Modelos Lineares Generalizados: Aplicação a Dados de Câncer

Mariana Costa Freitas

Introdução

O trabalho a seguir tem como objetivo ajustar modelos estatísticos para estimar o número de óbitos por câncer em diferentes regiões do Canadá, considerando os fatores de gênero, local do câncer e população da região.

O conjunto de dados utilizado, ccancer, contém informações sobre o número estimado de mortes por câncer em 2000 para as províncias de Ontário, Newfoundland e Quebec. O banco de dados apresenta 30 observações e cinco variáveis:

- Count: número estimado de óbitos para um determinado tipo de câncer;
- Gender: gênero dos pacientes (Feminino ou Masculino);
- Region: região onde o óbito foi registrado (Ontário, Newfoundland ou Quebec);
- Site: local do câncer (Pulmão, Colorretal, Mama, Próstata ou Pâncreas);
- Population: população estimada da região em 2000/2001.

Metodologia

A análise dos dados foi realizada por meio de modelos lineares generalizados (MLGs), uma extensão dos modelos de regressão linear que permite a modelagem de variáveis resposta que seguem distribuições distintas da normal. Os MLGs são definidos pela relação: $g(E[Y]) = X\beta$, onde $g(.)$ é a função de ligação, $E[Y]$ é o valor esperado da variável resposta, X é a matriz de covariáveis e β é o vetor de parâmetros a ser estimado.

Dado que o número de óbitos por câncer é uma variável de contagem, inicialmente consideramos a modelagem utilizando a distribuição de Poisson, cuja função de massa de probabilidade é dada por:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

onde λ representa o valor esperado da variável resposta. No entanto, a distribuição de Poisson assume que a variância é igual à média, o que pode não ser adequado caso haja superdispersão nos dados.

Para lidar com esse problema, consideramos o modelo Binomial Negativo, que também é apropriado para dados de contagem e permite a presença de superdispersão. A função de massa de probabilidade da distribuição Binomial Negativa é:

$$P(Y = y) = \frac{\Gamma(y + \theta)}{y!\Gamma(\theta)} \left(\frac{\mu}{\mu + \theta}\right)^{y} \left(\frac{\theta}{\mu + \theta}\right)^{\theta},$$

onde μ é a média da variável resposta e θ é o parâmetro de dispersão.

Após a seleção do modelo mais adequado, foi realizada uma avaliação de ajuste por meio da análise de resíduos. Foram utilizados:

- Resíduos de Pearson, que verificam discrepâncias entre os valores observados e ajustados;
- Medida H, que avalia a influência de cada observação no ajuste do modelo;
- Distância de Cook, que identifica observações influentes;
- Análises gráficas, incluindo gráficos de resíduos padronizados e envelopes simulados para verificar a adequação do modelo.

