

Universidade Federal do Paraná

Trabalho de Modelos Lineares Generalizados

Ananda Bordignon¹, Brendha Lima², Giovanna Lazzarin³

12 de Novembro de 2018

¹ GRR20149157

² GRR20149163

³ GRR20149088

SUMÁRIO

X. Resumo.....	3
1. Introdução.....	4
2. Material e métodos.....	4
2.1 Conjunto de dados.....	4
2.2 Recursos computacionais.....	5
2.3 Métodos	5
3. Resultados e discussão.....	6
3.1 Análise descritiva.....	6
3.2 Ajuste dos modelos.....	9
4. Considerações finais.....	14

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, compacidade, 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.

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:

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: compacidade 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:

Tabela 1 – Seis primeiras linhas 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 amplamente 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 $k=1, \dots, n$), pode-se expressar a probabilidade de sucesso conforme a fórmula abaixo.

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}.$$

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

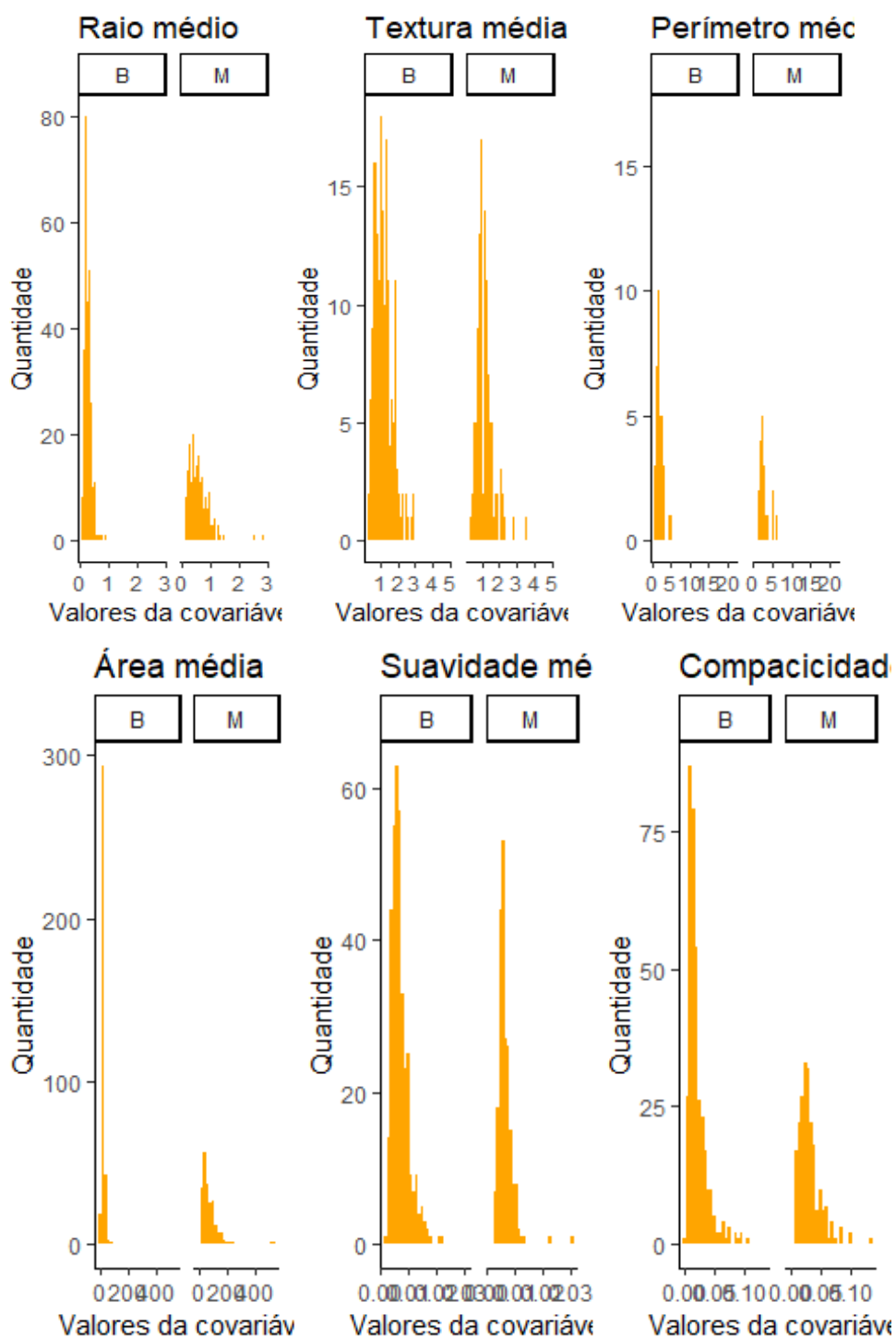
3.1 Análise descritiva

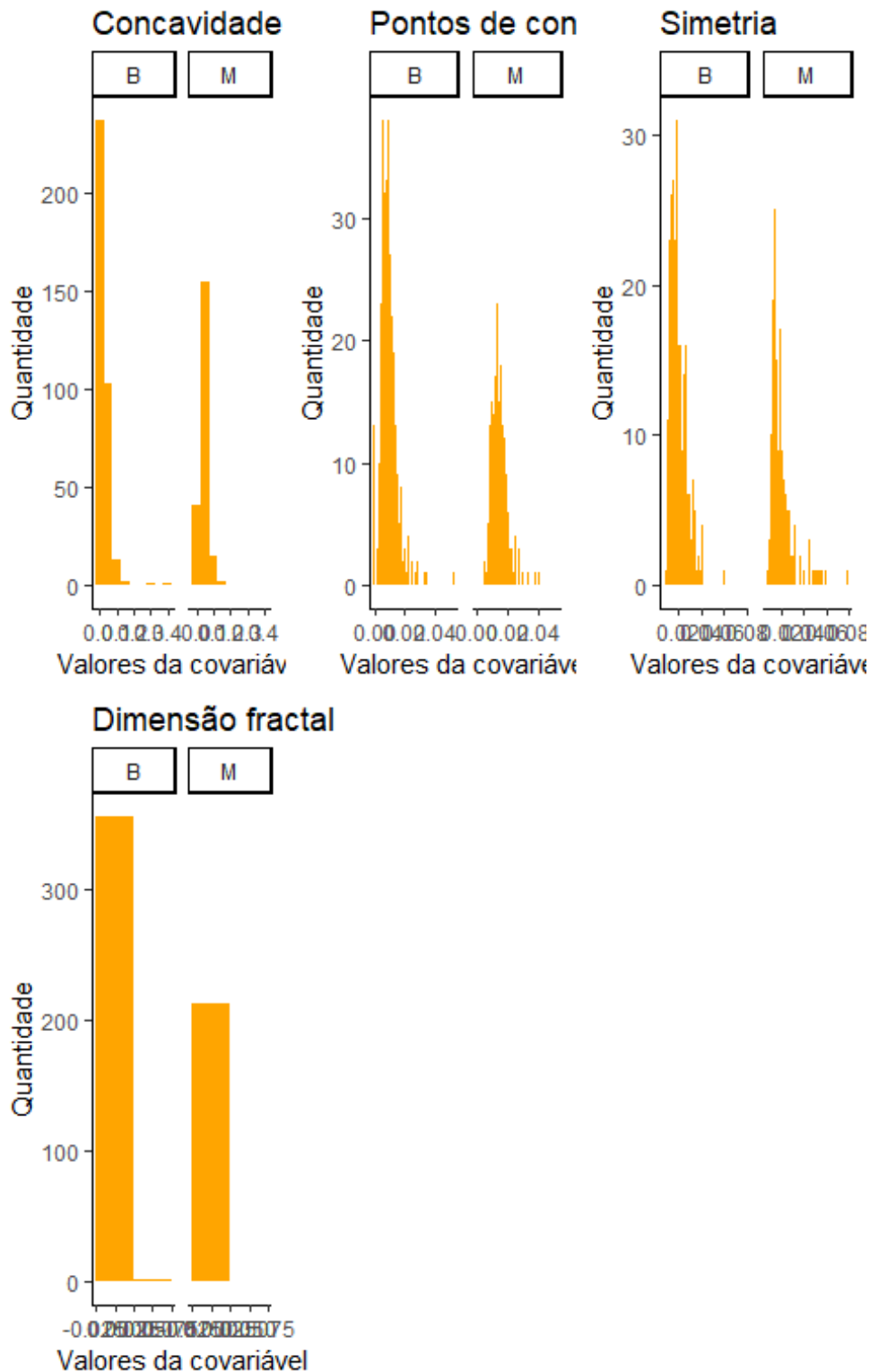
A análise descritiva dos dados, realizada antes dos ajustes dos modelos, teve como objetivo observar qual é o comportamento de todas as possíveis covariáveis em relação à variável resposta e se dá na seguinte forma:

Tabela 2 - Proporção de diagnósticos

Diag.	Prop.
B	0.63
M	0.37

Gráfico 1 – Histogramas descritivos





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. A seguir alguns resultados.

Tabela 3 - Resultados dos ajustes

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

Ao realizarmos o 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:

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

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

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

Quadro 1 - Ajuste do modelo utilizando Método Stepwise

```

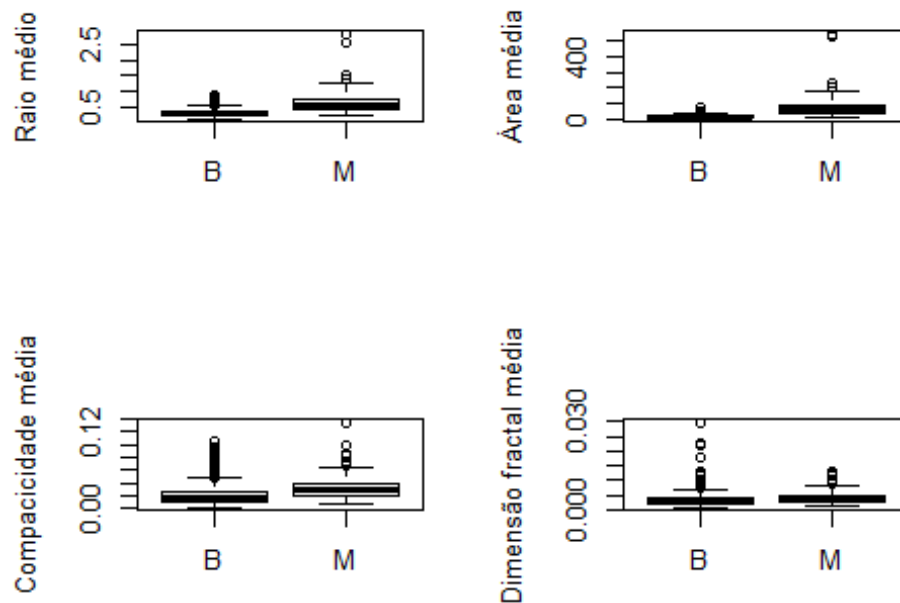
Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.5447     0.4760  -7.447 9.56e-14 ***
## mean_area       0.4039     0.0501   8.063 7.45e-16 ***
## mean_radius    -28.7337     4.6288  -6.208 5.38e-10 ***
## mean_compactness  74.3701    20.4062   3.644 0.000268 ***
## 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, compacidade e dimensão fractal, cujos comportamentos podem ser observados a seguir:

Gráfico 2 – Box Plot das variáveis selecionadas



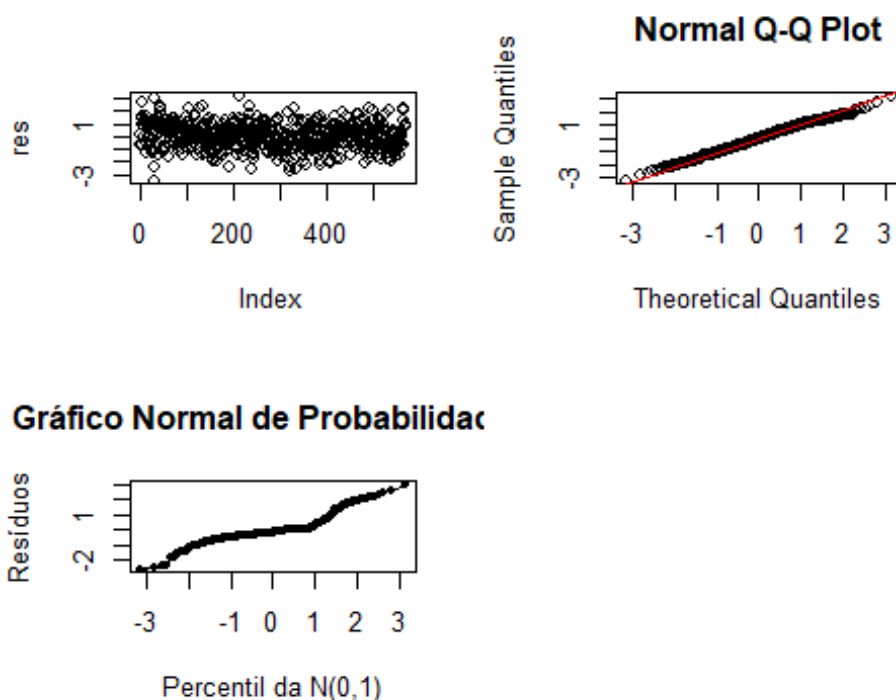
Qualidade do ajuste

A seguir temos alguns métodos para avaliar se este ajuste de modelo é o ideal, e será analisado através gráficos de resíduos.

Quadro 3 – Resultado do teste de Shapiro-Wilk

