Trabalho de Modelos Lineares Generalizados

Ananda Bordignon[*]*, Brendha Lima*[†]*, Giovanna Lazzarin*[‡]

12 de Novembro de 2018

Resumo

O objetivo deste trabalho é apresentar uma análise estatística, por meio de um modelo linear generalizado 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. As covariáveis são um conjunto de medidas retiradas de nódulos da mama, e envolvem raio, área, perímetro, textura, suavidade, compacicidade, pontos côncavos, simetria, concavidade e dimensão fractal, todos em forma de média por conta das irregularidades presentes nos caroços. Foram avaliados quais era os comportamentos de cada covariável quando o diagnóstico era de nódulo com tecido maligno para observar sua coerência e ligação direta com a resposta, houve estudo para selecionar qual seria a melhor função de ligação a ser utilizada entre as possíveis para a Distribuição Binomial e vários ajustes das covariáveis melhor descritos ao longo do documento. A seleção final indica que

1. Introdução

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.

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:

[*]GRR20149157
[†]GRR20149163
[‡]GRR20149088

1. Mean_radius: raio médio da distância do centro ao perímetro;
2. Mean_texture: textura média, irregularidades;
3. Mean_perimeter: perímetro médio do tumor;
4. Mean_area: área média;
5. Mean_smoothness: regularidade média (média da variação local);
6. Mean_compactness: compacicidade média;
7. Mean_concavity: concavidade média;
8. Mean_concave_points: média de pontos côncavos;
9. Mean_symmetry: simetria média;
10. Mean_fractal_dimension: média da dimensão fractal;
11. Diagnóstico (variável resposta): M para nódulo maligno (câncer) e B para nódulo benigno.

Os dados se dispuseram desta forma:

Table 1: Tabela 1 - Primeiras observações da base de dados

Diagnóstico	Raio médio	Textura média	Perím. médio	Área média	Suav. média
M	1.0950	0.9053	8.589	153.40	0.006399
M	0.5435	0.7339	3.398	74.08	0.005225
M	0.7456	0.7869	4.585	94.03	0.006150
M	0.4956	1.1560	3.445	27.23	0.009110
M	0.7572	0.7813	5.438	94.44	0.011490
M	0.3345	0.8902	2.217	27.19	0.007510

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, ggplot, gridExtra, entre outros.

2.3 Métodos

A proposta para modelar o número de diagnósticos foi o modelo linear generalizado com distribuição Binomial, pois é aplamente utilizada quando a variável de resposta é dicotômica. A construção do modelo Binomial se dá através de múltiplas repetições de Bernoulli e é caracterizada como a distribuição de probabilidades discreta do número de ocorrências de algum evento numa sequência de tentativas, tendo n ensaios realizados e k ocorrências do evento (com k=1, ..., n), pode-se expressar a probabilidade de sucesso conforme a fórmula abaixo.

$$\mathbb{P}[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

Será também utilizado uma função de ligação escolhida entre: logito, probito, complemento log-log e cauchy, que são funções de ligações possíveis para a distribuição Binomial.

