Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Modelos Lineares Generalizados Log Linear

Douglas de Paula Nestlehner

Abril, 2022

Capítulo 1

Problema Apresentado

Considerando os dados representados na Tabela 1.1:

–	Minneapolis		Dallas	
Idade	Casos	Pop	Casos	Pop
5-24	1	172675	4	181343
25-34	16	123065	38	146207
35-44	30	96216	119	121374
45-54	71	92051	221	111353
55-64	102	72159	259	83004
65-74	130	54722	310	55932
75-84	133	32185	226	29007
85+	40	8328	65	7538

Tabela 1.1: Dados do problema

- Faça o gráfico log(contagem/pop) x Idade por cidade (use diferentes cores para os pontos)
- Considere a variável Idade como quantitativa. Modele a taxa de incidência com respeito a idade, claro, incluindo a covariável Cidade. Faça o envelope, discute os resultados
- Considere a variável Idade como Fator. Modele a taxa de incidência com respeito a idade (incluindo Cidade). Faça o envelope, discute os resultados
- Para o melhor modelo, compare as duas cidades (use OR)

1.1 Gráfico

Afim de observar o comportamento das observações, plotamos o gráfico log(contagem/pop) x Idade por cidade, obtendo a Figura 1.1

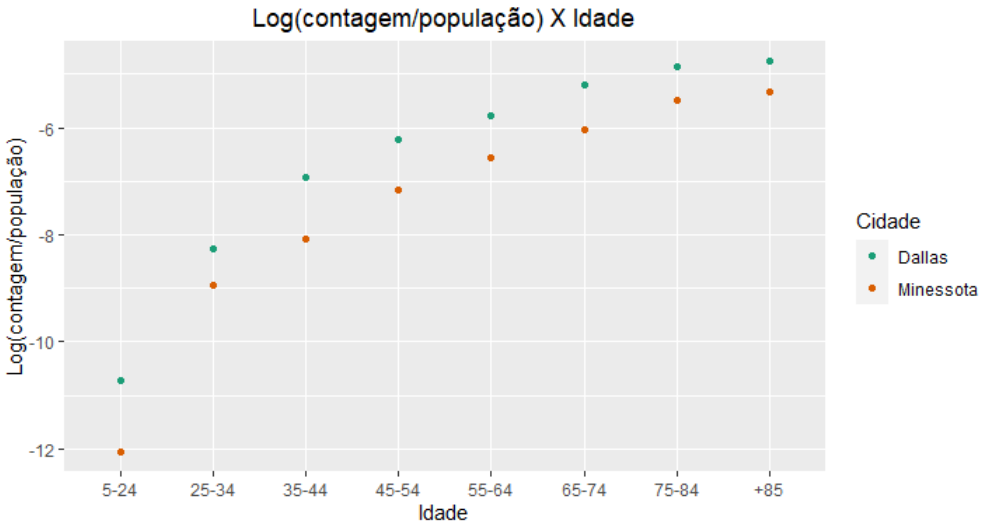

Figura 1.1: Gráfico Log(Taxa de Incidencia) X Idade

Observa-se que, o log da taxa de incidência (Numero de casos / Numero da população) para as duas cidades tem uma breve diferença em todas as faixas etárias, tendo sempre como maior log(Taxa) a cidade de Dallas.

Em ambas as cidades as curvas apresentam um formato de gráficos log-linear.

1.2 Ajuste do Modelo

Foi observado no gráfico das variáveis resposta um comportamento log-linear, e além disso estamos trabalhando com dados representado em uma tabela de contingencia, assim sendo, para o ajuste do modelo, iremos considerar modelos log-lineares.

A variável considerada como resposta é o log(taxa de incidência) para as cidades de Minneapollis e Dallas, trata-se portanto de uma variável quantitativa (taxa). Para problemas como este, devemos considerar uma distribuição que comporte tais características.

Desse modo, os modelos a serem ajustado são modelos log-lineares Poisson, no seguinte formato:

$$log(\text{Taxa de Incidencia}) = \beta_0 + \beta_i Idade_i + \beta_{i+1}\text{Cidade}$$

Em que i, é o numero de faixa etárias na variável Idade. Caso considerarmos a variável Idade como quantitativa, i=1.

Entretanto, o modelo Poisson é um tipo de modelo especifico para Contagens, e estamos trabalhando com Taxa. Porem podemos manipular o modelo definido acima, para um modelo de contagem.

$$log(\text{Taxa de Incidencia}) = \beta_0 + \beta_i Idade_i + \beta_{i+1}\text{Cidade}$$

$$log(\text{Casos/População}) = \beta_0 + \beta_i Idade_i + \beta_{i+1}\text{Cidade}$$

$$log(\text{Casos}) - log(\text{População}) = \beta_0 + \beta_i Idade_i + \beta_{i+1}\text{Cidade}$$

Ou seja, o modelo a ser ajustado é:

$$log(\text{Casos}) = \beta_0 + \beta_i Idade_i + \beta_{i+1}\text{Cidade} + log(\text{População})$$

Que nada mais é que o modelo poisson incluindo o offset, caso especifico para modelagem de taxas.

