Relatório: Modelo Poisson

Jeff Caponero, Michel Miler Rocha dos Santos e Camille Menezes dos Santos

Sumário

Introdução

Por muitos anos, os modelos de regressão linear foram amplamente empregados para tentar descrever uma variedade de fenômenos aleatórios, mesmo quando os pressupostos desses modelos eram violados. Transformações na variável resposta eram realizadas a fim de contornar esses problemas, principalmente ao que tange a violação da normalidade, como a transformação Box-Cox. Concomitante a isto, propor uma solução cuja a estimação dos parâmetros dependesse de um processo iterativo seria muito custoso computacionalmente à época.

A partir da década de 1970, onde ocorreu um desenvolvimento maior da capacidade computacional, foi possível propor outras abordagens para modelos de regressão. Surgiu uma ampliação conceitual dos modelos lineares, por exemplo, a classe dos Modelos Lineares Generalizados (MLGs), que englobam os modelos lineares tradicionais como casos particulares. Os MLGs permitem acomodar diversas distribuições de respostas, não se restringindo apenas à normal. Essa mudança de paradigma abriu espaço para a modelagem de dados que não seguem distribuições normais, como dados de contagem (Poisson), binomiais e proporções, entre outros.

A menção à "família exponencial uniparamétrica de distribuições" refere-se a um conjunto de distribuições que engloba várias distribuições comuns, incluindo a normal, Poisson, Binomial e Gama. Essa família de distribuições possui propriedades matemáticas vantajosas, conferindo eficiência e flexibilidade à modelagem estatística.

Modelo Poisson

Seja Y uma variável aleatória com distribuição Poisson de média μ, denotamos Y Poi(µ). A função densidade de Y é dada por

$$f(y;\mu;\phi) = exp\left\{y.log(\mu) - \mu - log(y!)\right\}$$

Logo, aplicando a definição de Família Exponencial adotada para MLG, temos que:

$\theta = log(\mu)$,

$\phi = 1$,

$b(\theta) = e^{\theta}$

$c(y,\phi) = -log(y!)$.

Para a estimação de parâmetros é possível estabelecer um procedimento iterativo, junto com a introdução do conceito de desvio, o qual tem sido amplamente empregado na avaliação da adequação dos Modelos Lineares Generalizados (MLGs). Esse conceito também desempenha um papel crucial no desenvolvimento de resíduos e métricas de diagnóstico.

Nesse processo iterativo, os parâmetros do modelo são ajustados repetidamente para otimizar a adaptação aos dados observados. O conceito de desvio, por sua vez, é uma medida que reflete a discrepância entre os dados observados e as previsões do modelo. Ao minimizar esse desvio, os parâmetros do modelo são calibrados de modo a se ajustarem melhor aos dados.

Além disso, o conceito de desvio tem uma importância significativa na avaliação da qualidade do ajuste dos MLGs. Medir o desvio entre os valores observados e os previstos pelo modelo é fundamental para determinar o quão bem o modelo se adapta aos dados. Essa avaliação é essencial para verificar se o modelo é apropriado para a situação em análise.

Os resíduos, que são as diferenças entre os valores observados e os valores ajustados pelo modelo, são derivados do conceito de desvio. Esses resíduos podem fornecer informações valiosas sobre a qualidade do ajuste do modelo e a presença de padrões não capturados pelo modelo. Através dos resíduos, é possível identificar possíveis discrepâncias entre as previsões do modelo e os dados reais.

Além disso, a utilização de medidas de diagnóstico baseadas nos desvios e resíduos é essencial para identificar possíveis problemas com o modelo, como valores atípicos, falta de ajuste ou violações das suposições do modelo. Essas medidas permitem a detecção de anomalias que poderiam afetar a confiabilidade das inferências feitas com base no modelo.

Esses resultados podem mais propriamente ser alcançados pelo uso de fuções de ligações canônicas, o que por sua vez, oferece uma série de vantagens, uma delas é a garantia de que a função de verossimilhança ($L(\beta)$) seja uma função côncava. Isso, por sua vez, simplifica a obtenção de diversos resultados assintóticos. A propriedade de concavidade da função de verossimilhança traz consigo implicações significativas, como a obtenção mais direta de resultados assintóticos.

Uma vantagem concreta é observada na garantia da unicidade da estimativa de máxima verossimilhança dos coeficientes (β's), desde que essa estimativa exista. Isso significa que, quando utilizamos ligações canônicas, há uma única estimativa que maximiza a verossimilhança dos dados observados. Isso torna o processo de estimação mais estável e confiável, pois não há ambiguidade na determinação dos parâmetros ótimos.

