## =}

title: “Trabalho de Modelos Lineares Generalizados” author: “Ananda Bordignon, Brendha Lima, Giovanna Lazzarin” date: “12 de Novembro de 2018” output: pdf\_document —

# 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 Mison através do Dr. William H. Wolberg. Nesses dados, os recursos são calculados a partir de uma imagem digitalizada de um aspirador de agulha fina (PAAF) de uma massa mamária.

# 1. Introdução

O objetivo deste trabalho é apresentar uma análise estatística, por meio de um modelo linear generalizado s, por meio de um modelo linear generalizado para dados binários (ou seja, com variável resposta do tipo dicotômica), dos dados dos 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 diagnósticos positivos há, por exemplo, área, raio, tamanho e espessura do nódulo, etc.

# 2. Material e métodos

## 2.1 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 569 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()

Primeiras linhas da base

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | diagnostico | radius | texture | perimeter |
| 842517 | 1 | 20.57 | 17.77 | 132.90 |
| 84300903 | 1 | 19.69 | 21.25 | 130.00 |
| 84348301 | 1 | 11.42 | 20.38 | 77.58 |
| 84358402 | 1 | 20.29 | 14.34 | 135.10 |
| 843786 | 1 | 12.45 | 15.70 | 82.57 |
| 844359 | 1 | 18.25 | 19.98 | 119.60 |

## 2.2 Recursos 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.

## 2.3 Métodos

A proposta para modelar o número de diagnósticos foi o modelo linear generalizado com distribuição Binomial, pois é comumente utilizada quando a variável de resposta é dicotômica, indicando probabilidade de sucesso ou fracasso.

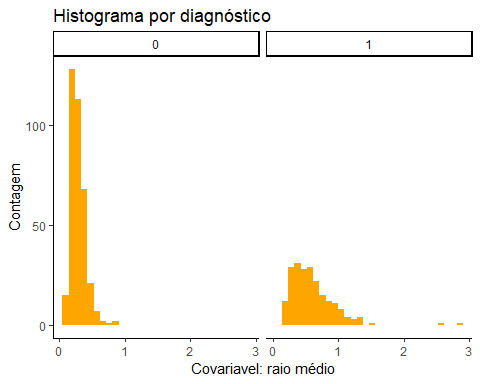
# 3. Resultados e discussão

## 3.1 Análise descritiva

A análise descritiva dos dados, realizada antes dos ajustes dos modelos se dá na seguinte forma:

##   
## 0 1   
## 0.63 0.37

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



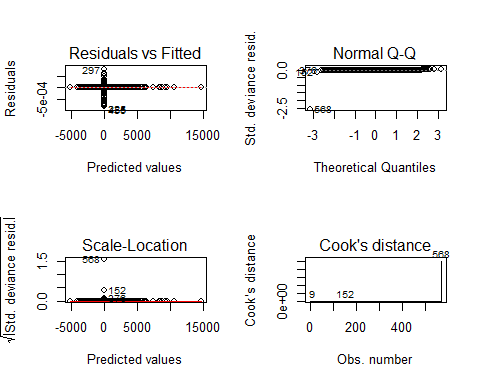
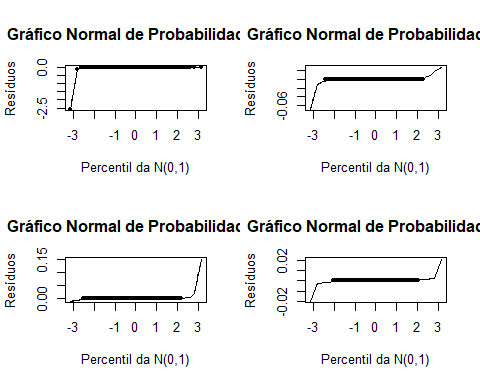
Acima pode se observar que os histogramas apresentam comportamentos diferentes para as duas classes, porque aparentemente as pessoas diagnosticadas com câncer possuem um raio médio menor do que as que não têm a doença (e etc).

## 3.2 Ajuste dos modelos

Foram testadas 4 funções para modelos lineares generalizados da família Binomial para modelar o número 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 64.00001 -3.844075e-06  
## 2 probito 784.87307 -3.604365e+02  
## 3 cloglog 856.96037 -3.964802e+02  
## 4 cauchy 1599.69922 -7.678496e+02

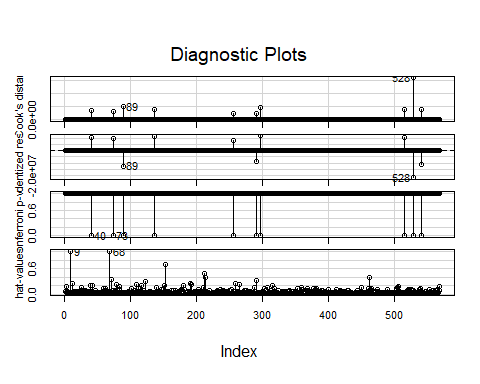


## Loading required package: carData

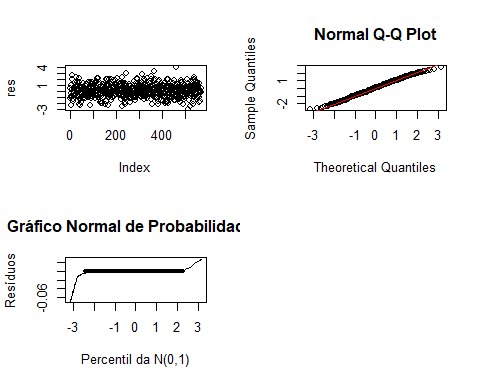
##   
## Attaching package: 'car'

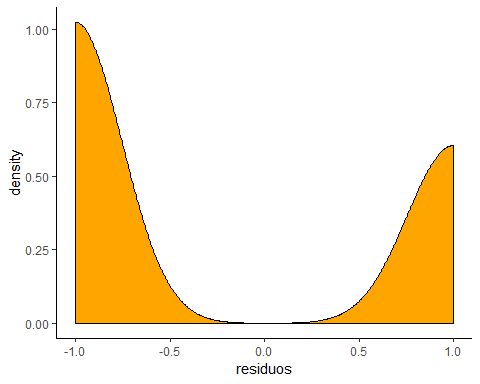
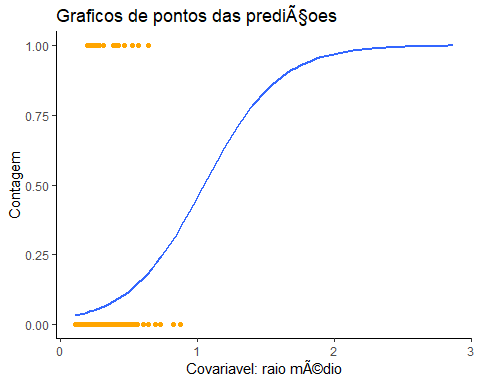
## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some



##   
## Shapiro-Wilk normality test  
##   
## data: residuos  
## W = 0.99629, p-value = 0.2082





# 4. Considerações finais