Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

ESTUDO SOBRE A TAXA ANUAL DE CÂNCER NASAL EM UMA REFINARIA DE NÍQUEL NO PAÍS DE GALES

CE225 - Modelos Lineares Generalizados

Eduardo Elias Ribeiro Junior

Curitiba, 19 de novembro de 2014

Contents

1 Introdução

Para estudar a associação entre a taxa anual de câncer nasal em trabalhadores de uma refinaria de níquel no País de Gales realizou-se um estudo onde foram coletadas as seguintes informações: idade no primeiro emprego com 4 níveis (1: <20, 2: 20-27, 3: 27.5-34.9 e 4: 35+ anos), ano do primeiro emprego com 4 níveis (1: <1910, 2: 1910-1914, 3: 1915-1919 e 4: 1920-1924), tempo decorrido desde o primeiro emprego com 5 níveis (1: 0-19, 2: 20-29, 3: 30-39, 4:40-49 e 5: > 50 anos), número de casos de câncer e o total de pessoas-anos de observação. Um total de 72 observações foram registradas, na tabela 1 são apresentados os resultados dos números de casos de câncer resultante das combinações das variáveis explicativas idade do trabalhador no primeiro emprego, ano do primeiro emprego e tempo decorrido desde o primeiro emprego. Perceba que nesta tabela não foram exibidos os valores do total de pessoas-ano de observação, cujo temos 72 valores. Esta variável não é de interesse no estudo servindo apenas de equiparação de resultados.

Table 1: Número de casos de câncer nasal

Idade	Ano	Tempo				
		0 - 19	20 - 29	30 - 39	40 - 49	> 50
< 20	< 1910	NA	0	0	0	0
	1910 - 1914	0	1	0	1	0
	1915 - 1919	0	0	0	0	0
	1920 - 1924	0	0	0	0	0
20 - 27	< 1910	NA	1	2	0	2
	1910 - 1914	0	1	4	2	2
	1915 - 1919	0	0	0	2	0
	1920 - 1924	0	1	0	3	0
27.5 - 34.9	< 1910	NA	3	1	1	1
	1910 - 1914	0	3	2	3	0
	1915 - 1919	0	2	1	0	1
	1920 - 1924	0	0	1	1	0
> 35	< 1910	NA	0	0	0	NA
	1910 - 1914	0	2	5	0	NA
	1915 - 1919	0	2	1	0	NA
	1920 - 1924	1	3	0	0	NA

2 Materiais e Métodos

Os dados apresentados na tabela 1 serão estudados e sumarizados com a teoria de Modelos Lineares Generalizados (MLG's). A variável resposta considerada neste estudo é o número de casos de câncer e nesta abordagem de MLG's é considerado como componente aleatório do modelo, cujo devemos associar uma distribuição de probabilidades que se encaixe na família exponencial bi-paramétrica de distribuições[1]. A distribuição de Poisson é a mais adequada para as características descritas. Outro componente importante para a definição de um MLG é a parte sistemática ou estrutural onde especificamos a relação linear dos efeitos das variáveis explanatórias e a função de ligação, que tem por objetivo linearizar a relação entre a média da distribuição associada e o preditor linear e delimitar os valores produzidos pelo modelo ao espaço paramétrico válido para a média. Abaixo definimos o modelo em notação matemática:

$$Y_{ijk} \sim Poisson(\mu_{ijk} = \lambda_{ijk}.t_{ijk}), \tag{1}$$

$$log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \delta_k + log(t_{ijk}) \tag{2}$$

Os parâmetros α_i, β_j e δ_k representam os efeitos das variáveis idade no primeiro emprego, ano do primeiro emprego e tempo decorrido desde o primeiro emprego nas categorias i = 1,2,3,4, j = 1,2,3,4 e k = 1,2,3,4,5. Na especificação do preditor linear foi incorporada a variável $log(t_{ijk})$ que representa apenas a padronização da taxa conforme o total de pessoas-anos de observação sem prâmetro relacionado, caracterizando um termo **offset** que após ajuste será devidamente testado. Nesta parametrização os parâmetros α_1, β_1 e δ_1 são iguais a zero, pois o efeito da primeira categoria de cada variável será incorporada no parâmetro μ. Ao todo o modelo teórico compreende 11 parâmetros a serem estimados.

