Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Análise Bayesiana de um ensaio clínico de betacaroteno para prevenir câncer de pele de células basais e espinocelulares.

Douglas de Paula Nestlehner

Setembro, 2022

Sumário

Capítulo 1

Problema Apresentado

Em 1990 foi publicado um artigo que consiste na análise de um ensaio clínico que objetiva a prevenção de câncer de pele. Os dados são do Skin Cancer Prevention Study, um ensaio clínico randomizado, duplo-cego e controlado por placebo de betacaroteno para prevenir câncer de pele não melanoma em indivíduos de alto risco.

Um total de 1.805 indivíduos foram randomizados para placebo ou 50mg de beta-caroteno por dia durante 5 anos. Os indivíduos foram examinados uma vez por ano e biopsiados se houver suspeita de câncer para determinar o número de novos cânceres de pele ocorridos desde o último exame. A variável resposta é uma contagem do número de novos cânceres de pele por ano.

1.1 Variáveis

Dados completos estão disponíveis em 1.683 indivíduos, compreendendo um total de 7.081 medições. As variáveis presentes na base são:

- **Tratamento**: A variável categórica Tratamento é codificada 1=beta-caroteno, 0=placebo.
- **Ano:** A variável Ano denota o ano de acompanhamento.
- **Gênero:** A variável categórica Gênero é codificada 1=masculino, 0=feminino.
- **Pele:** A variável categórica Pele indica o tipo de pele e é codificada 1=queimaduras, 0=caso contrário.
- **Exposição:** A variável Exposição é uma contagem do número de cânceres de pele anteriores.
- **Idade:** A variável Idade é a idade (em anos) de cada sujeito na randomização.

Capítulo 2

Resultados

Nessa seção, iremos apresentar os resultados obtidos nesse trabalho.

2.1 Análise descritiva

Apenas no intuito de conhecer o formato dos dados, temos representado na tabela 2.1 as primeiras observações da base.

Tabela de dados no formato Longo

Obs	ID	Centro	Idade	Pele	Gênero	Exposição	Y	Tratamento	Ano
1	100034	1	51	1	1	4	0	0	1
2	100034	1	51	1	1	4	1	0	2
3	100034	1	51	1	1	4	1	0	3
4	100034	1	51	1	1	4	1	0	4
5	100034	1	51	1	1	4	0	0	5
6	100045	1	68	1	0	2	0	0	1

Figura 2.1: Base de dados

Afim de saber o comportamento das principais variáveis, construímos gráficos de frequência em relação ao ano, permitindo assim, observar os resultados ao longo do tempo.

Na figura 2.3 temos representado os graficos de frequencia por ano, das variaveis gênero, tratamento, pele e centro.

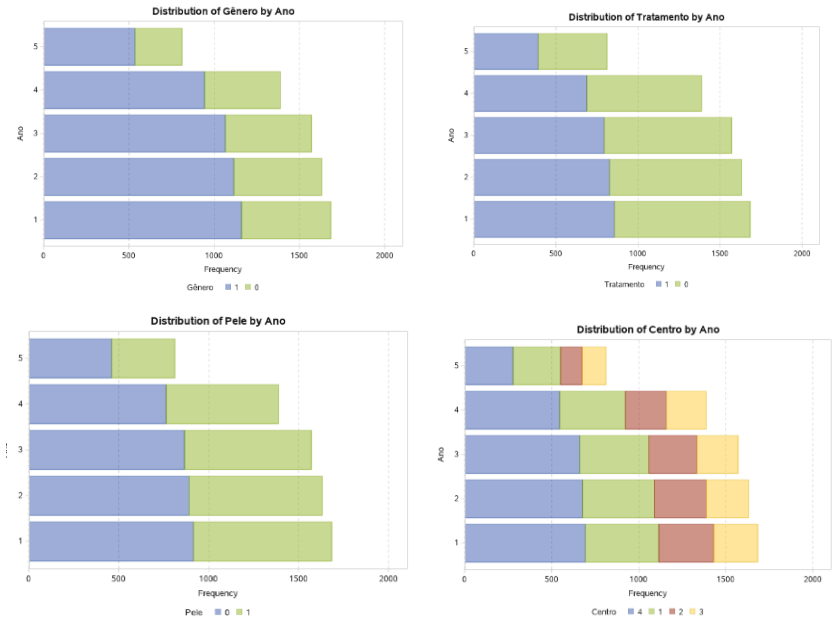

Figura 2.2: Gráficos de frequência das variáveis categóricas

Podemos notar, de modo geral, que a medida que o tempo passa o numero de observações vai diminuindo (o que é normal). Na variável gênero existem sempre um maior numero de observações do sexo masculino; em tratamento um maior frequência do tratamento tipo 1; na variável pele existe um maior equilíbrio para as duas classes; e na variável centro não conseguimos interpretar exatamente o que significa, pois a mesma não é especificada na base de dados.