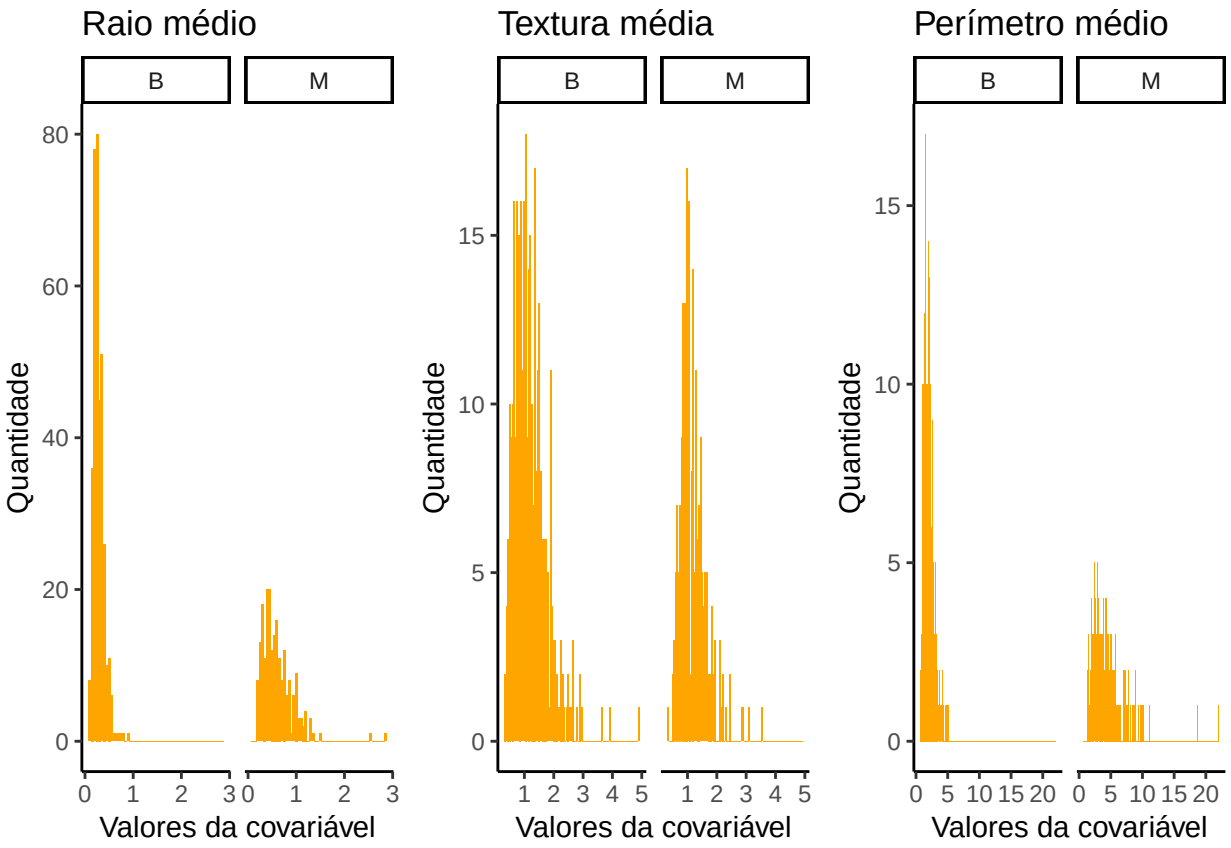
3. Resultados e discussão

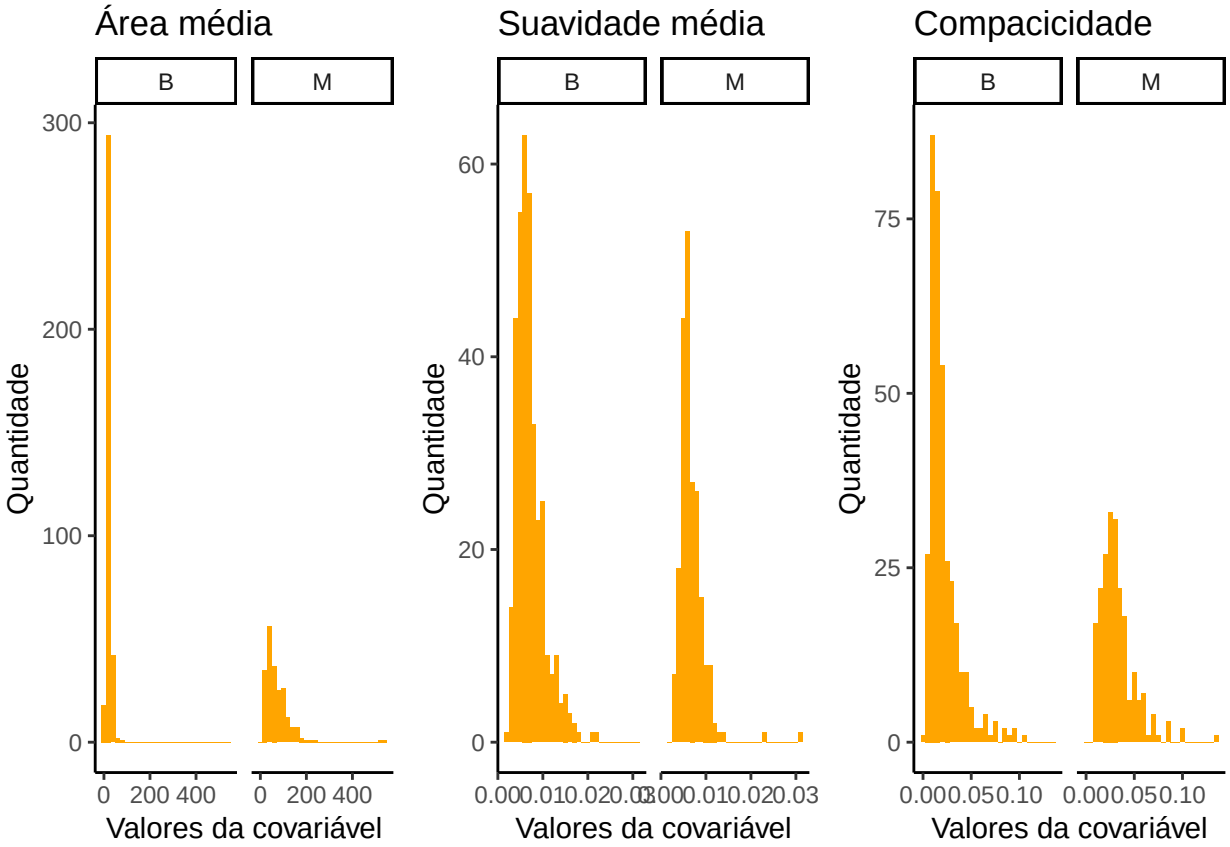
3.1 Análise descritiva

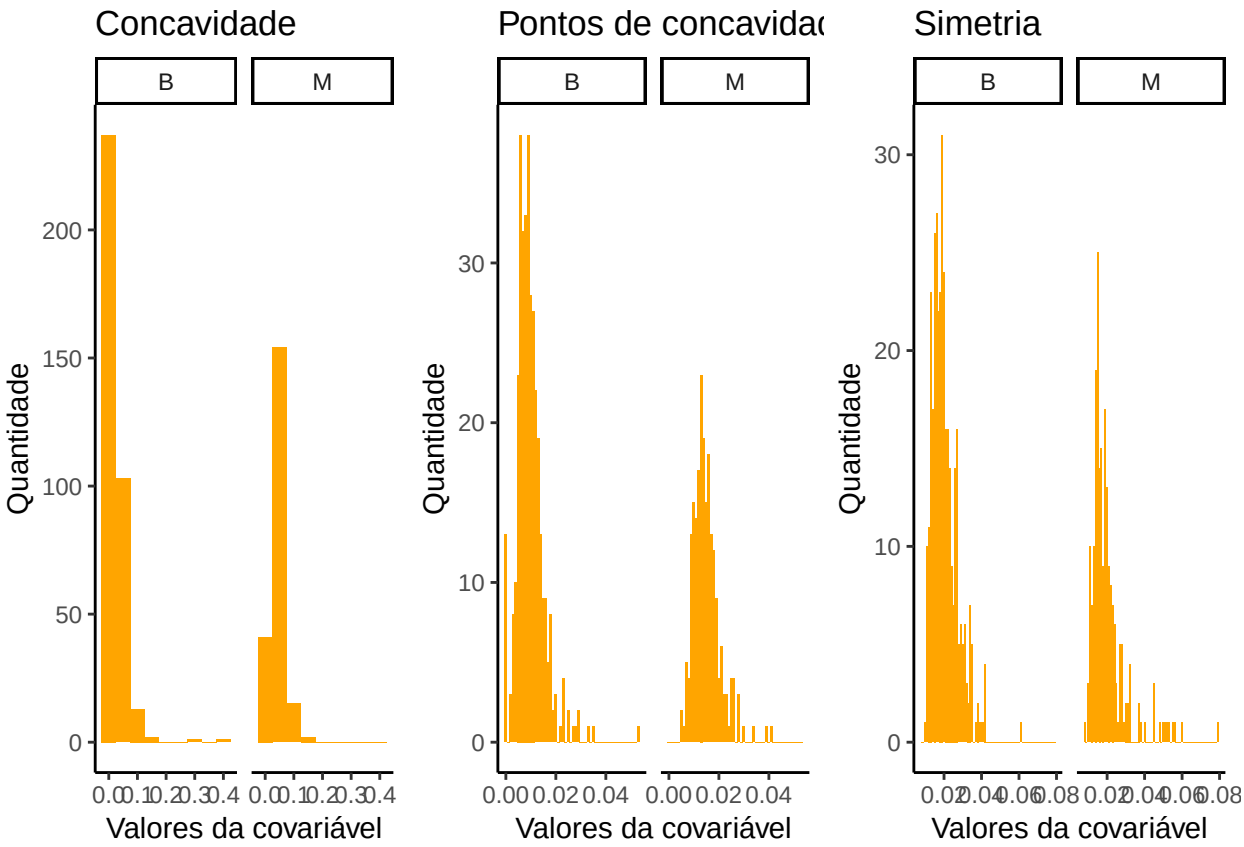
A análise descritiva dos dados, realizada antes dos ajustes dos modelos, teve como objetivo observar qual é o comportamento de todas as possíveis covariáveis em relação à variável resposta e se dá na seguinte forma:

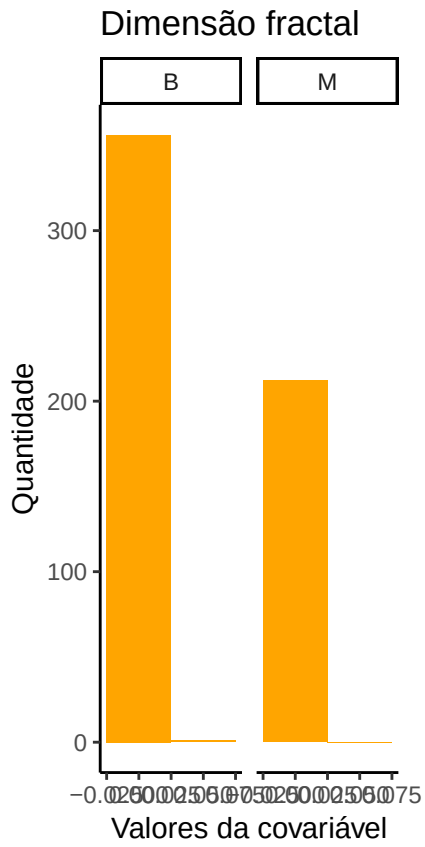
Table 2: Tabela 2 - Proporção de diagnósticos

Diag.	Prop.
B	0.63
M	0.37




Área média
Suavidade média
Compacicidade
B
M
Quantidade
Valores da covariável
0
100
200
300
20
40
60
25
50
75
0 200 400
0.00 0.01 0.02 0.03
0.000 0.050 0.10


Concavidade
Pontos de concavidad
Simetria
B
M
Quantidade
Valores da covariável
0
50
100
150
200
10
20
30



Acima pode se observar que os histogramas apresentam comportamentos diferentes para as duas classes, como por exemplo, pessoas diagnosticadas com câncer (tumor maligno) possuem um raio médio maior do que as que não têm a doença, tendendo a ser mais compactos e com perímetro maior do que aqueles que são benignos.

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

TABELA DEVIANCE

Table 3: Tabela 3 - Resultados dos ajustes

ajuste	aic	logLik
logito	287.9175	-132.9588
probito	286.7384	-132.3692
cloglog	295.0887	-136.5443
cauchy	300.0591	-139.0295

```
##
## Call:
## glm(formula = diagnostico ~ ., family = binomial(link = (link = "logit")),
##     data = dados1)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.55359  -0.39808  -0.17839   0.00744   3.12609
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.39225    0.63580  -5.335 9.53e-08 ***
## mean_radius             -28.22915    6.13786  -4.599 4.24e-06 ***
## mean_texture             -0.33386    0.42301  -0.789   0.4300
## mean_perimeter            0.07511    0.47621   0.158   0.8747
## mean_area                 0.39346    0.05710   6.891 5.55e-12 ***
## mean_smoothness           4.19959   98.69183   0.043   0.9661
## mean_compactness         57.25377   28.68811   1.996   0.0460 *
## mean_concavity            3.61415   11.95153   0.302   0.7623
## mean_concave_points      39.76686   64.17408   0.620   0.5355
## mean_symmetry            -6.76193   30.45752  -0.222   0.8243
## mean_fractal_dimension -318.73330  149.29356  -2.135   0.0328 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 265.92  on 558  degrees of freedom
## AIC: 287.92
##
## Number of Fisher Scoring iterations: 8
```

```
##
## Call:
## glm(formula = diagnostico ~ ., family = binomial(link = (link = "probit")),
##     data = dados1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48183  -0.39691  -0.13683   0.00013   3.00487
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.01768    0.34441  -5.858 4.68e-09 ***
## mean_radius             -14.83774    3.24042  -4.579 4.67e-06 ***
## mean_texture             -0.21580    0.22920  -0.942   0.3464
## mean_perimeter            0.02131    0.27123   0.079   0.9374
## mean_area                 0.21192    0.02882   7.354 1.93e-13 ***
## mean_smoothness          -9.97025   52.05094  -0.192   0.8481
## mean_compactness         36.94481   16.04071   2.303   0.0213 *
## mean_concavity            0.24138    6.46178   0.037   0.9702
## mean_concave_points      27.75452   34.90036   0.795   0.4265
## mean_symmetry             1.86956   15.92477   0.117   0.9065
## mean_fractal_dimension -198.01290   83.04106  -2.385   0.0171 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 264.74  on 558  degrees of freedom
## AIC: 286.74
##
## Number of Fisher Scoring iterations: 9
```

Ao realizarmos teste comparativo de AIC (Critério de Informação de Akaike), ambos links logito e probito obtiveram resultado praticamente igual (287,9 e 285,8) e c log-log e cauchy tiveram valores mais altos (295 e 300). Através da análise de Deviance, a qualidade do ajuste também é muito próxima (-132,95 e -132,37), optou-se então pelo link logito por ser mais simples de leitura e mais comumente utilizado.

