Modelando dados de incidência de câncer de próstata e fatores que influenciam no Antígeno Prostáico Específico

Jailson Rodrigues de souza, 364214

Universidade Federal do Ceará, Fortaleza, Ceará, Brasil

1 Introdução

Um grupo de pesquisadores de um determinado centro médico universitário está interessado em estudar a associação entre antígeno específico da próstata (PSA) e algumas medidas clínicas prognósticas em homens com câncer de próstata em estado avançado. Os dados foram coletados de 97 homens que estavam prestes a sofrer prostatectomias radicais. O conjunto de dados possui um número identificando o paciente e informações a respeito de 8 medidas clínicas.

2 Análise Descritiva

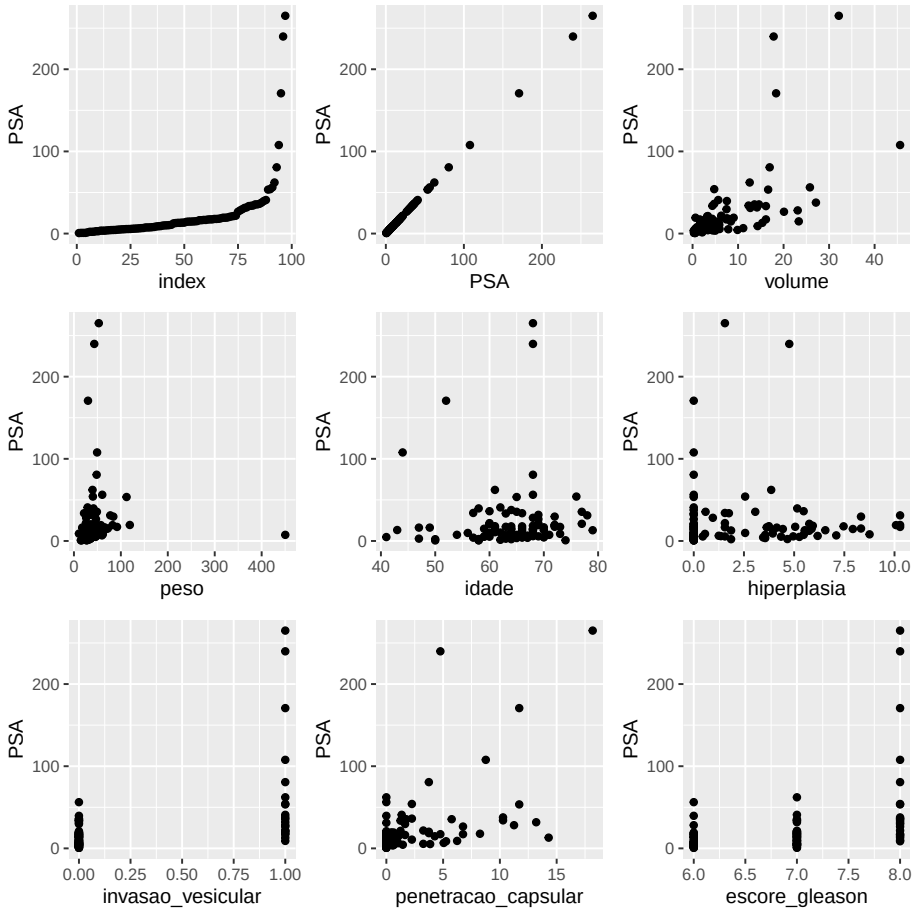
Table 1. Descrição das variáveis utilizadas no estudo

Número da variável	Nome da variável	Descrição
1	Número de identificação	1-97
2	Nível PSA	Nível sérico de antígeno prostático específico (ng/ml)
3	Volume câncer	Estimativa do volume do câncer (cc)
4	Peso	Peso da próstata (gm)
5	Idade	Idade do paciente (anos)
6	Hiperplasia prostática benigna	Quantidade de hiperplasia prostática benigna (cm^2)
7	Invasão da vesícula seminal	Presença ou ausência (1 se sim; 0 se não)
8	Penetração capsular	Grau de penetração capsular (cm)
9	Escore Gleason	Grau patologicamente determinado da doença (escores altos indicam pior prognóstico)

