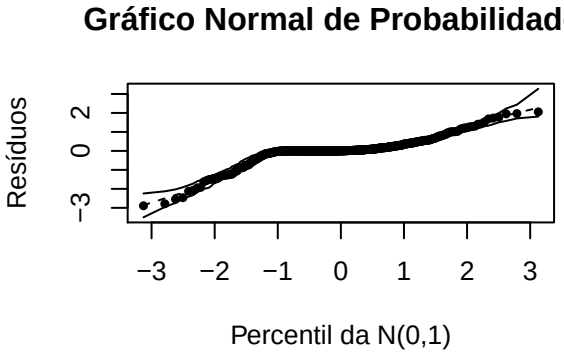
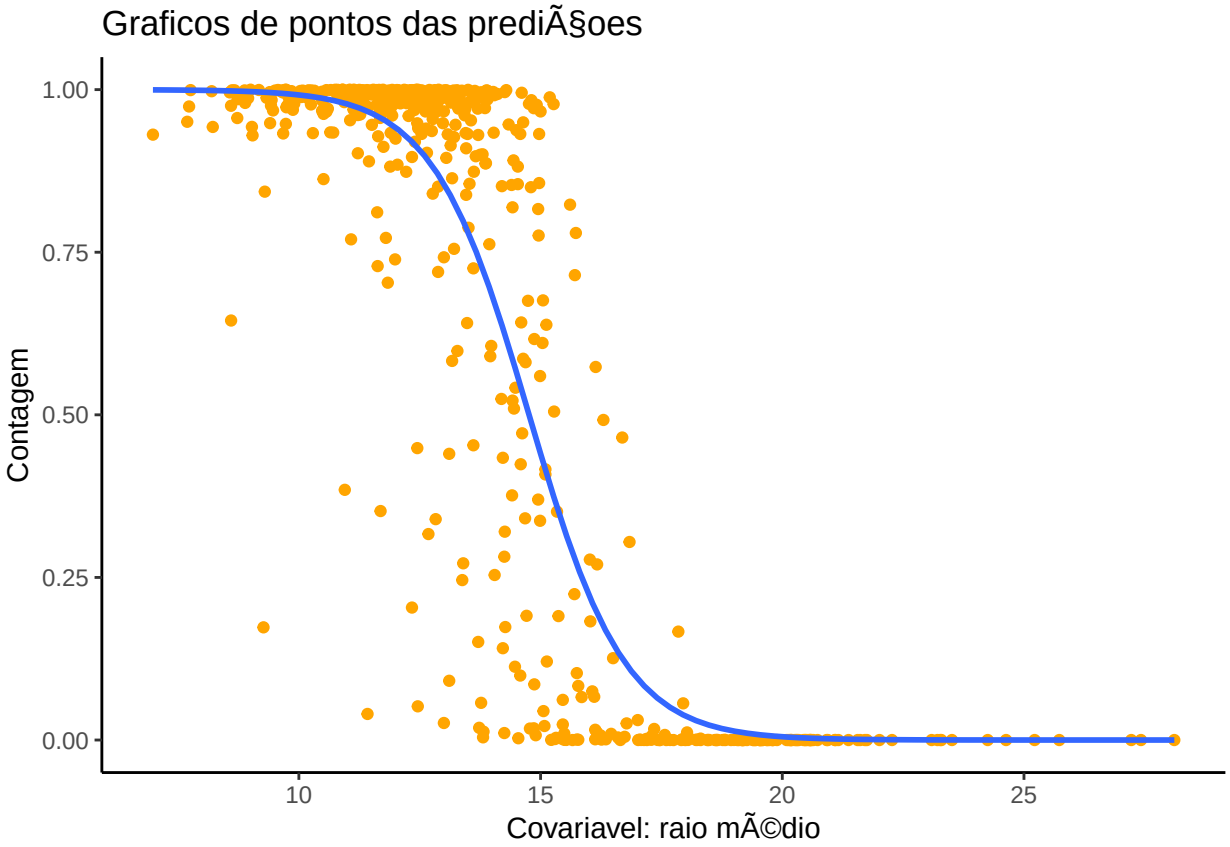
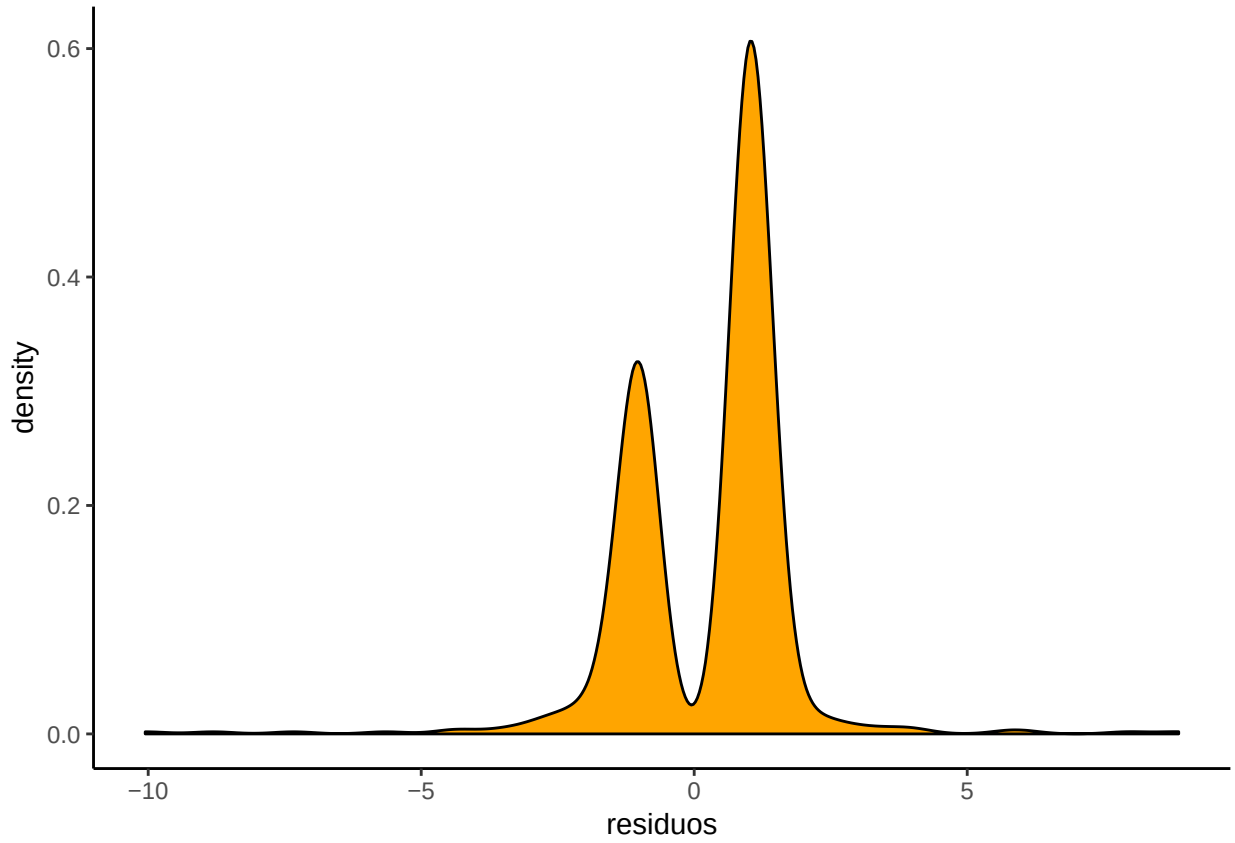


Gráfico Normal de Probabilidades
Resíduos
−3
0
2
−3
−2
−1
0
1
2
3
Percentil da N(0,1)


Graficos de pontos das prediÃ§oes
1.00
0.75
0.50
0.25
0.00
Contagem
10
15
20
25
Covariavel: raio mÃ©dio



Considerações finais