No entanto, quando se trata de ligações não canônicas, a situação é mais complexa. Em 1976, Wedderburn discutiu as condições sob as quais a concavidade da função ($L(\beta)$) ainda pode ser estabelecida em tais cenários. Essa discussão é importante, pois a concavidade da função de verossimilhança é um pressuposto fundamental para muitos resultados estatísticos assintóticos, que são cálculos aproximados que se tornam mais precisos com um grande número de observações.

Exemplo de aplicação - Anomalias Cromossômicas

Em 1976, Roy J. Purrott e Elaine Reeder realizaram uma pesquisa intitulada "The Effect of Changes in Dose Rate on the Yield of Chromosome Aberrations in Human Lymphocytes Exposed to Gamma Radiation." (Efeito da variação na taxa de dosagem na produção de anomalias cromossômicas em linfócitos humanos expostos a radiação gama.)

O estudo em questão aborda um tópico crucial na avaliação dos efeitos da exposição à radiação em organismos vivos, mais especificamente, o uso da dosimetria citogenética para quantificar e compreender as alterações cromossômicas que ocorrem como resultado da radiação ionizante. O foco principal recai sobre as anomalias cromossômicas dicêntricas em linfócitos humanos, que se tornaram um indicador valioso para avaliar a exposição à radiação e estabelecer limites seguros em situações de radiação ambiental ou acidentes nucleares.

O estudo teve suas raízes no trabalho pioneiro de Bender e Gooch, que propuseram que a frequência de anomalias cromossômicas dicêntricas em linfócitos humanos poderia ser utilizada como uma espécie de dosímetro biológico para a radiação. Desde então, a dosimetria citogenética evoluiu e se consolidou como uma técnica confiável na proteção radiológica. Ao longo dos anos, o laboratório responsável pelo estudo investigou mais de 200 casos de possíveis superexposições à radiação, demonstrando a utilidade e a aplicabilidade prática desse método.

A escolha das anomalias cromossômicas dicêntricas como alvo de estudo se justifica por sua frequência relativamente alta quando comparada a outras anomalias induzidas pela radiação, bem como por sua baixa incidência natural em células não irradiadas. Além disso, os dicêntricos possuem uma aparência característica e são frequentemente acompanhados por deleções acêntricas, o que fornece uma maneira adicional de confirmar sua identificação. No entanto, é importante ressaltar que a formação de dicêntricos é afetada pela taxa de dose da radiação, devido ao mecanismo de formação por quebra em duas etapas, o que significa que a proximidade das quebras em termos de espaço e tempo influencia sua formação.

Um aspecto crucial explorado no estudo é o tempo durante o qual os danos cromossômicos permanecem reativos. As estimativas variam consideravelmente, refletindo a diversidade de sistemas vegetais e animais estudados, bem como a falta de consenso para células humanas.

Diferentes pesquisadores encontraram resultados divergentes sobre o tempo necessário para o reparo das anomalias cromossômicas após a exposição à radiação. Essa variação pode ser atribuída às diferenças entre sistemas estudados e à complexidade dos processos de reparo celular.

A influência da dosimetria na formação de dicêntricos também é um aspecto fundamental abordado pelo estudo. Estudos iniciais nessa área foram prejudicados por culturas prolongadas e pela estimulação prévia das células antes da exposição à radiação. A metodologia foi aprimorada ao longo do tempo, estabelecendo que as células devem ser analisadas na primeira metáfase, que ocorre 48-54 horas após a exposição à radiação, e que as células devem ser irradiadas antes da estimulação. Experimentos realizados por diferentes grupos demonstraram que a frequência de dicêntricos é influenciada pela taxa de dose, apresentando padrões complexos em relação à dose total.

Um estudo específico dentro do escopo maior do trabalho examinou de forma detalhada como a taxa de dose afeta a formação de anomalias cromossômicas em linfócitos humanos. Diferentes doses de radiação foram administradas a taxas de dose variadas, e os resultados revelaram que tanto a frequência de dicêntricos quanto a de anomalias totais diminuem à medida que a taxa de dose diminui. A análise estatística dos dados foi realizada com base em um modelo matemático que considera a contribuição de diferentes componentes na formação das anomalias cromossômicas. Observou-se que a formação de dicêntricos diminuiu significativamente em taxas de dose mais baixas para doses mais altas. Além disso, as anomalias cromossômicas acentricas também mostraram padrões semelhantes, sugerindo que muitas delas são causadas por um processo de dois hits, ou seja, por duas lesões cromossômicas em momentos distintos.