Para observar as variáveis quantitativas (Y, exposição e idade) construimos um histograma para cada uma, representados na Figura **??**

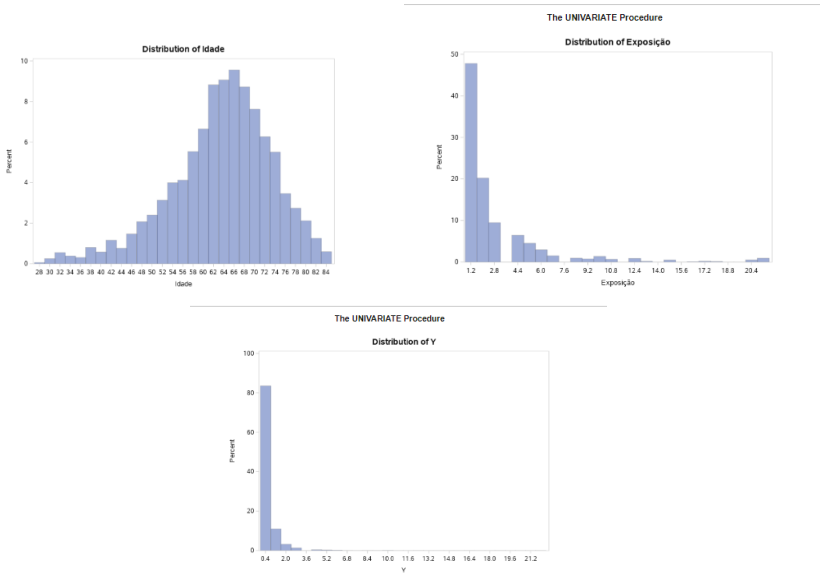

Figura 2.3: Gráficos de histograma das variáveis quantitativas

2.2 Modelo Linear Generalizado Misto Bayesiano

Podemos ter como interesse modelar o numero de novos cânceres de pele, e para isso, podemos ter que essa contagem pode ser dada por alguma taxa/relação do numero de cânceres anteriores (variável "Exposição") e o novo numero observado (variável "Y").

Para esse trabalho, iremos supor a seguinte taxa:

$$\text{Taxa} = \frac{\text{Novos caso (Y)}}{\text{Exposição}}$$

Desse modo, se formos modelar a taxa, teremos algo do tipo:

$$log(\text{Taxa}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

$$log\left(\frac{\text{Novos caso (Y)}}{\text{Exposição}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

$$log(\text{Novos caso (Y)}) - log(\text{Exposição}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

$$log(\text{Novos caso (Y)}) = log(\text{Exposição}) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Assim sendo, para o ajuste do modelo linear generalizado misto Bayesiano, devemos definir a distribuição da variável reposta, qual função de ligação iremos usar no ajuste (já consideramos como log para demonstrar o exemplo de acima), o preditor linear adequado, e as distribuições a priori para os parâmetros.

- **Distribuição de Y:** Por se tratar da contagem de cânceres de pele, atribuímos a distribuição **Poisson**;
- **Função ligação:** Como iremos usar a distribuição poisson, a função de ligação mais indicada é a logarítmica (log(Y));
- **Distribuições a priori parâmetros:**
  - **Vetor de coeficientes lineares:** Utilizamos a princípio, o padrão do PROC BGLIMM.
  - **Matriz de variâncias e covariâncias dos efeitos aleatórios:** Utilizamos a princípio, o padrão do PROC BGLIMM.
- **Parâmetros MCMC das cadeias de Markov**: definimos o burnin = 500 (é o numero simulações padrão, por se tratar de uma base relativamente pequena, acredita-se que é o suficiente), para o tamanho da cadeia definimos como sendo =

10000; e afim de controlar o afinamento da cadeia de Markov, definimos o thin = 1).

Realizando o ajuste considerando as definições anteriores, e utilizando o PROC BGLIMM do SAS, temos os seguintes resultados.

Na figura 2.4 temos as informações gerais sobre o ajuste.

The BGLIMM Procedure

Model Information	
Data Set	WORK.DADOS2
Response Variable	Y
Distribution	Poisson
Link Function	Log
Fixed Effects Included	Yes
Random Effects Included	Yes
Sampling Algorithm	Gamerman, Conjugate
Burn-In Size	500
Simulation Size	10000
Thinning	1
Random Number Seed	78145475
Number of Threads	1

Class Level Information		
Class	Levels	Values
Tratamento	2	0 1
Pele	2	0 1
Gênero	2	0 1
Ano	5	1 2 3 4 5
Centro	4	1 2 3 4

Number of Observations	
Number of Observations Read	7081
Number of Observations Used	7081

Figura 2.4: Informações gerais.

Onde é possível notar as definições sobre o modelo, distribuição da variável reposta, função de ligação, os parâmetros MCMC das cadeias de Markov, etc.

Na figura 2.5, temos as distribuições a priori dos parâmetros do modelo, que a principio, usamos o padrão do procedimento BGLIMM.

