Universidade Federal do Paraná
Departamento de Estatística

EFICIÊNCIA DE UM TRATAMENTO APLICADO A PACIENTES COM LEUCEMIA

CE225 - Modelos Lineares Generalizados

Jhenifer Caetano Veloso - GRR20137558
Rogério de Jesus H. Filho - GRR20137589

15 de agosto de 2017

Resumo

Leucemia é um tipo de câncer que atinge as células do sangue, conhecida também como uma doença dos glóbulos brancos. Para verificar a eficiência ou não de um tratamento em pacientes com leucemia, foi realizado um estudo observacional envolvendo seis variáveis explicativas, dado que o tratamento pode assumir duas respostas, a análise estatística se deu através de um modelo linear generalizado, mais especificamente, uma distribuição Binomial com função de ligação logito. Adicionalmente foi realizado uma seleção de variáveis, pelo Critério de Informação de Akaike (AIC), a fim de verificar se o melhor modelo é retirando alguma variável explicativa. Os resultados mostraram que o melhor modelo é aquele que considera três variáveis e uma interação e ainda, verificou-se a qualidade do ajuste final, o qual explica bem as variávies do modelo. Por conseguinte, foram apresentadas as estimativas das variáveis que contribuem para que o tratamento seja satisfatório ou não.

Palavras-chave: leucemia; distribuição Binomial; modelo linar generalizado

Sumário

1 Introdução

Leucemia é um tipo de câncer que atinge as células do sangue, conhecida também como uma doença dos glóbulos brancos, mas em alguns casos a leucemia pode atingir outras células do corpo humano. Existem quatro tipos, leucemia mielóide aguda e crônica, e linfóide aguda e crônica. É uma doença que tem origem na medula óssea passando para o sangue, podendo atingir fígado, baço, gânglios linfáticos, testículos, sistema nervoso central e diversos outros sistemas. Os principais sintomas são anemia, infecções e hemorragias constantes. De acordo com o Instuto Nacional de Câncer José Alencar Gomes da Silva (INCA), para o ano de 2016 no Brasil, estimam-se 5.540 casos novos da doença para homens e 4.530 para mulheres.

Existem diversos métodos de tratamento para a doença, o qual depende do grau de severidade, entre eles estão a quimioterapia, imunoterapia, transplante de medula ou radioterapia. Devido a alta mortalidade para esse diagnóstico, é de vital importância o estudo de aprimoramento de tratamentos ou mesmo novos tipos de tratamento.

Neste contexto, este trabalho foi realizado com o objetivo de verificar a eficácia de um tratamento para um tipo de leucemia aguda.

2 Material e Métodos

2.1 Material

O estudo foi realizado com 51 pacientes adultos, previamente diagnosticados com um tipo agudo de leucemia. Para cada paciente as variáveis observadas foram:

idade: idade do paciente na época do diagnóstico (em anos)
mdd: mancha diferencial da doença (em %)
im: infiltração na medula (em %)
cl: células com leucemia na medula (em %)
md: malignidade da doença ($x10^3$)
tmax: temperatura máxima antes do tratamento (x10°F)
trat: tratamento (1: satisfatório, 0: não satisfatório)
tsobre: tempo de sobrevivência após o diagnóstico (em meses)
sit: situação do paciente (1: sobrevivente, 0: não sobrevivente).

A tabela 1 apresenta as seis primeiras observações do conjunto de dados.

Tabela 1: Conjunto de dados

	idade	mdd	im	cl	md	tmax	trat	tsobre	sit
1	20	78	39	7	0.60	990	1	18	0
2	25	64	61	16	35.00	1030	1	31	1
3	26	61	55	12	7.50	982	1	31	0
4	26	64	64	16	21.00	1000	1	31	0
5	27	95	95	6	7.50	980	1	36	0
6	27	80	64	8	0.60	1010	0	1	0

Pela descrição das variáveis, percebe-se que as variáveis *trat*, *tsobre* e *sit* são as denominadas variáveis dependentes, porém como o objetivo do estudo é avaliar a eficiência do tratamento, as variáveis *tsobre* e *sit* serão desconsideradas da análise.