Análise dos dados

Análise Descritiva

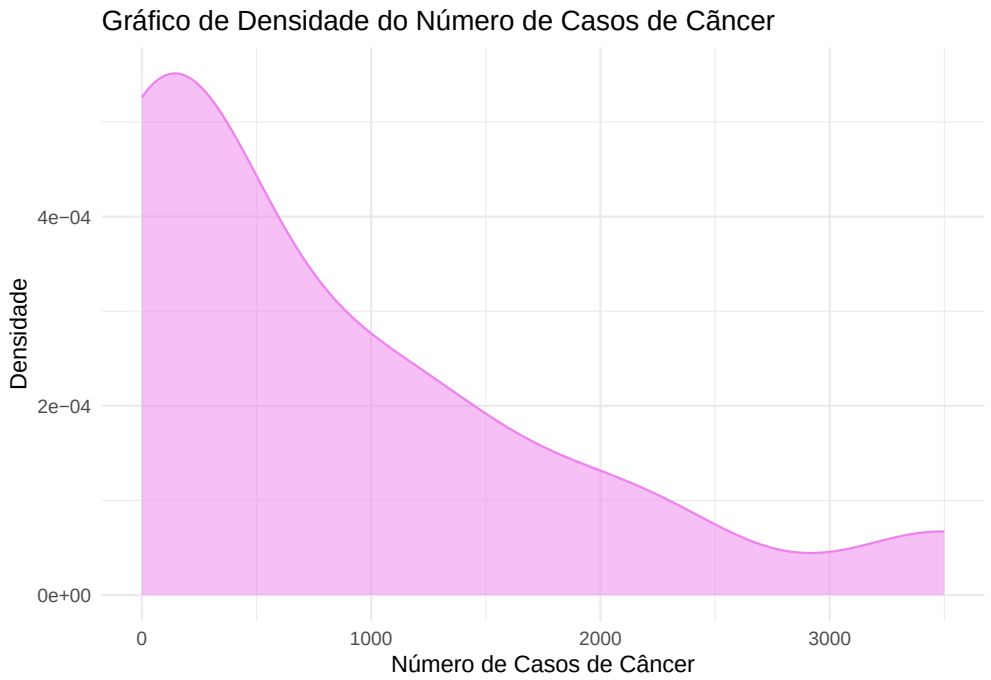
Para escolher um vom modelo inicial, é necessário primeiro fazer uma análise descritiva dos dados, a fim de melhor compreender a sua natureza e, então, encontrar também um modelo adequado para descrever o conjunto de dados. A seguir, temos a apresentação de medidas de resumo para ambas as variáveis quantitativas e qualitativas relacionadas ao número de casos de câncer.

Variável	Média	Mediana	Desvio	Mínimo	1º Quartil	3º Quartil	Máximo
Número de casos de câncer	814.8333	400	1,019.601	0	31.25	1,212.5	3,500
População	6,606,233.333	7,410,500	4,744,320.433	533,800	533,800.00	11,874,400.0	11,874,400

Contagem	Proporção (%)	Categoria
15	0.50	Gênero: Femino
15	0.50	Gênero: Masculino
6	0.20	Local do câncer: Mama
6	0.20	Local do câncer: Colo retal
6	0.20	Local do câncer: Pulmão
6	0.20	Local do câncer: Pâncreas
6	0.20	Local do câncer: Próstata
10	0.33	Região: Newfoundland
10	0.33	Região: Ontário
10	0.33	Região: Quebec

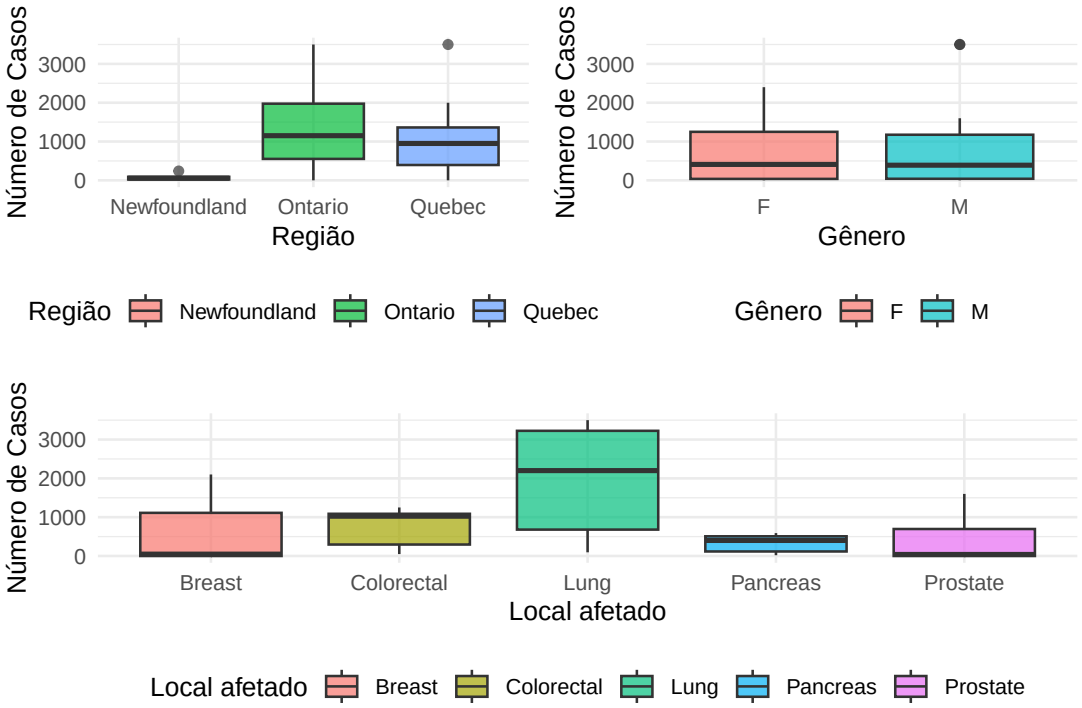
A partir da primeira tabela, é possível notar que o número de casos de câncer apresenta um alto desvio em comparação com a média, com 25% das contagens abaixo ou igual a 31.25 e 25% acima de 1212.5. Já a população ainda apresenta desvio alto, porém menor que a média, sendo que 25% dos dados têm população entre 533,800 e 25% acima de 11,874,400.0. Na segunda tabela, observamos que todas as variáveis quantitativas do conjunto de dados apresentam o mesmo número de observações por nível do fator.