1.2.1 Ajuste do Modelo (Idade Quantitativa)

Para fins de comparação iremos ajustar dois modelos, sendo um considerando a variável Idade como quantitativa, e outro considerando como qualitativa (considerando os intervalos).

Como temos na variável Idade, faixa etárias (qualitativa), optamos por representar essa variável pela média dos intervalos, para assim, transforma-la em quantitativa.

–	Idade (Qualitativa)	Idade(Quantitativa)
1	5-24	15
2	25-34	30
3	35-44	40
4	45-54	50
5	55-64	60
6	65-74	70
7	75-84	80
8	85+	85

Tabela 1.2: Variável Idade

Ajustando o modelo, obtemos as seguintes estimativas:

—	Estimate Std.	Error	z-value	$Pr(>\|z\|)$
(Intercept)	-9.605493	0.092736	-103.58	$< 2e-16 * **$
idades	0.061831	0.001398	44.23	$< 2e-16 * **$
cidadeMinneapolis	-0.818878	0.052181	-15.69	$< 2e-16 * **$

Tabela 1.3: Coeficientes estimados

AIC = 225,92

Inicialmente o modelo aparenta estar bem ajustado, porem precisamos verificar por meio da analise de diagnostico. Nesse caso iremos utilizar apenas o gráfico de envelope para a tomada de decisão.

Plotando o gráfico de envelope, obtemos o resultado representado na Figura 1.2:

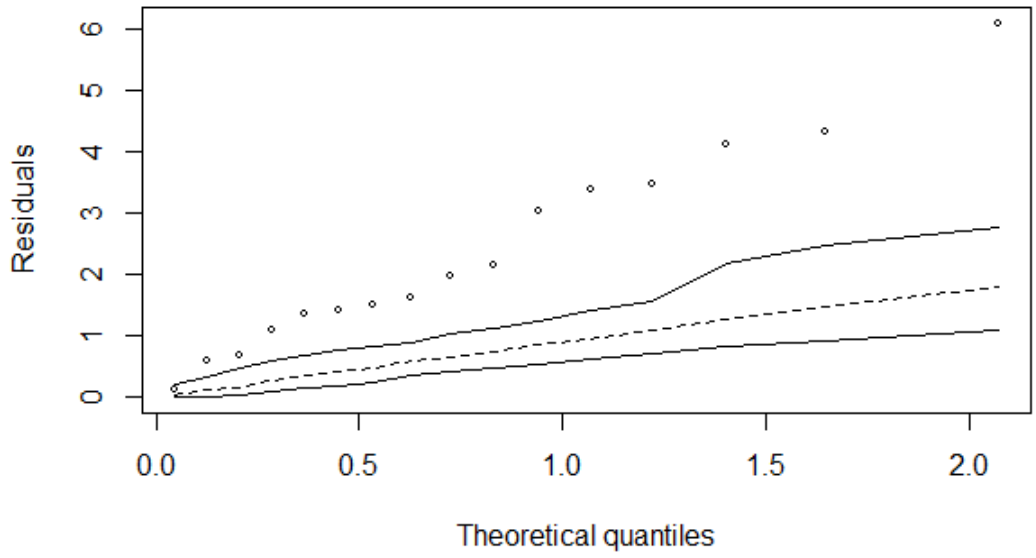

Figura 1.2: Gráfico Envelope para o modelo considerando Idade como quantitativa.

Em que é possível observar que praticamente todos os pontos estão fora do envelope, o que nos indica que o modelo não é adequado.

1.2.2 Ajuste do Modelo (Idade Qualitativa)

Apos ajustar o modelo considerando a variável Idade como quantitativa, e checando que o modelo não adequado. Realizamos o ajuste do modelo considerando Idades como Qualitativa, o que realmente é o mais adequado, pois estamos trabalhando com faixa etárias.

Obtemos então as seguintes estimativas para o modelo:

—	Estimate Std.	Error	z-value	$Pr(>\|z\|)$
Intercept	-4.67541	0.09911	-47.176	$< 2e-16$ ***
idades25-34	-3.54800	0.16749	-21.184	$< 2e-16$ ***
idades35-44	-2.33084	0.12747	-18.286	$< 2e-16$ ***
idades45-54	-1.58300	0.11384	-13.906	$< 2e-16$ ***
idades5-24	-6.17819	0.45774	-13.497	$< 2e-16$ ***
idades55-64	-1.09091	0.11091	-9.836	$< 2e-16$ ***
idades65-74	-0.53277	0.10862	-4.905	$9.35e-07$ ***
idades75-84	-0.11964	0.11095	-1.078	0.281
cidadeMinneapolis	-0.80428	0.05221	-15.406	$< 2e-16$ ***

Tabela 1.4: Coeficientes estimados

AIC $= 120.44$

Quase todas as variáveis do modelo são significantes, aparentando ser um bom modelo. Para poder verificar se realmente o modelo foi bem ajustado, construímos o gráfico do envelope, representado na Figura 1.3.