2.2 Métodos

Dada a natureza binária da variável resposta *trat* (0 ou 1), utilizou-se os conceitos de modelos lineares generalizados. Em regressão, para este tipo de dados a distribuição Binomial é a principal alternativa como componente aleatório do modelo, o componente sistemático é dado pela combinação linear das variáveis explicativas e para a função de ligação foi considerada a função `logit`. A definição do modelo com as características citadas é descrito abaixo:

$$Y_i|x_i \sim Binomial(m_i, \pi_i)$$

$$g(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Em que Y é a variável resposta, $x_{i1}, x_{i2}, \ldots, x_{in}$ as i-ésimas realizações das respectivas variáveis explicativas $X_1, X_2, \ldots, X_n$, $m_i = 1 \; \forall \; i$ (resultando em um conjunto de variáveis de Bernoulli) e $ln\left(\frac{\pi_i}{1-\pi_i}\right)$ é a função de ligação na escala do preditor.

A seleção de variáveis foi realizada através do algoritmo *stepwise* considerando como critério de seleção o AIC (Critério de Informação de Akaike), cujo a fórmula é dada por:

$$AIC_{model} = -2log(L) + 2p$$

Em que L é a verossimilhança e p o número de parâmetros do modelo. Este algoritmo parte de um modelo especificado e realiza sucessivas atualizações na inclusão ou exclusão de variáveis pertencentes ao modelo até que se atinja o menor AIC possível.

Após ajustado o modelo, avaliou-se a qualidade do ajuste através da estatística de Hosmer-Lemeshow, que avalia o modelo ajustado comparando as frequências observadas e as esperadas. O teste associa os dados as suas probabilidades estimadas da mais baixa à mais alta, então faz um teste qui quadrado para determinar se as frequências observadas estão próximas das frequências esperadas.

3 Modelagem Estatística

3.1 Análise Descritiva e Exploratória

Para observar o comportamento e particularidades das variáveis em estudo, nesta seção serão dicutidos algumas medidas e gráficos descritivos. Como primeira visualização será exibido uma tabela com as medidas resumo das variáveis em estudo.

Tabela 2: Medidas resumo

	idade	mdd	im	cl	md	tmax	trat
Min.	20.00	26.00	8.00	1.00	0.00	980.00	0.00
1st Qu.	35.00	53.50	37.50	6.00	1.00	986.00	0.00
Median	50.00	69.00	61.00	9.00	2.60	990.00	0.00
Mean	49.86	65.94	58.20	9.80	9.37	996.14	0.47
3rd Qu.	61.00	81.00	72.50	13.50	9.95	1005.00	1.00
Max.	80.00	97.00	95.00	20.00	115.00	1038.00	1.00

De acordo com a tabela 2 temos indícios de assimetria na disposição dos valores da variável malignidade da doença (indo de 0 até 115). Observações muito dispersas nas variáveis explicativas podem acarretar em pontos influentes no estudo.

O interesse neste caso foi avaliar a variável tratamento em função das demais por meio de um modelo linear generalizado, portanto, uma análise exploratória visando avaliar preliminarmente se há relação entre essas variáveis é imprescindível.

Na figura 1 são exibidos gráficos que ilustram a relação de variáveis combinadas duas a duas separadas pelo tratamento.