Uma das conclusões importantes desse estudo é que a taxa de dose de radiação ionizante de baixa TLE (Transferência Linear de Energia) tem um impacto substancial na formação de anomalias cromossômicas. Especificamente, os resultados indicam que em taxas de dose abaixo de 150 rad por hora, a frequência de anomalias cromossômicas é afetada de maneira significativa. Isso é relevante porque muitas vezes a dosimetria citogenética é usada para estimar a dose equivalente total do corpo em casos de superexposição à radiação. Em situações de exposição a radiações de doses baixas, a influência da taxa de dose é menos pronunciada, pois a maioria das anomalias cromossômicas é induzida por trilhas únicas de partículas.

O estudo contribui para a compreensão mais ampla dos efeitos da radiação ionizante nas células humanas e fornece insights valiosos para o desenvolvimento de estratégias de proteção radiológica e avaliação de riscos. Além disso, ressalta a importância de considerar a taxa de dose ao usar a dosimetria citogenética como ferramenta de avaliação em cenários de exposição à radiação, especialmente em situações de baixas doses e taxas de dose variáveis. Isso pode ter implicações importantes em ambientes de risco radiológico e segurança nuclear.

Reanálise dos dados

Sobre o conjunto de dados

Os dados se referem a 27 experimentos publicados no trabalho mencionado anteriormente e é composto pelas seguintes variáveis:

ca - Quantidade de cromossomos com anomalia;
cells - Número de células amostradas;
doseamt - Quantidade total de radiação a que as células foram expostas;
doserate - Taxa de administração da radiação gama.

Análise descritiva

A tabela a seguir apresenta uma breve análise descritiva desses dados.

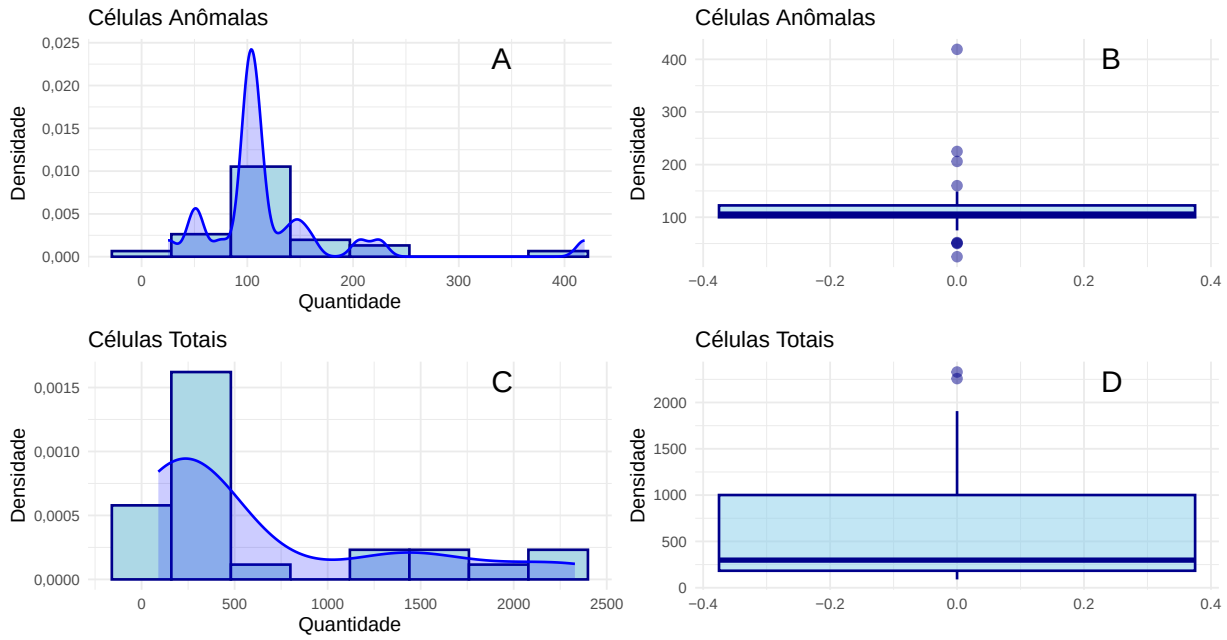
Tabela 1: Medidas resumo dos dados

	Min	Q1	Median	Mean	Q3	Max	Std.Dev	CV	Skewness	Kurtosis
ca	25,0	100,0	106,0	120,44	132,0	419	73,47	0,61	2,47	7,53
cells	90,0	182,0	298,0	640,70	1.238,0	2.329	701,62	1,10	1,23	0,02
doseamt	1,0	1,0	2,5	2,83	5,0	5	1,68	0,59	0,28	-1,61
doserate	0,1	0,5	1,5	1,65	2,5	4	1,29	0,78	0,41	-1,13