Ao assumir que a variável resposta tem distribuição de probabilidade binomial, para adequar a resposta média ao modelo linear será usada a função de ligação logito:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}},\ i = 1, \ldots, n,$$

Com n= 569 observações, xi como valor da variável explicativa e yi número de ocorrências do evento. A função de ligação também pode ser escrita na forma abaixo:

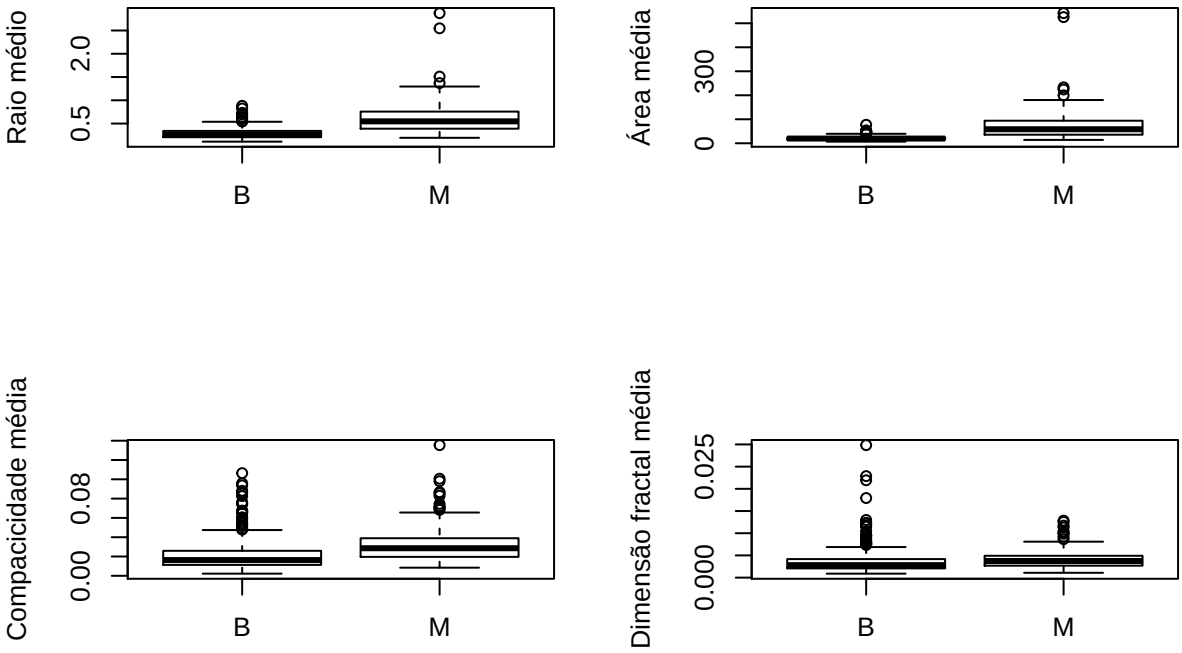
$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