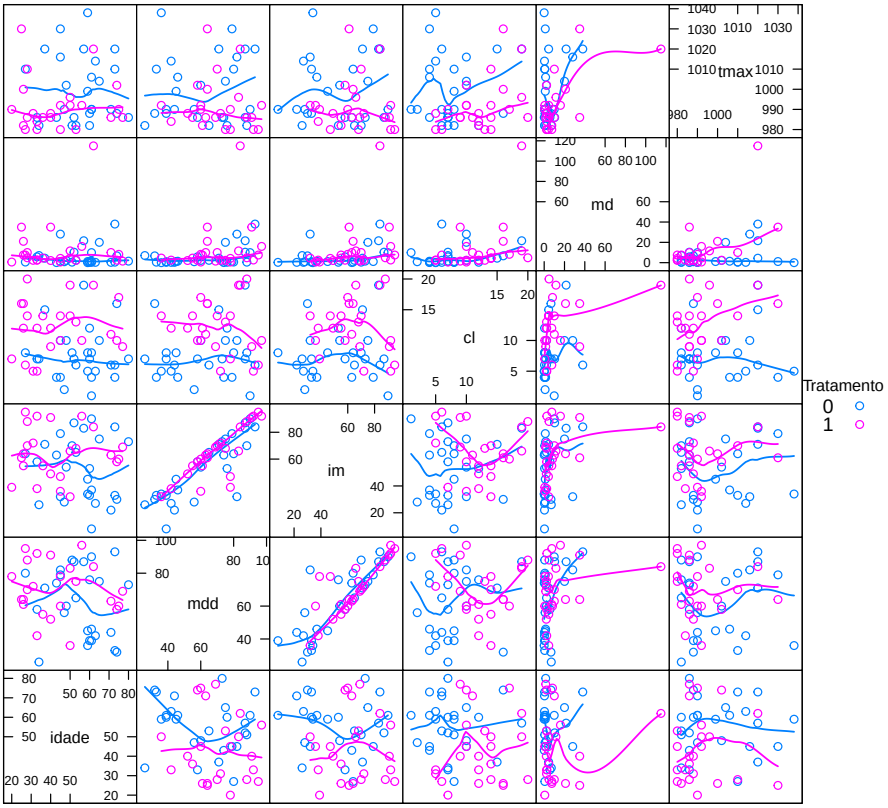

Figura 1: Representação Gráfica das Relações entre as Variáveis

Pela figura 1, percebe-se que as categorias do tratamento satisfatório (cor rosa) e não satisfatório (cor azul) não estão se sobreponto na maioria das variáveis, mas ainda assim não a uma clara relação entre elas. Dentre todos os gráficos, o único que aparentemente apresenta uma tendênca é a porcentagem de mancha diferencial da doença e a porcentagem de infiltração na medula, quando maior a mancha maior será a infiltração.

A fim de quantificar a relação linear entre as variáveis, para que não tenhamos problemas de colinearidade, vamos explorar a matriz de correlação entre elas.

Tabela 3: Matriz de Correlação

	idade	mdd	im	cl	md	tmax	trat
idade	1.00	-0.20	-0.14	-0.12	0.05	0.08	-0.35
mdd	-0.20	1.00	0.85	0.10	0.33	-0.03	0.22
im	-0.14	0.85	1.00	0.14	0.34	-0.01	0.26
cl	-0.12	0.10	0.14	1.00	0.38	0.07	0.49
md	0.05	0.33	0.34	0.38	1.00	0.36	0.19
tmax	0.08	-0.03	-0.01	0.07	0.36	1.00	-0.26
trat	-0.35	0.22	0.26	0.49	0.19	-0.26	1.00

Na tabela 3, todos os elementos da diagonal principal equivalem a 1, pois a correlação de

uma variável com ela mesma é perfeita. O valor da correlação varia de -1 a 1, quanto mais próximo de 1, mais forte é a correlação positiva e quanto mais próximo de -1, mais forte é a correlação negativa. O maior valor foi de 0.85, indicando uma forte correlação positiva entre porcentagem de infiltração na medula e mancha diferencial da doença. Mas sem o ajuste de um modelo linear, não é recomendável realizar inferências a partir desses dados.

3.2 Seleção de Variáveis

Um ponto importante no processo de modelagem é a seleção de variáveis para compor o modelo que melhor explica a variável resposta, isto é, dentre as seis variáveis explicativas disponíveis, devemos encontrar um subconjunto de variáveis importantes para o modelo.

Esta seleção de variáveis foi realizada conforme o algoritmo stepwise, que tem como critério de seleção o AIC descrito na seção 2 e considerou-se a direção *forward* (passo a frente, iniciando com um modelo nulo e inserindo variáveis, uma a uma, até que se encontre o menor AIC tendo como limite um modelo completo especificado). Foi considerado como modelo completo o modelo aditivo com todos os efeitos principais e todas as interações simples, somando ao todo 22 parâmetros.

O algoritmo resultou, com AIC = 49.46, no conjunto de variáveis células com leucemia, temperatura máxima, idade, infiltração na medula, e as interações entre temperatura máxima e infiltração na medula, e idade e infiltração. Com isso, foi considerado somente as variáveis indicadas pelo algoritmo para especificação do modelo.

3.3 Modelo Proposto

O modelo ajustado considerando somente as variáveis selecionadas na subseção anterior fica dado por:

$$Trat_i|x_i \sim Binomial(1, \pi_i)$$

$$g(\pi_i) = ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 cl_i + \widehat{\beta}_2 tmax_i + \widehat{\beta}_3 idade_i + \widehat{\beta}_4 im_i + \widehat{\beta}_5 tmax_i im_i + \widehat{\beta}_6 idade_i im_i$$

A tabela 4 complementa o modelo com os valores estimados dos parâmetros e seus respectivos erros padrão:

Tabela 4: Resumo das Estimativas para o Modelo Ajustado

Efeito	Parâmetro	Estimativa	E. Erro Padrão	Estatística Z	Pr($>\|z\|$)
constante	β_0	-164	129.40	-1.268	0.2050
cl	β_1	0.6156	0.2036	3.024	0.0025
tmax	β_2	0.1671	0.1311	1.274	0.2026
idade	β_3	-0.2109	0.1150	-1.834	0.0666
im	β_4	4.516	2.255	2.003	0.0452
tmax:im	β_5	-0.00462	0.00230	-2.008	0.0446
idade:im	β_6	0.00238	0.00172	1.382	0.1671

Pela tabela 4, nota-se que mesmo após utilizado o algoritmo de seleção de variáveis, tem-se, ao nível de significância de 5%, os efeitos das variáveis tmax, idade e interação entre idade e im não significativos, então ajustou-se um novo modelo sem essas variáveis.