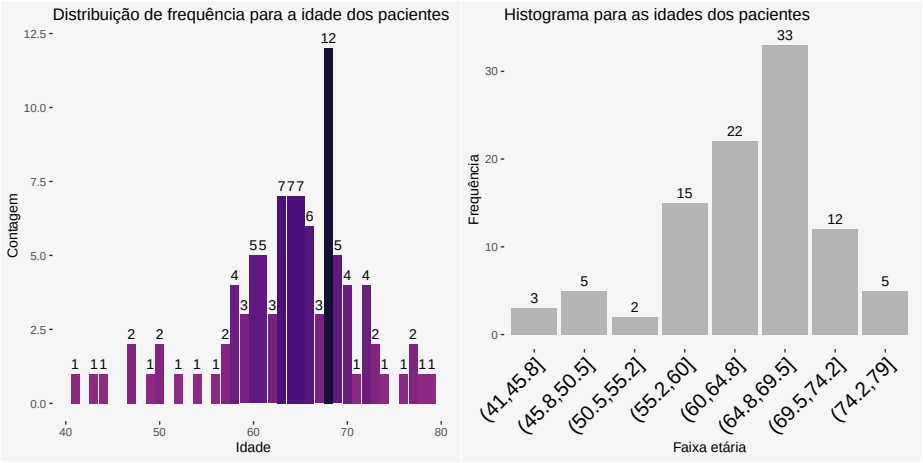
Table 2. Estatísticas descritivas para as variáveis do estudo

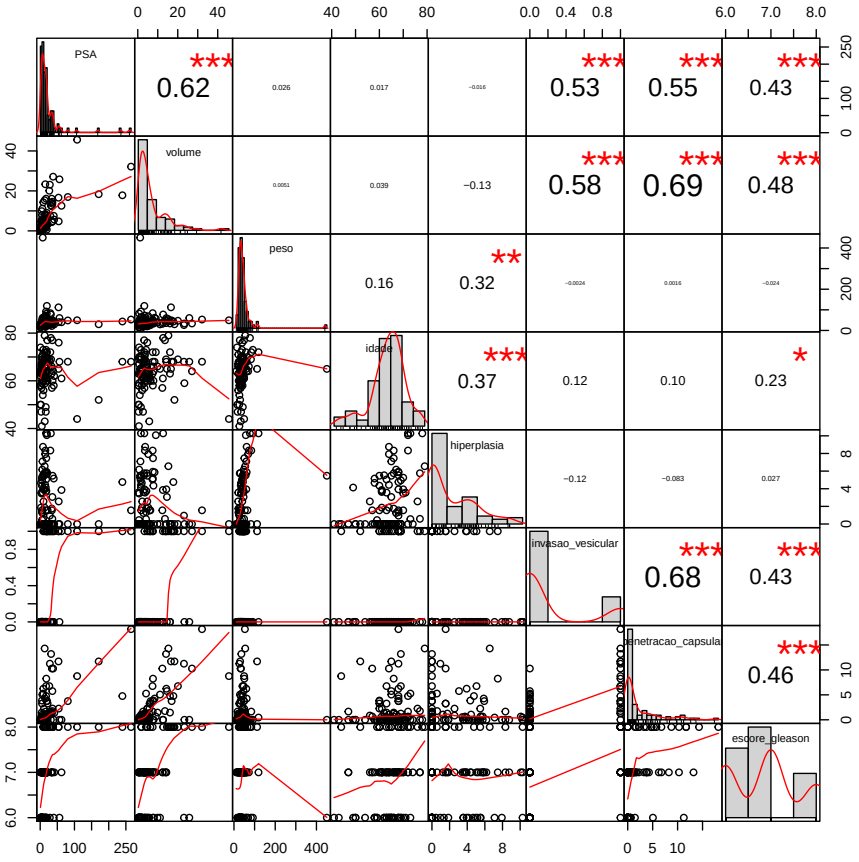
variable	Media	DesvioPadrao	CV	qrt1	qrt2	qrt3	Minimo	Maximo
PSA	23.73	40.78	1.72	5.64	13.33	21.33	0.65	265.07
volume	7.00	7.88	1.13	1.67	4.26	8.41	0.26	45.60
peso	45.49	45.71	1.00	29.37	37.34	48.42	10.70	450.34
idade	63.87	7.45	0.12	60.00	65.00	68.00	41.00	79.00
hiperplasia	2.53	3.03	1.20	0.00	1.35	4.76	0.00	10.28
invasao_vesicular	0.22	0.41	1.91	0.00	0.00	0.00	0.00	1.00
penetracao_capsular	2.25	3.78	1.68	0.00	0.45	3.25	0.00	18.17
escore_gleason	6.88	0.74	0.11	6.00	7.00	7.00	6.00	8.00