A seguir, foi criado um gráfico de densidade para a variável de interesse, número de casos de câncer, a fim de melhor identificar um modelo adequado para os dados.

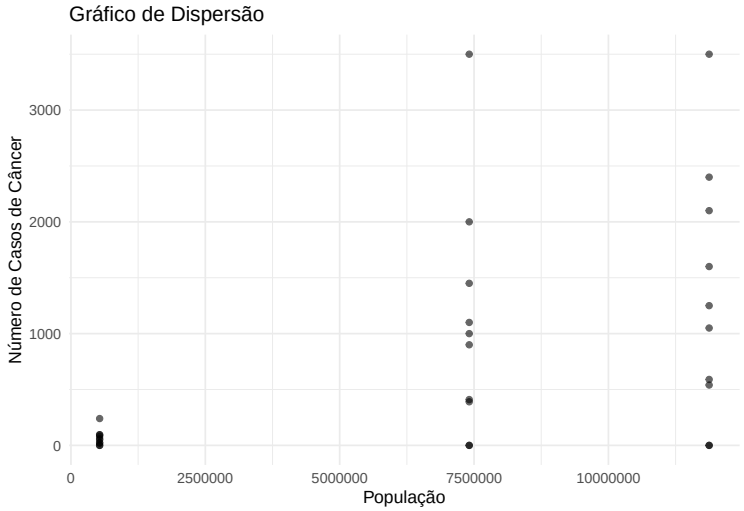


A curva de densidade nos mostra que a distribuição apresenta uma assimetria à direita, que pode indicar que há uma concentração das observações em valores mais baixos, mas há também a presença de valores mais altos e alguns possíveis *outliers*, responsáveis pela longa cauda à direita. O gráfico condiz com a alta variabilidde dos dados de contagem mostrada anteriormente ao apresentar as medidas resumo.

Outra importante análise inicial, é observar como a variável de interesse, no caso, número de casos de câncer, se comporta em diferentes níveis ou valores do restante das variáveis. A seguir, os boxplots indicam essa relação de acordo com as variáveis qualitativas e o gráfico de dispersão com a única variável quantitativa, população.