[1] Ver **Modelos Lineares Generalizados** por M. Turkman e G. Silva, página 5, disponível em http://docentes.deio.fc.ul.pt/maturkman/mlg.pdf

2.1 Análise Descritiva e Explorátória

Como visualização dos dados apresentaremos inicialmente três tabelas com medidas descritivas da variável reposta número de casos de câncer estratificadas pelas categorias das variáveis explanatórias.

Table 2: Média e Desvio-Padrão do Número de Casos de Dengue por Idade

<20	20-27	27.5-34.9	>35
0.1053 (0.3153)	1.0526 (1.2236)	1.0526 (1.0788)	0.9333 (1.4864)

Table 3: Média e Desvio-Padrão do Número de Casos de Dengue por Ano

<1910	1910-1914	1915-1919	1920-1924
0.7333 (0.9612)	1.3684 (1.5352)	0.4737 (0.7723)	0.5263 (0.9643)

Table 4: Média e Desvio-Padrão do Número de Casos de Dengue por Tempo

0-19	20-29	30-39	40-49	>50
0.0833 (0.2887)	1.1875 (1.1673)	1.0625 (1.5262)	0.8125 (1.1087)	0.5 (0.7977)

Nas tabelas 2, 3 e 4 podemos notas algumas categorias com uma médias de número de casos de câncer mais elevadas, porém os respectivos erros padrões das médias obervadas (valores entre parênteses nas tabelas) foram bem elevados, o que pode ser resultante do alto número de zeros na amostra. Na figura 1 temos alguns gráficos de caixas que podem auxiliar a visualizar a disposição do número de casos nas categorias das variáveis explanatórias.