Vamos começar plotando as correlações marginais para buscar indicios que nos levem a encontrar fatores iniciais para a analise



Embora as correlações marginais entre o PSA e as outras variáveis não seja elevado, estudos realizados anteriormente constataram que o PSA varia quase linearmente com a idade e o o volume do câncer.

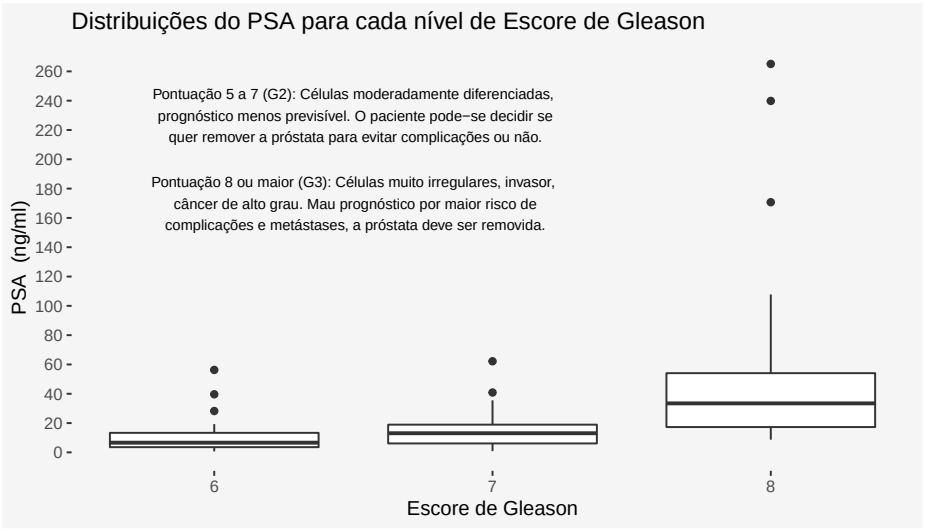

Distribuição de frequência para a idade dos pacientes
Contagem
Idade
Histograma para as idades dos pacientes
Frequência
Faixa etária
(41.45.8]
(45.8,50.5]
(50.5,55.2]
(55.2,60]
(60,64.8]
(64.8,69.5]
(69.5,74.2]
(74.2,79]