Analisando com base na região, notamos que Newfoundland apresenta uma média de casos de câncer extremamente baixa e variância do número de casos de câncer bem baixa também; já Ontario e Quebec apresentam média do número de casos parecida, porém Ontario apresenta terceiro quartil mais alto e variância também mais elevada. Em relação ao gênero, as medidas do número de casos para ambos os gêneros é bem parecida, com média e quartis bem próximos. Quanto ao local afetado pelo câncer, a maioria dos locais têm variabilidade baixa e quartis próximos, exceto para o local 'Pulmão', com alta média e variabilidade do número de casos quando comparada aos outros locais.



Como estamos avaliando o número de casos em três regiões, Ontario, Newfoundland e Quebec, temos apenas três valores distintos para a população no eixo X, enquanto o número de casos se distribui no eixo Y. É

possível observar que, para o menor valor de população, o numero de casos se concentra em valores bem pequenos, e conforme a população aumenta, a variabilidade do número de casos aumenta, apresentando tanto valores altos e baixos, mas se concentrando um pouco mais em valores medianos.

A análise descritiva das variáveis é importante na escolha do modelo mais adequado aos dados, visto que nos ajuda a melhor compreender a distribuição, a relação entre as variáveis e as características dos dados, o que é necessário para selecionar a família de distribuição e a função de ligação corretas. A nossa variável de interesse aqui, número de casos de câncer, representa uma contagem, situação em que, em geral, é utilizada uma distribuição Poisson ou uma distribuição Binomial Negativa. Para decidir qual será usada nesse caso, precisamos verificar se o pressuposto de Poissson de média igual a variância é atendido. Para isso, foi criada a tabela a seguir com as informações de média e variância para a variável de interesse.

Table 3: Média e Variância da variável 'mpg'

Estatística	Valor
Média	814.8333
Variância	1,039,585.3161

A partir dessas informações, inferimos que os dados apresentam sobredispersão, ou seja, a variância é muito maior que a média. Logo, não podemos usar a distribuição Poisson e vamos optar pela distribuição Binomial Negativa, que é capaz de se adequar a essa sobredispersão.

Ajuste do modelo

Decidido o modelo a ser utilizado, é hora de determinar a função de ligação para o modelo. Para modelos de contagem, a ligação logarítmica é mais comum e apropriada para modelos de contagem, pois garante que os valores preditos sejam sempre positivos; mas as funções de ligação identidade e raíz quadrática também são utilizadas para esses casos. Já as funções de ligação logito, probito e complementary log log não podem ser utilizadas para dados de contagem, pois assumem variáveis respostas binárias ou no intervalo [0,1], não sendo adequadas nesse caso.

Apesar da possibilidade de usá-las para dados de contagem, quando testadas aqui, as funções de ligação identidade e raíz quadrática não encontraram uma combinação de coeficientes válidos para ajuste do modelo. Assim, foi usada a função de ligação logarítmica.

A seguir, vamos testar algumas possibilidades desse modelo, testando variáveis e a interação entre elas, para verificar qual seria a melhor escolha nesse caso.

Id	Modelo	Desvio	Graus de liberdade	2xLog-verossimilhança
1	1	36.87	29	-418.0826
2	gender	36.87	28	-418.0327
3	gender + region + site	9,375.13	22	-9,565.9311
4	gender + region + site + population	9,375.13	22	-9,565.9311
5	gender * site + region + population	21.95	18	-255.7487
6	gender * site + region	21.95	18	-255.7487

Id	Modelo	Desvio	Graus de liberdade	2xLog-verossimilhança
7	gender + site * region + population	9,191.16	14	-9,381.9592
8	gender * site * region + population	0.00	0	-190.8263

Na tabela acima, temos as informações das variáveis e interações que descrevem cada modelo, e suas respectivas métricas de ajuste, como o desvio, os graus de liberdade residuais e o logaritmo da verossimilhança.

O Modelo 1, que inclui apenas o intercepto, foi descartado por ser muito simples. Com um desvio de 36.87 e um logaritmo da verossimilhança de -418.0826, ele não consegue capturar a variabilidade dos dados de forma adequada. O Modelo 2, que adiciona a variável "gender", também foi descartado, pois não houve melhoria significativa no ajuste em relação ao Modelo 1, mantendo um desvio alto e um logaritmo da verossimilhança muito próximo ao anterior.