O novo modelo fica dado por:

$$g(\pi_i) = ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 cl_i + \widehat{\beta}_2 tmax_i + \widehat{\beta}_3 im_i + \widehat{\beta}_4 im_i tmax_i$$

E novamente, a tabela 5 apresenta o resumo do novo modelo especificado:

Tabela 5: Resumo das Estimativas para o Modelo Ajustado

Efeito	Parâmetro	Estimativa	E. Erro Padrão	Estatística Z	$Pr(>\|z\|)$
constante	β_0	-109.3	107.5	-1.017	0.30907
cl	β_1	0.4491	0.1427	3.148	0.00164
tmax	β_2	0.1039	0.1073	0.968	0.33295
im	β_3	3.038	1.823	1.667	0.09559
tmax:im	β_4	-0.00303	0.001833	-1.653	0.09827

Como estudo envolve um doença grave, mais especificamente um câncer, não utilizou-se o nível de significância de 5% para dizer quais efeitos são significativos ou não, por esta razão, considerou-se significativos os efeitos apresentados na tabela 5. Foi ainda considerado o efeito da variável *tmax* mesmo que não significativo, porque sua interação com a variável *im* é significativa.

3.4 Análise de Diagnóstico

Antes de validar qualquer análise previamente feita com o modelo, a análise de diagnóstico verificará a adequação do modelo aos dados, verifica se todos os pressupostos são atedidos, bem como observar informações influentes. Abaixo, a verificação dos pressupostos:

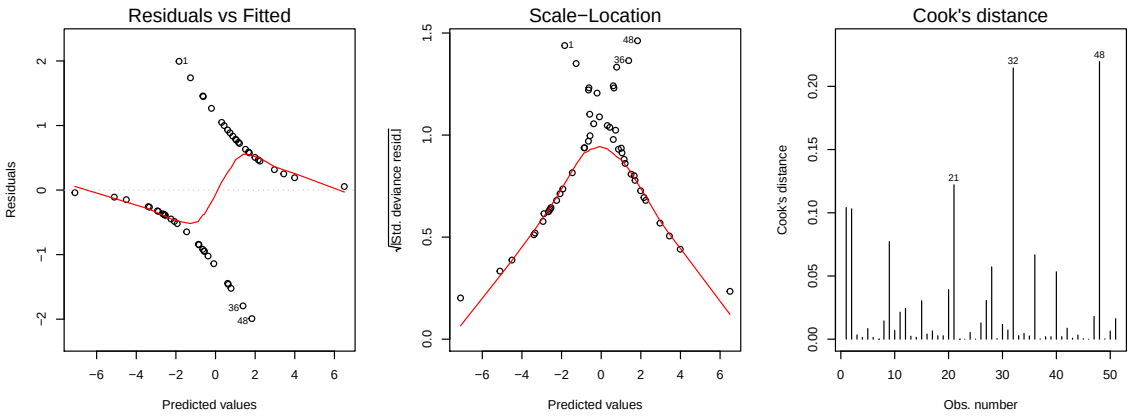

Figura 2: Análise de Diagnóstico do Modelo Proposto - Pressupostos

De acordo com a figura 2 não observa-se fuga aos pressupostos. No primeiro gráfico, há uma forma não linear devido ao fato da variável resposta ser dicotômica, apresentando apenas valores 0 ou 1. No segundo gráfico, não se tem evidência de tendência e novamente percebe-se uma disposição aleatória dos pontos, não caracterizando uma relação média variância. E no terceiro, tem-se a distância de Cook, observa-se que as observações 32 e 48 se destavam, porém elas não evidenciam um ponto de alavancagem. Como complemento, a figura 3 apresenta o gráfico quantil-quantil com envelope simulado.