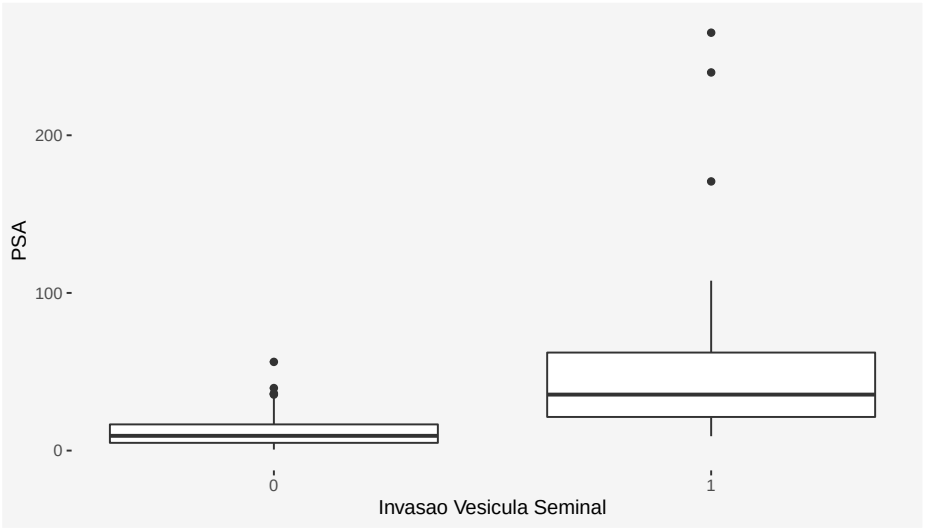

PSA
volume
peso
idade
hiperplasia
invasao_vesicular
penetracao_capsula
escore_gleason
0.62
0.53
0.55
0.43
0.58
0.69
0.48
0.16
0.32
0.37
0.12
0.10
0.23
0.68
0.43
0.46

No gráfico acima: A distribuição empírica de cada variável é mostrada na diagona. Abaixo da diagonal: Os diagramas de dispersão com uma curva ajustada. Acima da diagonal estão os valores das correlaçõpes conjuntamente com seus níveis de significância: Cada nível de significância está associado a uma certa quantidade de estrelas: i.e:

$$\{(0,"***"),(0.001,"***"),(0.01,"**"),(0.05,"*"),(0.1,"."),(1,""))\}$$

O boxplot abaixo mostra O nível de PSA para cada escore de Gleason. Podemos observar que pacientes com Pontuação 8 no escore de Gleason possuem um PSA muito alto. De acordo com o instituto Oncoguia, se o nível do PSA é muito alto, a doença provavelmente está disseminada.

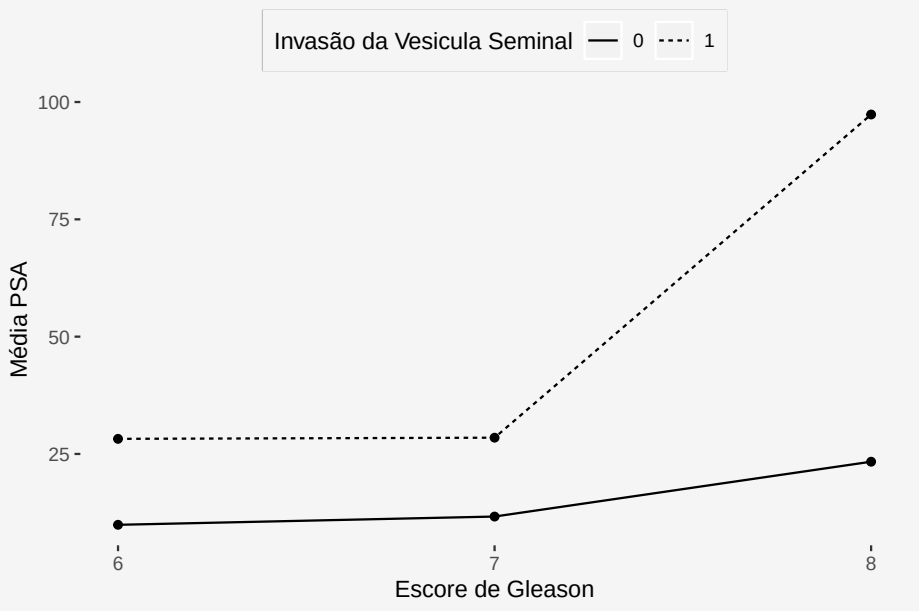




Dados os boxplots acima, podemos perceber que em pacientes que tiveram invasão vesicular, a mediana do nível de PSA é maior.

Table 3. Mediana de PSA entre pacientes com Invasão Vesícular (1) e sem (0).

invasao_vesicular	Mediana_PSA
0	9.356
1	35.517



3 Inferência e Modelagem

Como não há indícios premiliminares de variáveis preditoras eficientes para a modelagem linear para o PSA, iremos ajustar, a princípio, um modelo completo, ou seja, com todas as variáveis:

$$V2 = \beta_0 V1 ... + \beta_9 V9 \tag{1}$$

Os fatores de inflação de variância em conjunto com as correlações entre as variáveis explicativas (VIF) não sao elevados e, portanto, nao temos indícios de multicolinearidade.