Ajuste 1 - Modelo utilizando Método Stepwise

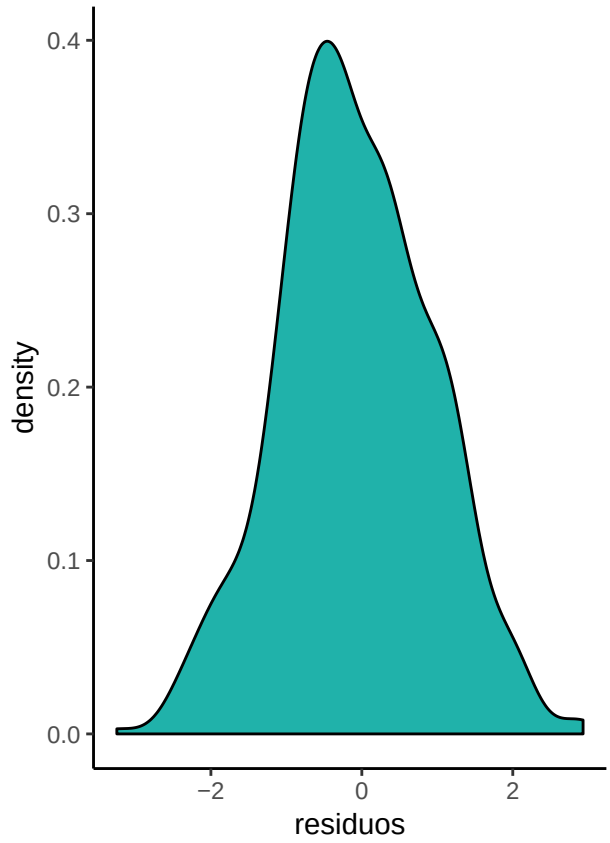
```
##
## Call:
## glm(formula = diagnostico ~ mean_area + mean_radius + mean_compactness +
##     mean_fractal_dimension, family = "binomial", data = dados1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.58130  -0.39905  -0.18599   0.00732   3.00982
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.5447     0.4760  -7.447 9.56e-14 ***
## mean_area                  0.4039     0.0501   8.063 7.45e-16 ***
## mean_radius              -28.7337     4.6288  -6.208 5.38e-10 ***
## mean_compactness          74.3701    20.4062   3.644 0.000268 ***
## mean_fractal_dimension -356.1314   146.6835  -2.428 0.015187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 267.34  on 564  degrees of freedom
## AIC: 277.34
##
## Number of Fisher Scoring iterations: 8
```

Após realizar os ajustes de modelos utilizando o método Stepwise, a seleção final de variáveis significativas, que mais ajudam a explicar se o tipo do nódulo é maligno ou benigno são: raio médio, área média, compacicidade

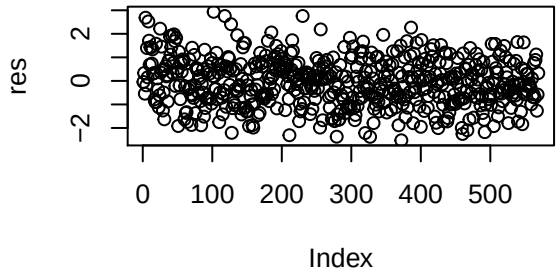
e dimensão fractal, cujos comportamentos podem ser observados abaixo:

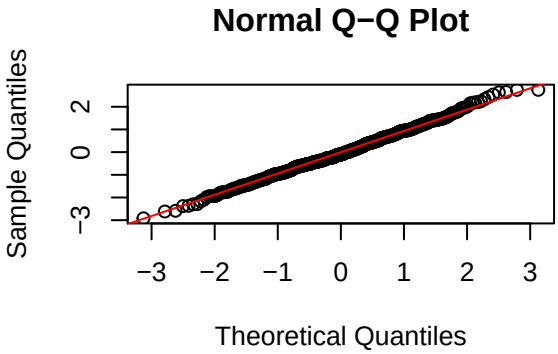


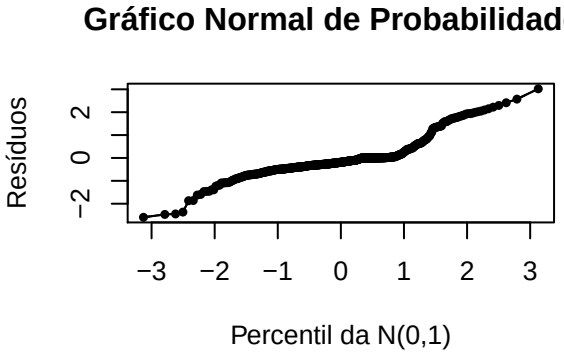
Qualidade do ajuste A seguir temos alguns métodos para avaliar se este ajuste de modelo é o ideal, e será analisado através gráficos de resíduos, com sua dispersão e através de envelope.




```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.99721, p-value = 0.4479
```