Os Modelos 3 e 4, que incluem as variáveis "gender", "region", "site" e, no caso do Modelo 4, "population", apresentaram desvios extremamente altos (9375.13) e logaritmos da verossimilhança muito negativos (-9,565.9311). Esses valores indicam que esses modelos têm um ajuste muito ruim aos dados, mesmo com a inclusão de mais variáveis. Assim, ambos foram descartados.

Por outro lado, o Modelo 7, que inclui a interação entre "site" e "region", além de "gender" e "population", foi descartado devido ao seu desvio muito alto (9191.16) e ao logaritmo da verossimilhança muito negativo (-9,381.9592). A complexidade adicional dessa interação não melhorou o ajuste do modelo.

O Modelo 5, que inclui a interação entre "gender" e "site", além das variáveis "region" e "population", apresenta desvio de 21.95 e um logaritmo da verossimilhança de -255.7487, tendo então um ajuste significativamente melhor em comparação aos modelos anteriores. Porém, o Modelo 6, que retira a variável "population", ainda considerando a interação entre "gender" e "site", apresenta as mesmas medidas de ajuste que o Modelo 5, porém com uma variável a menos, o que diminui a complexidade do modelo, sendo então o modelo mais adequado.

Já o Modelo 8, que inclui a interação tripla entre "gender", "site" e "region", além de "population", foi descartado por apresentar um desvio zero, o que sugere sobreajuste, ou seja, é inadequado para generalização e provavelmente terá um desempenho ruim em novos dados.

Com o modelo já selecionado, agora vamos interpretar seus coeficientes mostrados abaixo.

Variáveis	Coeficientes
Intercepto	4.56
Gênero: Masculino	-41.71
Local: Colo retal	-0.55
Local: Pulmão	0.18
Local: Pâncreas	-1.27
Local: Próstata	-43.56

Variáveis	Coeficientes
Região: Ontário	3.03
Região: Quebec	2.79
Interação: Masculino e Colo retal	41.85
Interação: Masculino e Pulmão	42.29
Interação: Masculino e Pâncreas	41.63
Interação: Masculino e Próstata	84.95

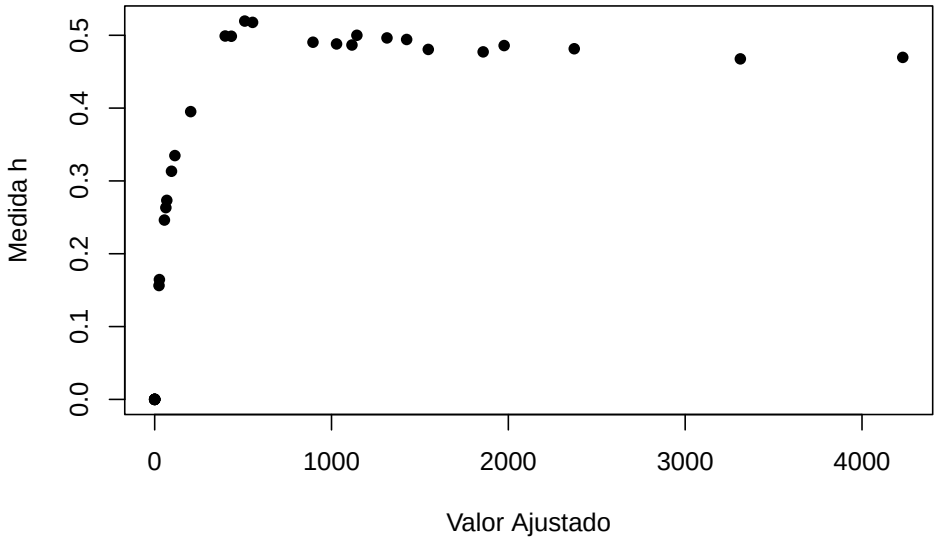
O intercepto é o valor esperado da variável resposta quando todas as variáveis preditoras são iguais a zero, ou seja, quando o gênero é feminino, local do câncer é mama, e a região é Newfoundland, o logaritmo do número de casos de câncer é 4.56. Assim, o número esperado de casos de câncer nessas condições é $e^{4.56} = 95.6$. Já os coeficientes das variáveis principais indicam o efeito de cada variável, enquanto os coeficientes de interação mostram como o efeito de uma variável muda dependendo do valor de outra.