Table 4. Correlações entre todas as variáveis explicativas e seus respectivos valores p.

row	column	cor	p
index	PSA	0.603	0.0000000
index	volume	0.621	0.0000000
PSA	volume	0.624	0.0000000
index	peso	0.114	0.2673002
PSA	peso	0.026	0.7988241
volume	peso	0.005	0.9604030
index	idade	0.197	0.0536538
PSA	idade	0.017	0.8672043
volume	idade	0.039	0.7038049
peso	idade	0.164	0.1077533
index	hiperplasia	0.165	0.1062806
PSA	hiperplasia	-0.016	0.8726613
volume	hiperplasia	-0.133	0.1933412
peso	hiperplasia	0.322	0.0013056
idade	hiperplasia	0.366	0.0002239
index	invasao_vesicular	0.567	0.0000000
PSA	invasao_vesicular	0.529	0.0000000
volume	invasao_vesicular	0.582	0.0000000
peso	invasao_vesicular	-0.002	0.9813051
idade	invasao_vesicular	0.118	0.2510649
hiperplasia	invasao_vesicular	-0.120	0.2434580
index	penetracao_capsular	0.477	0.0000008
PSA	penetracao_capsular	0.551	0.0000000
volume	penetracao_capsular	0.693	0.0000000
peso	penetracao_capsular	0.002	0.9877539
idade	penetracao_capsular	0.100	0.3319426
hiperplasia	penetracao_capsular	-0.083	0.4188958
invasao_vesicular	penetracao_capsular	0.680	0.0000000
index	escore_gleason	0.538	0.0000000
PSA	escore_gleason	0.430	0.0000113
volume	escore_gleason	0.481	0.0000006
peso	escore_gleason	-0.024	0.8139337
idade	escore_gleason	0.226	0.0261235
hiperplasia	escore_gleason	0.027	0.7942291
invasao_vesicular	escore_gleason	0.429	0.0000119
penetracao_capsular	escore_gleason	0.462	0.0000020

Table 5. Tolerância e VIF

Variables	Tolerance	VIF
dados$volume	0.46	2.16
dados$peso	0.89	1.13
dados$idade	0.81	1.24
dados$hiperplasia	0.76	1.31
dados$invasao_vesicular	0.50	2.01
dados$penetracao_capsular	0.40	2.52
dados$escore_gleason	0.69	1.46

Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	62202.34	62202.34	64.03	0.00
1	84.66	84.66	0.09	0.77
1	19.82	19.82	0.02	0.89
1	814.66	814.66	0.84	0.36
1	7365.96	7365.96	7.58	0.01
1	932.86	932.86	0.96	0.33
1	1794.05	1794.05	1.85	0.18
89	86457.37	971.43	NA	NA

3.1 Selecção de Variáveis

```
##
## Call:
## lm(formula = PSA ~ volume + factor(invasao_vesicular), data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.145  -7.535  -1.129   4.256 170.018
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.060      4.231   0.251   0.8027
## volume                         2.477      0.495   5.003 2.62e-06 ***
## factor(invasao_vesicular)1    24.647      9.423   2.616   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.09 on 94 degrees of freedom
## Multiple R-squared:  0.431,Adjusted R-squared:  0.4189
## F-statistic:  35.6 on 2 and 94 DF,  p-value: 3.098e-12
```

Table 6. Tabela ANOVA para o modelo completo

Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	62202.34	62202.34	64.03	0.00
1	84.66	84.66	0.09	0.77
1	19.82	19.82	0.02	0.89
1	814.66	814.66	0.84	0.36
1	7365.96	7365.96	7.58	0.01
1	932.86	932.86	0.96	0.33
1	1794.05	1794.05	1.85	0.18
89	86457.37	971.43	NA	NA

Table 7. Tabela ANOVA para o modelo completo

Estimate	Std..Error	t.value	Pr...t..
1.060368	4.23	0.25	0.8026785
2.476724	0.50	5.00	0.0000026
24.647065	9.42	2.62	0.0103769