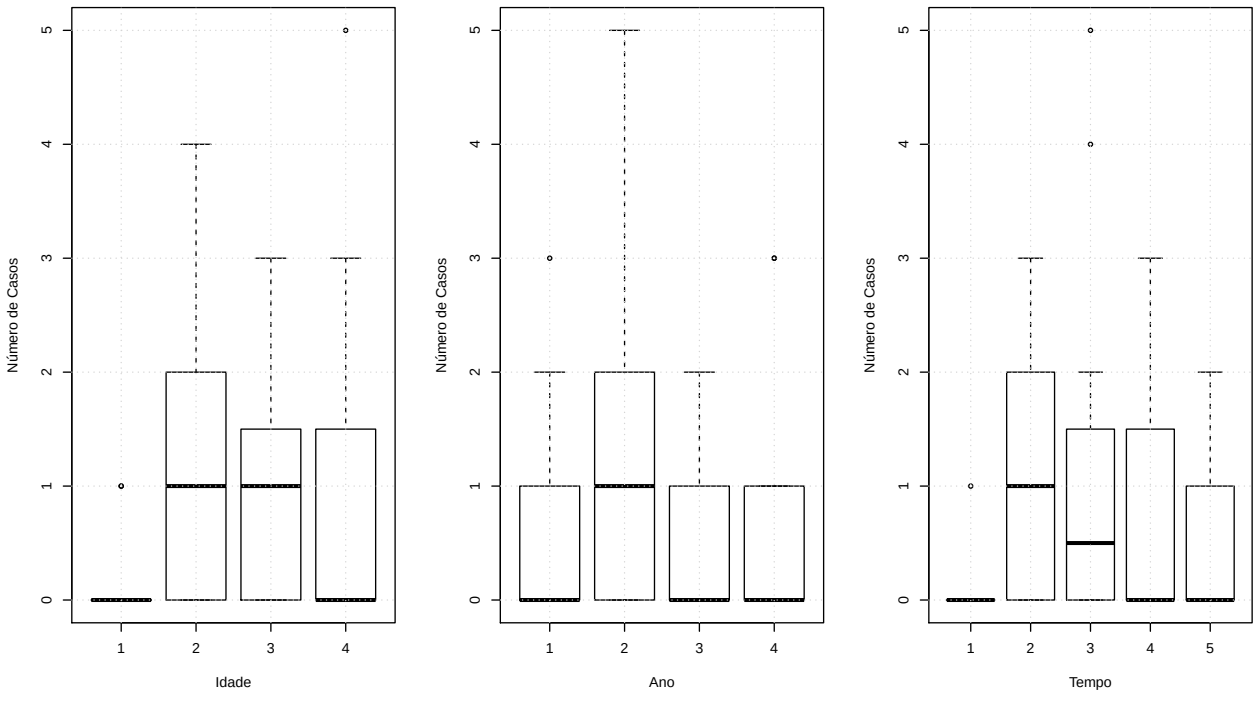

Figure 1: Box-Plots da Variável Resposta

Nos gráficos da figura 1 a dificuldade causada pelo excesso de zeros no conjunto de dados é mais evidente, note a mediana (linha horizontal em negrito nos gráficos de caixa)8 vezes fixada em zero. Mas ainda com o excesso de zeros presentes podemos notar um destaque das categorias 2 (20-27 anos) e 3 (30-39) de idade no primeiro emprego com relação as demais, para ano do primeiro emprego na categoria 2 (1910-1914) o número de casos de câncer se dispõe de forma bem distinta das demais categorias desta variável e para o tempo decorrido desde o primeiro emprego a categoria 2 (20-29) também parece ter se destacado. Perceba que ainda não se pode

realizar conclusões, pois além desta ser uma análise preliminar, a variável total de pessoas-ano de obervação ainda não foi incluída para equiparação dos resultados, no decorrer do estudo faremos es equiparação.

2.2 Modelo Ajustado

Com as 72 obsevaçoes exibidas na tabela 1 ajustamos o modelo, descrito por (1) e (2). Perceba que na tabela 1 exitem combinações entre as variáveis explanatória não avaliadas (NA), as combinações de idade no primeiro emprego igual na categoria e tempo decorrido desde o primeiro emprego igual a categoria 5 não foram observadas como também a combinação ano do primeiro emprego <1910 e tempo decorrido de 0-19. Isso impossibilita o ajuste de um modelo saturado ou de um modelo com as interações entre tempo decorrido e ano do primeiro emprego ou idade. Na tabela 5 é apresentado o quadro de diferença de deviances dos modelos sequenciais.

Table 5: Análise de Diferenças de *Deviances* nos Modelos Sequenciais

Modelos	gl_s	*Deviances*	Diferença de gl_s	Diferença de *Deviances*	Valor p
Nulo	71	135.68	-	-	-
Idade	68	109.07	3	26.61	0.0000
Ano | Idade	65	70.78	3	38.29	0.0000
Tempo | Ano, Idade	61	58.17	4	12.61	0.0133
Idade*Ano|Tempo, Ano, Idade	52	49.08	9	9.08	0.4298

Perceba pela tabela 5 que o modelo com apenas os efeitos principais (Tempo | Ano, Idade) produziu um valor de significância > 0.02 justificando a inclusão das três variáveis explicativas no modelo. Já o modelo com a interação Idade*Ano apresenta um nível de significância de 0.42 evidênciando que a inclusão deste interação não proporciona uma diferença significativa das deviances. Portanto o modelo a ser ajustado continua sendo descrito pelas equações (1) e (2). Na tabela 6 são apresentadas as estimativas dos parâmetros deste modelo.

Table 6: Resumo das Estimativas para o Modelo Ajustado

Efeito	Parâmetro	Estimativa	E. Erro Padrão	Estatística Z	Pr($>\|z\|$)
Constante	μ	-9.2718	1.3189	-7.03	0.0000
Idade	α_2	1.6728	0.7521	2.22	0.0261
Idade	α_3	2.4817	0.7591	3.27	0.0011
Idade	α_4	3.4282	0.7816	4.39	0.0000
Ano	β_2	0.6189	0.3713	1.67	0.0955
Ano	β_3	0.0541	0.4681	0.12	0.9079
Ano	β_4	-1.1261	0.4529	-2.49	0.0129
Tempo	δ_2	1.5982	1.0475	1.53	0.1271
Tempo	δ_3	1.7512	1.0552	1.66	0.0970
Tempo	δ_4	2.3549	1.0701	2.20	0.0278
Tempo	δ_5	2.8176	1.1183	2.52	0.0117

Notamos pelas estimativas do modelo que há uma estimativa de incidência de casos de câncer maior nas categoriais que representam uma maior idade do indivíduo no primeiro emprego o mesmo ocorre com o tempo decorrido desde o primeiro emprego. Por outro lado, para o ano do primeiro emprego temos que a incidência de casos é menor para datas maiores.