Por exemplo, o coeficiente para o local Pulmão 0.18 indica a diferença no log-odds para o local "Pulmão" em relação ao local "Mama". Um valor positivo sugere que, em comparação com o número de casos de câncer de mama, o local "Pulmão" está associado a um log-odds maior, ou seja, maior probabilidade de casos de câncer. Já no caso de uma interação, o coeficiente 41.85 representa a interação entre o gênero ser masculino e o local de câncer ser colo retal. Ele indica que o efeito de ter essas duas características é 41.8492504 unidades maior do que o efeito base delas individualmente. Seguindo o mesmo raciocínio, é possível interpretar os outros coeficientes.

Análise de resíduos

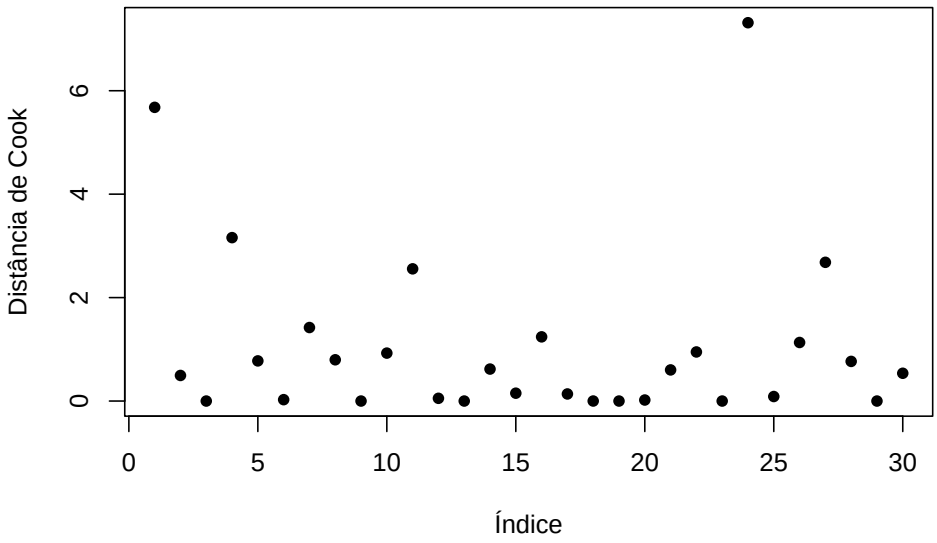
Escolhido o modelo, agora faremos uma análise de resíduos, com o intuito de verificar a adequação do modelo aos dados e identificar possíveis problemas.

A seguir, vamos analisar a presença de possíveis pontos de alavanca, ou seja, observações que, por possuirem valores extremos nas variáveis preditoras, têm um impacto maior no modelo do que outras observações, podendo então distorcer os resultados do modelo. Para identificá-los, vamos usar a medida H. A medida H varia entre 0 e 1, sendo que um valor de H próximo de 1 indica que a observação tem um alto potencial de alavancagem no modelo, enquanto um valor próximo de 0 indica pouca. Abaixo temos um gráfico de dispersão que apresenta os valores ajustados e o valor H.