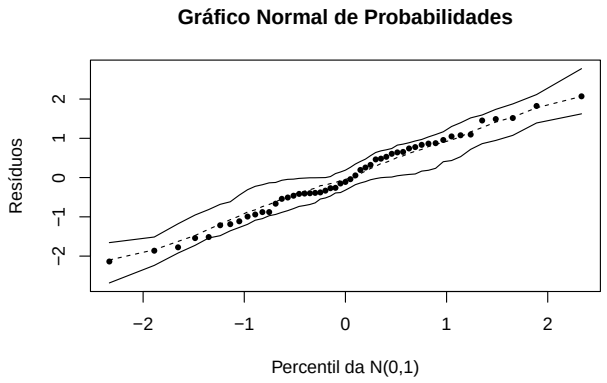

Figura 3: qqplot com Envelope Simulado

Na figura 3, é apresentado o gráfico dos resíduos e envelopes simulados para a banda de 95% de confiança, com ele é possível verificar possíveis problemas nas especificações do modelo, o que não ocorreu nesse caso, todos os resíduos estão dentro dos intervalos simulados.

E por fim, para complementar a análise de diagnóstico, verificamos se há medidas de influência através dos 3 gráficos abaixo:

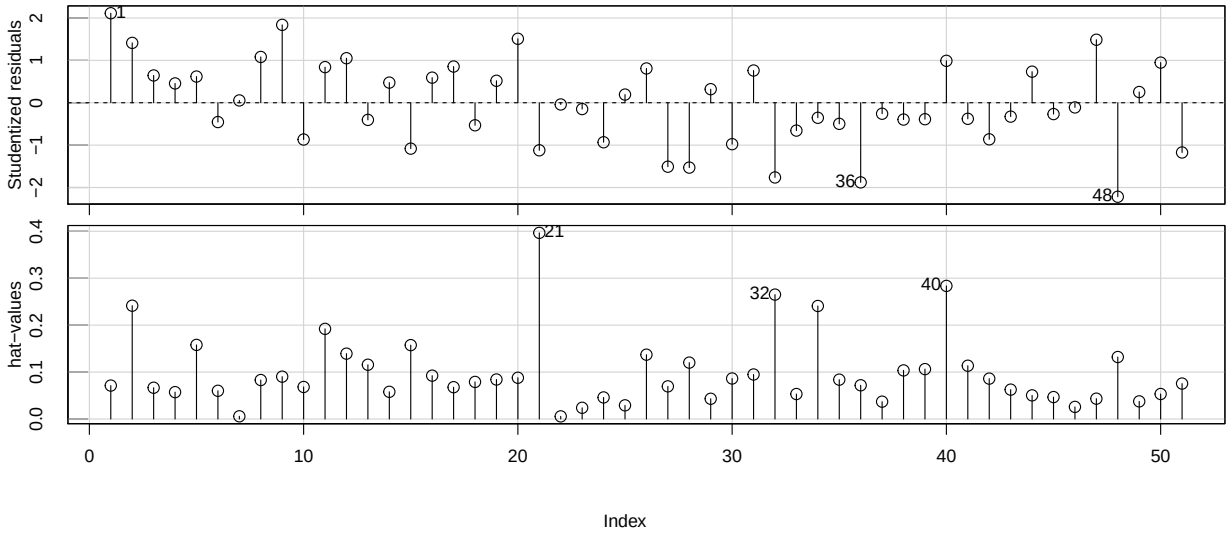

Figura 4: Medidas de Influência

Nos gráficos da figura 4 são apresentadas duas medidas de influência: resíduos studentizados e valores de alavancagem h. Ambos indicam que valores com grandes magnitudes, em relação aos demais, podem se apresentar como observações influentes. Foram ajustados modelos sem as observações identificadas, porém as estimativas e componentes do modelo não apresentaram diferenças significativas, então as observações 32, 40 e 48 continuaram presentes na análise.

3.5 Resultados

Com o modelo especificado e verificado o prossupostos, é possível avaliar a qualidade deste ajuste. Para isso foi utilizado a estatística de Hosmer e Lemeshow mencionada na seção 2, a qual é dada por:

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k n_k^{'} \bar{\pi}_k\right)^2}{n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Considerando a hitpótese nula de que "o modelo explica bem os dados", o teste resultou em $\hat{C} = 9.458237$ com um $p - valor = 0.305125$, ou seja, o modelo está bem ajustado.

4 Conclusão

Conclui-se que os percentuais de células infectadas com leucemia e de infiltração na medula óssea, a temperatura máxima do paciente antes do tratamento, bem como a interação entre a infiltração e a temperatura contribuem para que o tratamento seja satisfatório. Como a estimativa da interação apresenta valor negativo, ela incorre uma contribuição menor do que as outras.