2.3 Análise de Diagnóstico

Nesta seção iremos avaliar o modelo para posteriormente realizarmos inferências a partir dele. Primeiramente avaliando a função desvio temos um valor avaliado de 58.166 (61 graus de liberdade) com um nível descrito de 0.579, indicando um bom ajuste. Para os gráficos da figura 2, não percebemos evidências de má especificação do modelo. Algumas observações foram destacadas em alguns gráficos, mas sua exclusão da análise não apresentou diferenças nas interpretações, portanto decidiu-se permanecer com as 72 observações na análise.

```
## Error in na.omit(modelo$data):  object 'm3o' not found
```

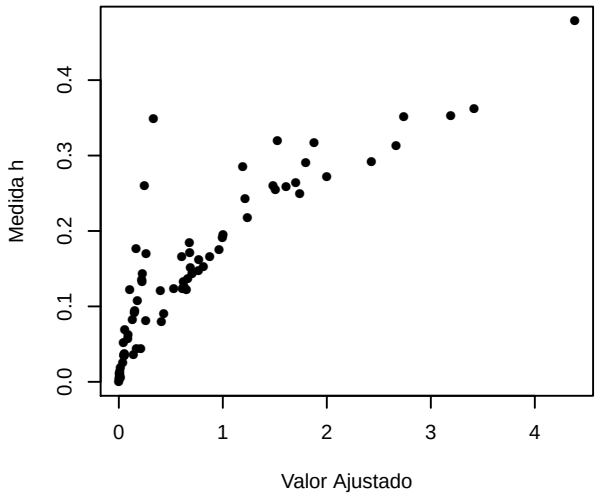

Figure 2: Análise de diagnóstico

3 Resultados

Com o modelo ajustado e avaliado. Podemos estudar a sumarização realizada por ele. Vamos inicialmente ver os resultados estimados pelo modelo para as combinações das variáveis explicativas na tabela 7. Assim podemos observar o ajuste a partir da comparação entre tabela 1 e tabela 7.

Table 7: Número Estimados de Casos de Câncer Nasal

Idade	Ano	Tempo				
		0 - 19	20 - 29	30 - 39	40 - 49	> 50
< 20	< 1910	NA	0.009	0.038	0.052	0.052
	1910 - 1914	0	0.151	0.181	0.223	0.228
	1915 - 1919	0.007	0.131	0.153	0.22	0.262
	1920 - 1924	0.009	0.052	0.055	0.086	0.09
20 - 27	< 1910	NA	0.432	1.506	1.609	1.193
	1910 - 1914	0.004	2.429	2.666	2.74	1.525
	1915 - 1919	0.044	0.663	0.629	0.621	0.402
	1920 - 1924	0.166	0.874	0.814	1.003	0.681
27.5 - 34.9	< 1910	NA	0.65	1.702	1.799	0.607
	1910 - 1914	0.008	3.415	3.193	2	0.683
	1915 - 1919	0.059	0.996	0.766	0.689	0.017
	1920 - 1924	0.248	1.238	0.964	0.706	0.26
> 35	< 1910	NA	0.212	0.611	0.528	NA
	1910 - 1914	0.017	4.385	1.741	0.412	NA
	1915 - 1919	0.106	1.879	1.214	0.141	NA
	1920 - 1924	0.333	1.483	0.768	0.17	NA

Note que os valores estimados na tabela 7 foram próximos dos observados (comparação com a tabela 1. Seguindo a mesma tendência de frequências mais elevadas para certas combinações. Para melhor visualização do efeito de cada variável explanatória são apresentados, na figura 3 as probabilidades de ocorrência de um grid de números de casos para cada categoria de cada variável presente no modelo.