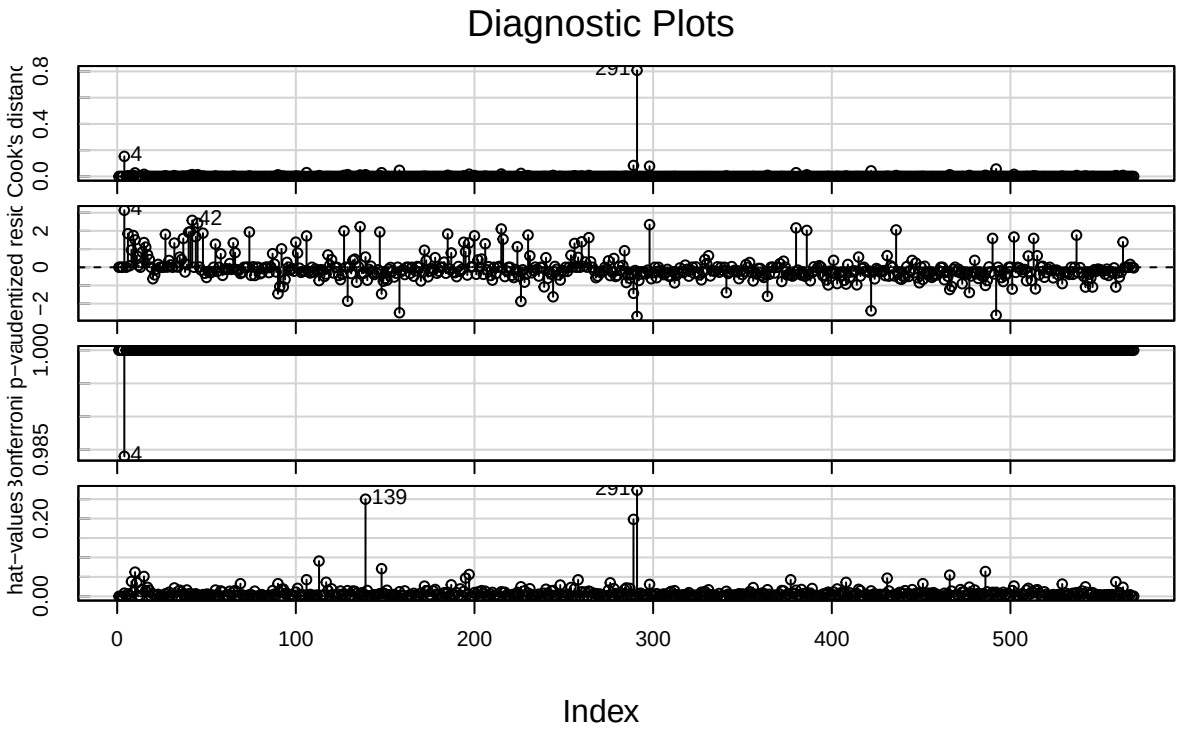




A variância dos resíduos aproximadamente homogênea, gráfico quantil quantil bem ajustado (apresenta linearidade dos dados) e teste de Shapiro Wilk corrobora com a teste de normalidade dos resíduos.