Table 8. Tabela ANOVA para o modelo reduzido

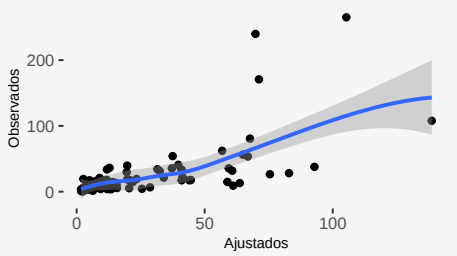
Df	Sum Sq	Mean Sq	F value	Pr(>F)
1	62202.34	62202.34	64.354257	0.0000000
1	6612.59	6612.59	6.841357	0.0103769
94	90856.77	966.56	NA	NA

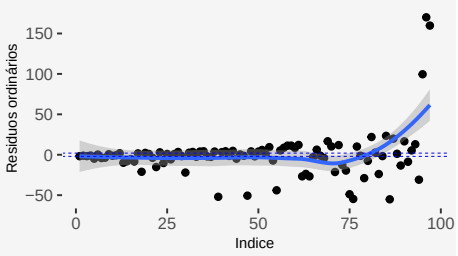
Embora a idade seja um fator relevante para explicar o PSA, nossos dados não capturaram bem essa característica. Sendo assim iremos utilizar um modelo reduzido para explicar a variabilidade do PSA.

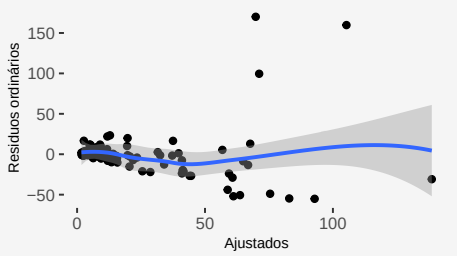
$$E[PSA_i] = \beta_0 volume\_cancer + \beta_1 invasao\_vesicular \tag{2}$$

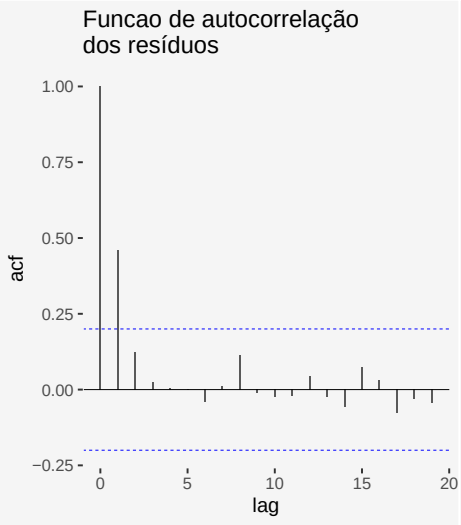
4 Diagnóstico

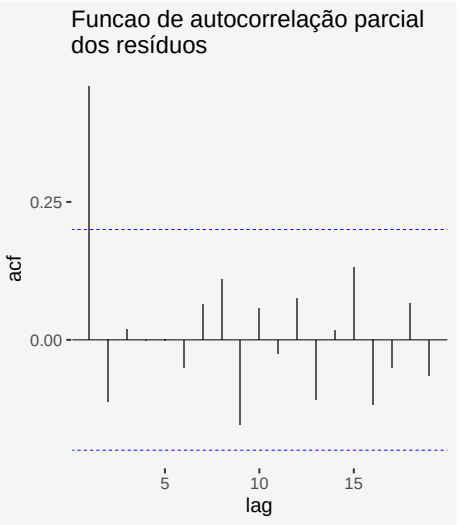
Nesta sessão iremos realizar a anáçise de diagnóstico. A Análise de diagnóstico é importante para verificar se o modelo proposto, quando ajustado aos dados, satisfaz as suposições do modelo linear normal. Por mínima que seja a fulga dessas suposiçõpes, as consequências podem ser catastróficas, pois podem tornar o modelo viesado, e as estimativas não serão mais eficientes no sentido de que suas variâncias não serão as menores na classe dos estimadores BLUE.