Priors for Fixed Effects	
Parameter	Prior
Intercept	Constant
Idade	Constant
Ano 1	Constant
Ano 2	Constant
Ano 3	Constant
Ano 4	Constant
Tratamento 0	Constant
Pele 0	Constant
Centro 1	Constant
Centro 2	Constant
Centro 3	Constant

Priors for Scale and Covariance Parameters	
Parameter	Prior
Random Var	Inverse Gamma (Shape=2, Scale=2)

Figura 2.5: Distribuições a priori

Na figura 2.6 temos as estimativas obtidas no ajuste.

Posterior Summaries and Intervals					
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
Intercept	10000	-3.6117	0.3052	-4.2056	-3.0147
Idade	10000	0.0152	0.00452	0.00646	0.0239
Ano 1	10000	-0.0852	0.0828	-0.2516	0.0755
Ano 2	10000	-0.2018	0.0856	-0.3691	-0.0372
Ano 3	10000	-0.1226	0.0840	-0.2916	0.0416
Ano 4	10000	-0.0768	0.0857	-0.2460	0.0879
Ano 5	0	.	.	.	.
Tratamento 0	10000	-0.0897	0.0823	-0.2552	0.0630
Tratamento 1	0	.	.	.	.
Pele 0	10000	-0.1841	0.0843	-0.3449	-0.0155
Pele 1	0	.	.	.	.
Centro 1	10000	-0.3464	0.1124	-0.5678	-0.1235
Centro 2	10000	0.3108	0.1066	0.1121	0.5330
Centro 3	10000	0.2054	0.1180	-0.0315	0.4320
Centro 4	0	.	.	.	.
Random Var	10000	1.0497	0.0925	0.8719	1.2322

Figura 2.6

Afim de verificar se as cadeias convergiram, fizemos os gráficos de disgnósticos representados na Figura 2.7. Vale ressaltar que o diagnostico é feito para cada efeito no modelo, mas apresentamos no relatório apenas o do efeito aleatório, tendo em vista que todos os outros apresentaram comportamento parecido.

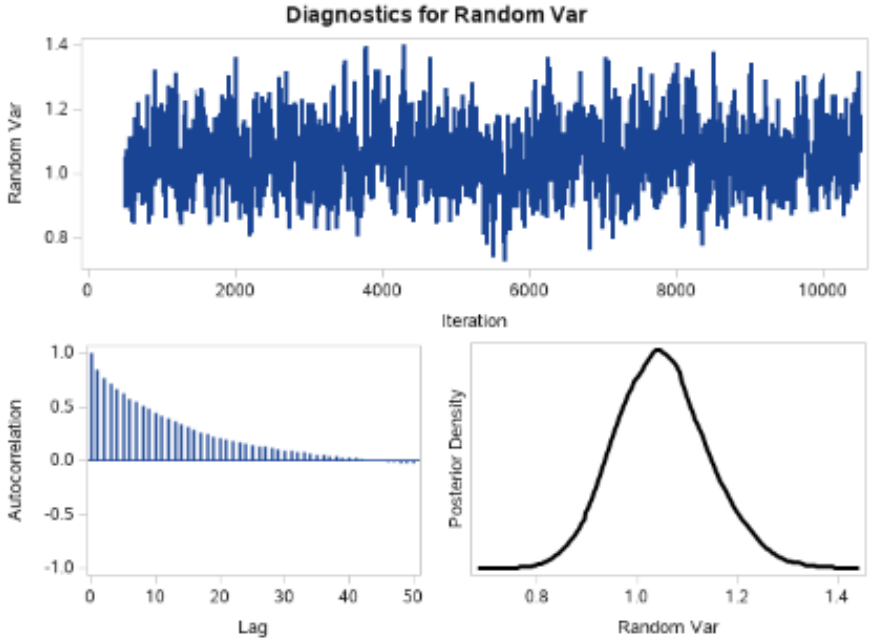

Figura 2.7: Diagnostico variável aleatória.

O primeiro gráfico (acima) tem como objetivo mostrar a estabilidade da cadeia de Markov ao longo da simulação, aparentemente o resultado apresentado é adequado. O segundo gráfico (autocorrelação) tem como objetivo mostrar a diminuição nas autocorrelações, o que também está adequado. De modo geral, a análise de diagnostico apresentou bons resultados.

Capítulo 3

Conclusão

Conseguimos ajustar um modelo linear generalizado misto bayesiano para o problema apresentado, entretanto alguns pontos de destaques foram observados no processo:

- A modelagem da contagem de novos casos sem considerarmos alguma taxa (que não tínhamos definida no problema), se tornou algo bem complicado utilizando o BGLIMM;
- O custo computacional no ajuste dos modelos foi bem alto, sempre quando precisávamos ajustar o modelo (realizando alguma alteração / testes), o tempo para se obter os resultados foi algo entre 5-8min (para cada ajuste).
- A definição das distribuições a priori foi outro ponto de dificuldade, sempre quando testávamos alguma outra sem ser o padrão, o tempo para os ajustes demorava ainda mais.