Pontos influentes

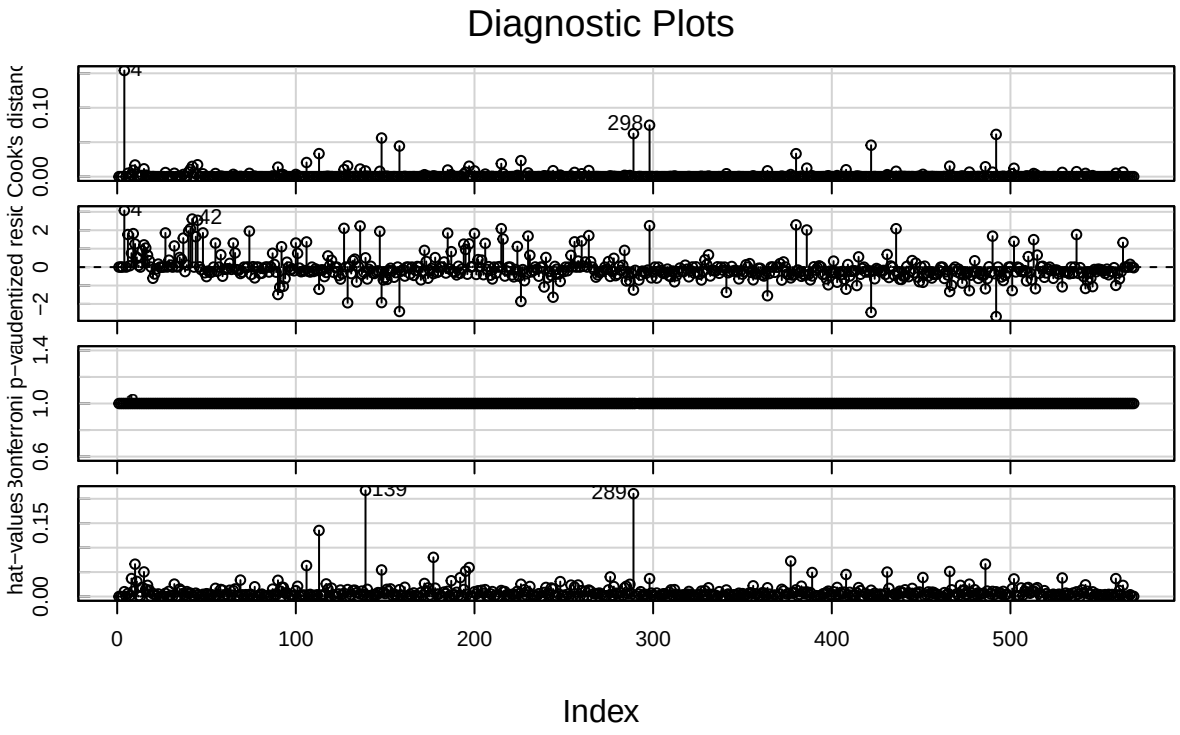
Utilizou-se distância de Cook para identificar pontos de alavancagem, ou seja, aqueles pontos extremos que podem estar interferindo na estimação de coeficientes da regressão.




```
## Calls:
## 1: glm(formula = diagnostico ~ mean_area + mean_radius +
##   mean_compactness + mean_fractal_dimension, family = "binomial", data =
##   dados1)
## 2: glm(formula = diagnostico ~ mean_area + mean_radius +
##   mean_compactness + mean_fractal_dimension, family = "binomial", data =
##   dados1[-which(rownames(dados1) == "291"), ])
##
##                         Model 1 Model 2
## (Intercept)              -3.545  -4.058
## SE                        0.476   0.523
##
## mean_area                0.4039  0.4227
## SE                       0.0501  0.0521
##
## mean_radius              -28.73  -29.65
## SE                         4.63    4.72
##
## mean_compactness           74.4    56.5
## SE                         20.4    20.5
##
## mean_fractal_dimension    -356    -142
## SE                         147     152
##

##
## Call:
```

```
## glm(formula = diagnostico ~ mean_area + mean_radius + mean_compactness +
##     mean_fractal_dimension, family = "binomial", data = dados1[-which(rownames(dados1) ==
##     "291"), ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.63046  -0.38161  -0.18688   0.00541   2.94687
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.05835    0.52309  -7.758 8.60e-15 ***
## mean_area                  0.42269    0.05211   8.111 5.01e-16 ***
## mean_radius              -29.64975    4.72214  -6.279 3.41e-10 ***
## mean_compactness          56.54148   20.51543   2.756  0.00585 **
## mean_fractal_dimension -141.49874  151.68599  -0.933  0.35090
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 750.51  on 567  degrees of freedom
## Residual deviance: 260.16  on 563  degrees of freedom
## AIC: 270.16
##
## Number of Fisher Scoring iterations: 8
```



!!!!!!!!!!!!!! Achar quais são eles, quais variáveis possuem. Realizar um ajuste com e outro sem, para comparar coeficientes. Pode ser que valores extremos naturalmente façam parte do conjunto de dados, temos que argumentar.

FÓRMULA FINAL (MUDAR PARÂMETROS!!!!!!!!!!!!!!!!!!!!!!!!!!!!)

Escala do preditor: ????????????(????ˆ) = 1,323 ??? 0,059. ????2 ??? 0,823. ????31 ??? 0,823. ????32 + 0,021. ????4 + 0,012. ????5 + + 0,237. ????8 Escala da resposta: ????ˆ = ???? 1,323???0,059.????2???0,823.????31???0,823.????32+0,021.????4+0,012.????5+0,013.????6+0,237.????8

4. Considerações finais

Em um primeiro momento, foi observado que a área, raio, compacicidade e dimensão fractal médios estavam diretamente ligados ao fato do nódulo mamário ser benigno ou maligno, portanto de todas as 10 covariáveis possíveis, apenas 4 entraram no modelo ideal.

O gráfico de resíduos versus valores ajustados e o gráfico normal de probabilidades com envelope simulado não apresentaram problemas, sendo isto considerado uma evidência de bom ajuste do modelo.

Após análise de pontos influentes, foi ajustado um modelo retirando estas observações que eram extremas e chegou-se a conclusão de que sua retirada não mudava muito as estimativas dos coeficientes da regressão (RODAR E OBSERVAR SE INFLUENCIA OU NÃO E COLOCAR AQUI).

A conclusão final é de que quanto maior a área média, raio médio e compacicidade média, maior a probabilidade do nódulo ser maligno. Além disso, para o aumento de XXXX unidades, em média, se aumenta em XXX a probabilidade de malignidade.