```
##  
## Shapiro-Wilk normality test  
##  
## data:  residuos  
## W = 0.99647, p-value = 0.2422
```

Gráfico 3 – Q-Q plot do modelo

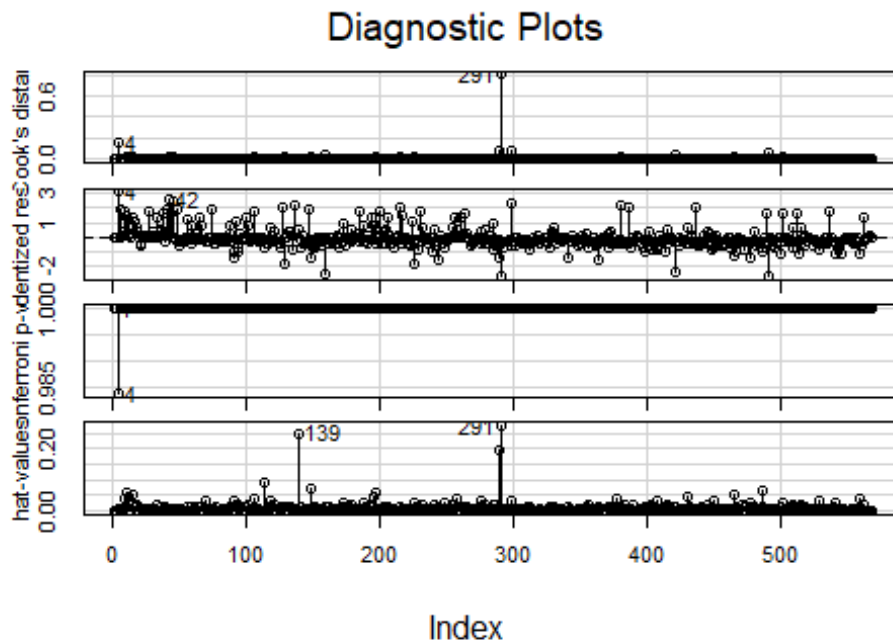


A variância dos resíduos aproximadamente homogênea, gráfico quantil quantil bem ajustado (apresenta linearidade dos dados) e teste de Shapiro Wilk corrobora com a teste de normalidade dos resíduos.

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.

Gráfico 4 – Gráfico de diagnóstico



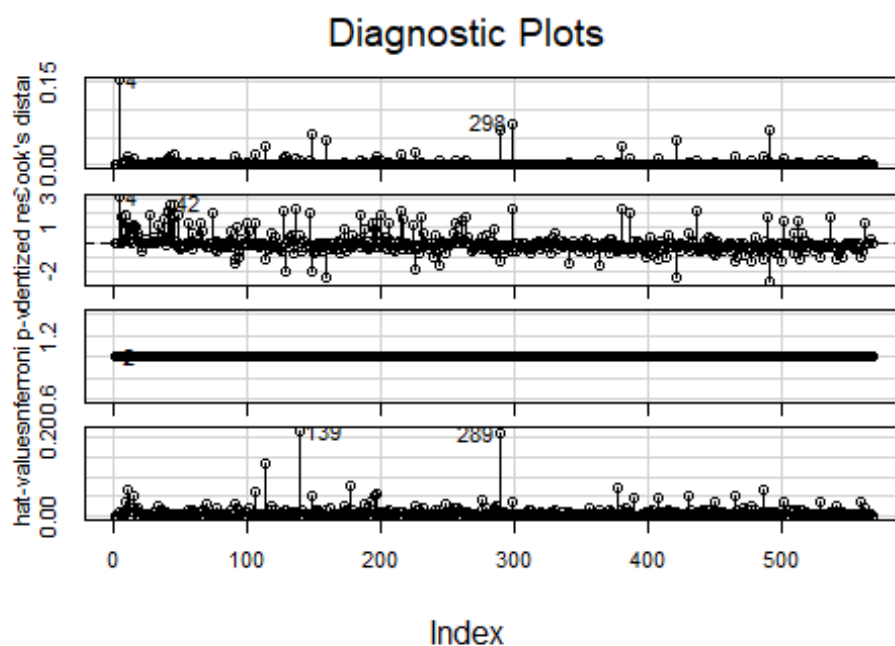
Existem indicativos de out-liers ou observações influentes. Verificando o indivíduo 291, que produziu maior valor para a distância de Cook, observamos que é uma pessoa com algumas características de câncer maligno, como, por exemplo, o raio médio, área média e compacidade, porém seu diagnóstico foi benigno. Fizemos uma comparação entre o Modelo 1 com base completa e Modelo 2 base sem a observação 291.

Quadro 3 – Comparação Modelo1 x Modelo2

##	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
##		

Após realizar uma nova análise sem a observação 291 verificamos que a última variável, média da dimensão fractal, deixa de ser significativa. Abaixo temos um novo gráfico de distância de Cook.

Gráfico 5 – Gráfico de diagnóstico 2



Quadro 3 – Ajuste do Modelo Final