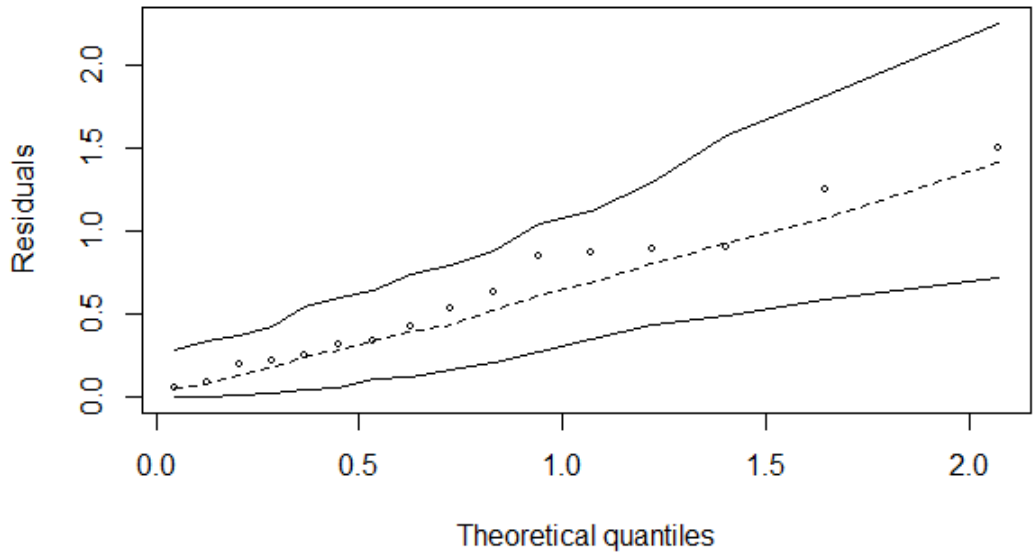

Figura 1.3: Gráfico Envelope para o modelo considerando Idade como qualitativa.

Diferentemente do primeiro modelo ajustado (considerando Idade com quantitativa), esse gráfico de envelope tem todos os pontos dentro do envelope, forte indicativo que esse modelo é adequado.

Outros grafico/testes poderiam ser elaborados para o diagnostico do modelo, entretanto considerei o resultado obtido na Figura 1.3, suficiente para definir o modelo com variável Idade qualitativa, como o melhor modelo entre os dois ajustados.

Desse modo, temos como modelo final:

$$log(casos) = -4.6754 - 3.5480 * I(\text{Idade 25-34}) - 2.3308 * I(\text{Idade 35-44})$$
$$-1.5830 * I(\text{Idade 45-54}) - 6.1781 * I(\text{Idade 5-24})$$
$$-1.0909 * I(\text{Idade 55-64}) - 0.5328 * I(\text{Idade 65-74})$$
$$-0.1196 * I(\text{Idade 75-84}) - 0.8043 * I(\text{Cidade = Minneapolis})$$
$$+log(População)$$

Equivalente a:

$$log(TaxadeIncidência) = -4.6754 - 3.5480 * I(\text{Idade 25-34}) - 2.3308 * I(\text{Idade 35-44})$$
$$-1.5830 * I(\text{Idade 45-54}) - 6.1781 * I(\text{Idade 5-24})$$
$$-1.0909 * I(\text{Idade 55-64}) - 0.5328 * I(\text{Idade 65-74})$$
$$-0.1196 * I(\text{Idade 75-84}) - 0.8043 * I(\text{Cidade = Minneapolis})$$

Em que I representa a função indicadora, exemplo I(Idade 25-34): caso a faixa etária seja 25-34 a função assume valor igual a 1, caso contrario assume 0.

1.3 Comparação Cidades

Apos encontrar o melhor modelo, temos o interesse de realizar a comparação entre as duas cidades presentes no estudo, Minneapolis e Dallas.

Para isso iremos utilizar o Odds Ratio (OR), o qual é calculado por:

$$OR = exp(coeficients)$$

Resultados obtidos:

—	OR
Intercept	0.009321693
idades25-34	0.028782249
idades35-44	0.097214490
idades45-54	0.205357593
idades5-24	0.002074182
idades55-64	0.335911786
idades65-74	0.586974542
idades75-84	0.887238308
cidadeMinnapolis	0.447412180

Tabela 1.5: Coeficientes estimados OR

Interpretando os resultados, temos na primeira linha que o risco de uma pessoa em Dallas ter um caso (doença) é de 0,009.

Na ultima linha temos que, comparado com Dallas o risco de um caso em Minneapolis é 0,4474 vezes maior (no caso seria menor se multiplicarmos) que Dallas.