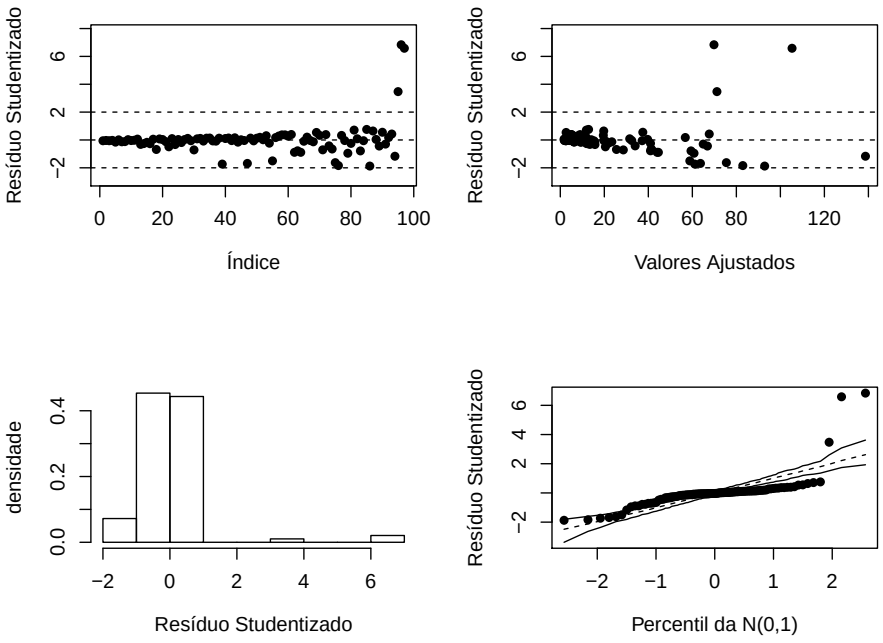


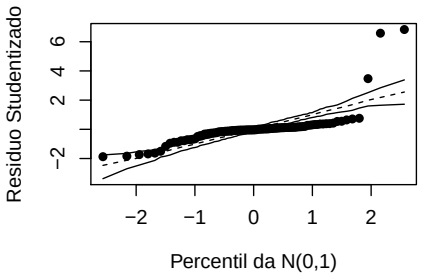






Podemos observar que a FAC para os resíduos indicam um decaimento exponencial, com duas barras ultrapassando o limite. Sugerindo, assim, que exista autocorrelação nos resíduos e que essa autocorrelação pode ser modelada por um AR(2). Já, a FACP sugerem um modelo MA(1). Sendo assim, os residuos podem ser modelados por um processo ARMA(2,1).





5 Conclusões

Durante o estudo foram obtidos diversos indícios de que um modelo de regressão Linear possa não ser adequado para modelar esse tipo de dados, tendo em vista que os graficos de diagnostico corroboram com essa afirmação, podemos rejeitar e propor outros modelos da mais generalizados que possam se adequar bem aos dados sem fugir das suposições.

Não obtante, um possivel modelo inicial que resolvi escolher é dado por:

$$E(PSA|Volume, Invasao) = \begin{cases} 1.060 + 2.477Volume + 24.647, \text{ se houve invasão da vesicula seminal} \\ 1.060 + 2.477Volume, \text{ Caso Contrário} \end{cases}$$

Ou seja, o fato de haver invasão da vesicula semina causa uma variação no intercepto de 23,25 % no nível de PSA do paciente quando o volume do cancer é o mínimo possivel.

References

1. SEARLE, Shayli R. ***Linear Models***, 2nd edition, Hoboken, N.J.: Wiley Interscience, 2003.
2. AMIT GUPTA, CORINNE ARAGAKI, MOMOKAZU GOTOH, NAOYA MASUMORI, SHINICHI OHSHIMA, TAIJI TSUKAMOTO, CLAUS G. ROEHRBORN. ***RELATIONSHIP BETWEEN PROSTATE SPECIFIC ANTIGEN AND INDEXES OF PROSTATE VOLUME IN JAPANESE MEN***, The Journal of Urology, Volume 173, Issue 2, 2005, Pages 503-506, ISSN 0022-5347