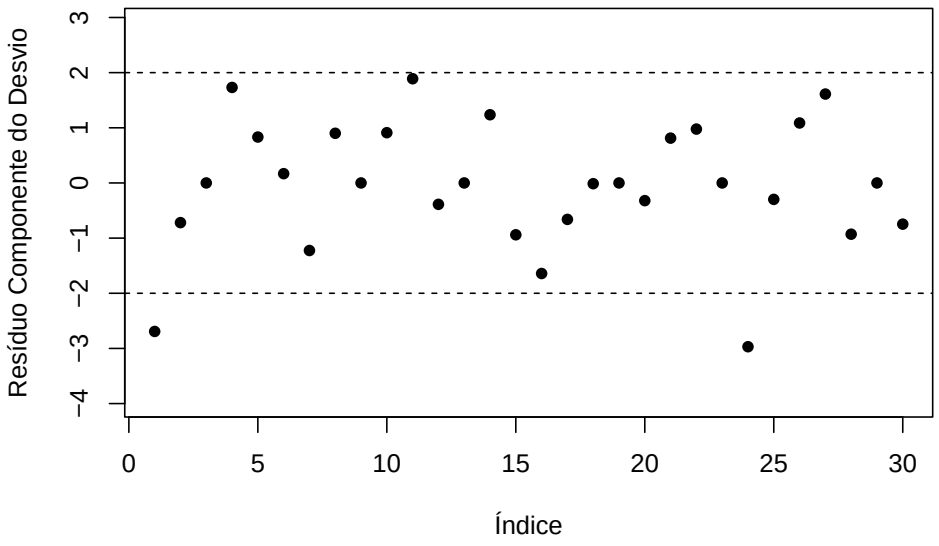
Podemos observar que as medidas H estão bem próximas de zero, sendo que o maior valor H obtido foi pouco mais de 0.08. Assim, não identificamos pontos de alavanca.

Outra métrica de ajuste importante é a Distância de Cook, utilizada para detectar observações influentes que podem estar distorcendo o modelo. Valores altos indicam que a remoção da observação do conjunto de dados causaria uma mudança significativa nas estimativas dos coeficientes do modelo.



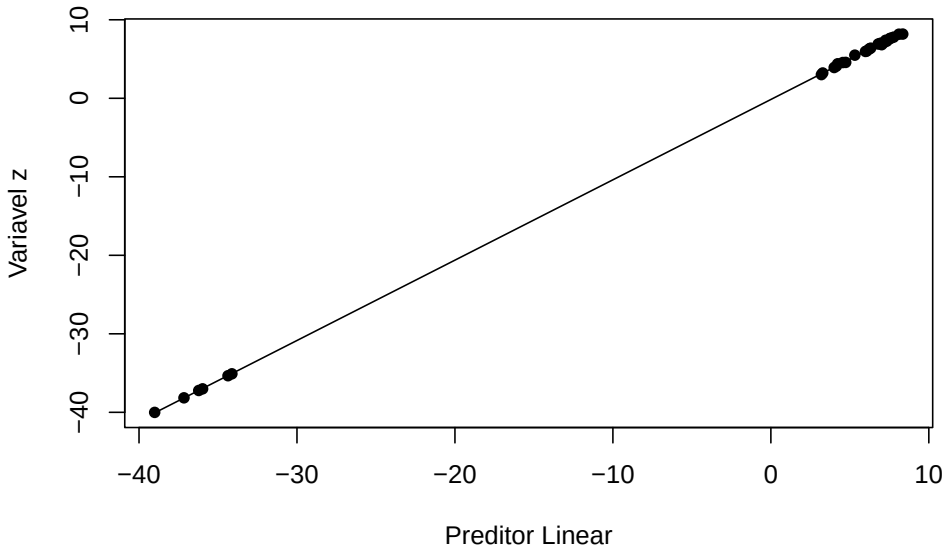
Em geral, os pontos apresentaram uma baixa Distância de Cook, sugerindo que têm pouca influência no modelo. Logo, não afetam significativamente as estimativas dos coeficientes.