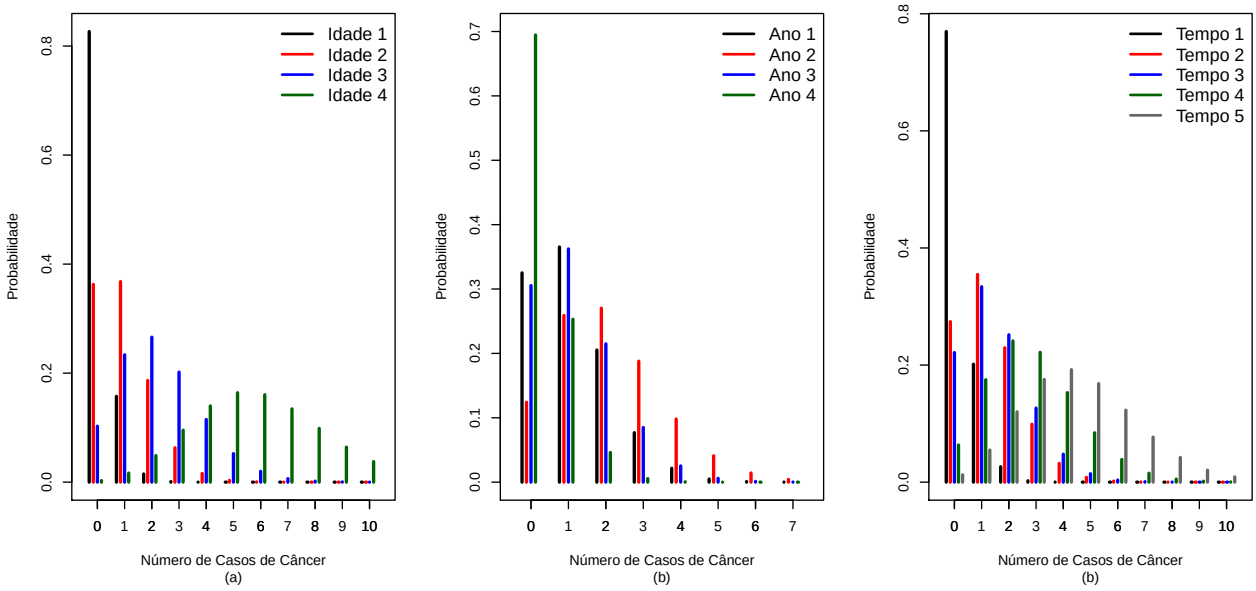

Figure 3: Probabilidades Estimadas de Ocorrência de Casos de Câncer

Os gráficos apresentados na figura 3 foram contruídos sob a fixação de idade no primeiro emprego em 20-27 ano, ano na categoria 1910-1914, tempo decorrido >50 e o total de pessoas-ano de observações na média da combinação destas categorias fixadas, para cada gráfico variou-se conforme as categorias da variável exibida. Perceba que para a idade do indivíduo no primeiro emprego são estimadas probabilidades mais elevadas para as primeiras categorias, figura 3 (a). Note a grade diferença entre a distribuição sob a categoria 1, < 20 anos, e a categoria 4, > 35 anos, a taxa de incidência de casos de câncer na categoria 4 é estimada ser aproximadamente

30 vezes a taxa sob a categoria 1. Para o ano do primeiro emprego do indivíduo apresentado na figura 3 (b) não uma ordenaçao sistemática quanto a taxa, porém notamos que a taxa mais elevada é observada na categoria 2 (1910-1914) e a mais baixa está sob a categoria 4 (1920-1924). Sob estas categorias estima-se que a taxa na categoria 2 é aproximadamente 6 vezes a taxa estimada sob a categoria 4. No último gráfico, figura 3 (c), percebemos que para as categorias da variável tempo decorrido desde o primeiro emprego temos taxas crecentes em relação a ordem das categorias, categorias que representam um tempo decorrido mais elevado apresentam taxas, também, mais elevadas. A razão estimada entre as taxas da última categoria em relação a primeira é de aproximadamente 17, ou seja, estima-se que a taxa de número de casos para a última categoria do tempo decorrido (>50 anos) é aproximadamente 17 vezes a taxa de número de casos sob a categoria 1 (0-19 anos).

Contudo, percebmos que as relações puderam ser estudados e os dados sumarizados. Como sugestão para estudos futuros é recomendável que as variáveis idade do indivíduo no primeiro emprego e tempo decorrido desde o último emprego sejam coletadas em mensuração intervalar, haja visto que os efeitos estimados são crecentes seguindo a mesma ordem das categorias.