```
## Call: glm(formula = diagnostico ~ mean_area + mean_radius +
##          family = "binomial", data = dados1[-which(rownames(dados1) ==
##            "291"), ])
##
## Coefficients:
##      (Intercept)      mean_area      mean_radius mean_compactness
##          -4.136           0.434          -30.661           39.807
##
## Degrees of Freedom: 567 Total (i.e. Null); 564 Residual
## Null Deviance:      750.5
## Residual Deviance: 261.1    AIC: 269.1
```

Preditor Linear =

$$-4,136 + 0,434 \cdot \text{área média} - 30,661 \cdot \text{raio médio} + 30,807 \cdot \text{compacidade média}$$

4. Considerações finais

Em um primeiro momento, foi observado que a área, raio, compacidade 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, verificamos uma observação que era extrema. Fizemos uma comparação do modelo completo com um modelo 2 sem a observação 291 e como os valores diferiram, refizemos o ajuste reajustando os betas. Observamos, novamente, o gráfico de Cook sem o ponto influente e concluímos que não havia mais nenhum ponto de alavancagem interferindo no ajuste.

Através de um teste de qualidade de ajuste por distribuição Qui Quadrado, observa-se que o valor tabelado com 95% de confiança e 565 graus de Liberdade é de 619,31, enquanto que o valor obtido do modelo ajustado é de 261,1. Portanto não se rejeita a hipótese nula de que o modelo está bem ajustado.