Também vamos observar o gráfico de resíduos do componente do desvio, que mostra esses resíduos em função do índice das observações. Se o modelo estiver adequado, os resíduos devem se distribuir aleatoriamente em torno de zero.



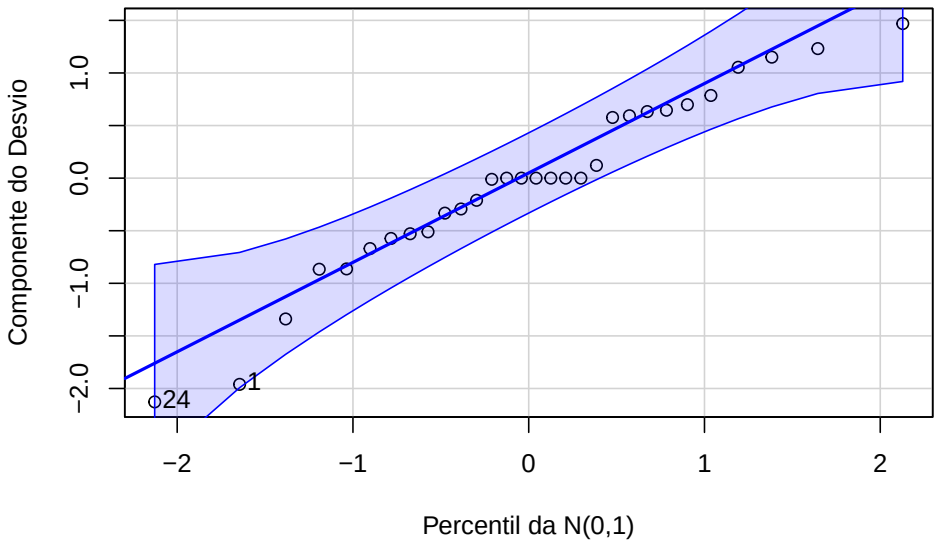
No gráfico, os pontos estão espalhados sem um padrão claro, sugerindo que o modelo está capturando bem a estrutura dos dados e que não há alguma tendência nos resíduos.

Outra análise importante é a comparação do preditor linear com a Variável z, $z = \eta + \frac{\text{resíduo de Pearson}}{\sqrt{\text{pesos do modelo}}}$. O valor z é útil para verificar se o modelo está bem ajustado porque, sob um bom ajuste do modelo, ele deve se alinhar com o preditor linear η.



O gráfico sugere que o modelo se ajustou bem aos dados, já que os pontos seguem a reta de referência.

Por último, vamos analisar um gráfico de envelope para o nosso modelo, que é uma versão do gráfico de quantis-quantis que inclui uma faixa de confiança simulada. Ele é utilizado para avaliar visualmente se os resíduos do modelo seguem uma distribuição teórica esperada, no caso a normal padrão.




```
## [1] 24  1
```

Observamos que os resíduos estão razoavelmente alinhados com a reta, o que sugere que o modelo binomial negativo está ajustado adequadamente. Além disso, em geral, os pontos estão dentro do envelope, ou seja, não há evidências de violação da normalidade dos resíduos.

Conclusão

A análise dos dados revelou que o modelo Binomial Negativo apresentou o melhor ajuste para descrever o número de óbitos por câncer nas regiões estudadas. A distribuição de Poisson, inicialmente considerada, mostrou superdispersão, o que justificou a adoção da distribuição Binomial Negativa. A análise de resíduos indicou um ajuste satisfatório, sem a presença de padrões sistemáticos que comprometessem a validade do modelo. Assim, a modelagem estatística, permite a identificação de fatores associados à mortalidade por câncer.

Referências

PAULA, Gilberto A. Modelos de regressão com apoio computacional. São Paulo: IME/USP, 2013.

Gauss M. Cordeiro e Clarice G.B. Demétrio. Modelos Lineares Generalizados e Extensões. Piracicaba: USP, 2008.