Verifica-se que a quantidade total de células e a quantidade de cromossomos com anomalia têm uma distribuição assimétrica à direita, uma vez que as suas médias são maiores que as suas medianas. O coeficiente de variação da quantidade total de células é alto, desse modo, há uma grande variabilidade nessa variável, sendo até maior que a variabilidade presente na variável quantidade de cromossomos anômalos.

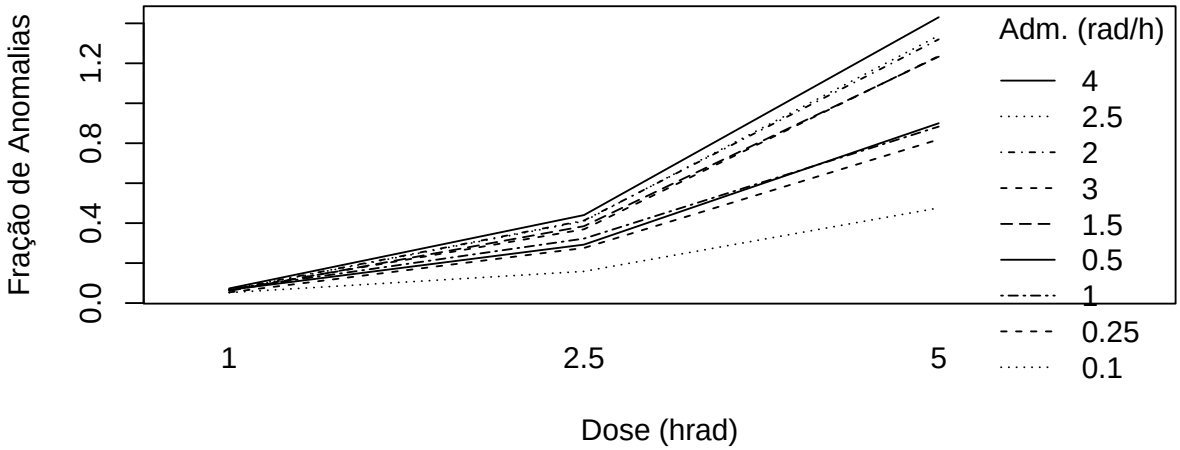
Para facilitar a compreenção das medidas de células anômalas e celulas totais, apresentadas na Tabela 1, a Figura 1 mostra graficamente estas distribuições.

Figura 1: Histogramas e Boxplots das variáveis em análise.



Os gráficos confirmam as análises de simetria e variabilidade feitas anteriormente. Alem disso, em ambas as distribuições, é possível observar outliers, sobretudo na variável dos cromossomos anômalos.

Figura 2: Interação entre a fração de anomalias e a dosimetria aplicada observadas.



Através da Figura 2, é possível notar que com o aumento da dosagem total e da taxa de

administração da radiação há um aumento da fração de anomalias.

Modelo Linear Generalizado

Função de ligação canônica

Os dados foram submetidos a uma análise de ajuste ao modelo linear generalizado com distribuição Poisson. Nesse contexto, a função de ligação utilizada foi $\eta = log(\mu)$. Os coeficientes obtidos a partir desse ajuste estão disponíveis na Tabela 2.

Tabela 2: Ajuste segundo o Modelo de Poisson com função de ligação $\eta = log(\mu)$.

	Coeficientes	EP	Estatística z	Valor-p
Intercepto	-2,818	0,056	-50,028	0,000
doserate	0,048	0,028	1,715	0,086
doseamt-2.5	1,443	0,083	17,403	0,000
doseamt-5	2,500	0,074	33,883	0,000
doserate:doseamt-2.5	0,115	0,039	2,969	0,003
doserate:doseamt-5	0,156	0,034	4,542	0,000

Todos os coeficientes, à exceção do coeficiente associado à variável taxa de administração (doserate), são significativos ao nível de 5%. Embora a variável doserate não tenha demonstrado significância por si só, é importante notar que suas interações com os níveis 2.5 e 5 da variável dose total (doseamt) foram estatisticamente significativas. Portanto, há evidência de interação entre essas duas variáveis, o que implica que a variável doserate não pode ser excluída do modelo.

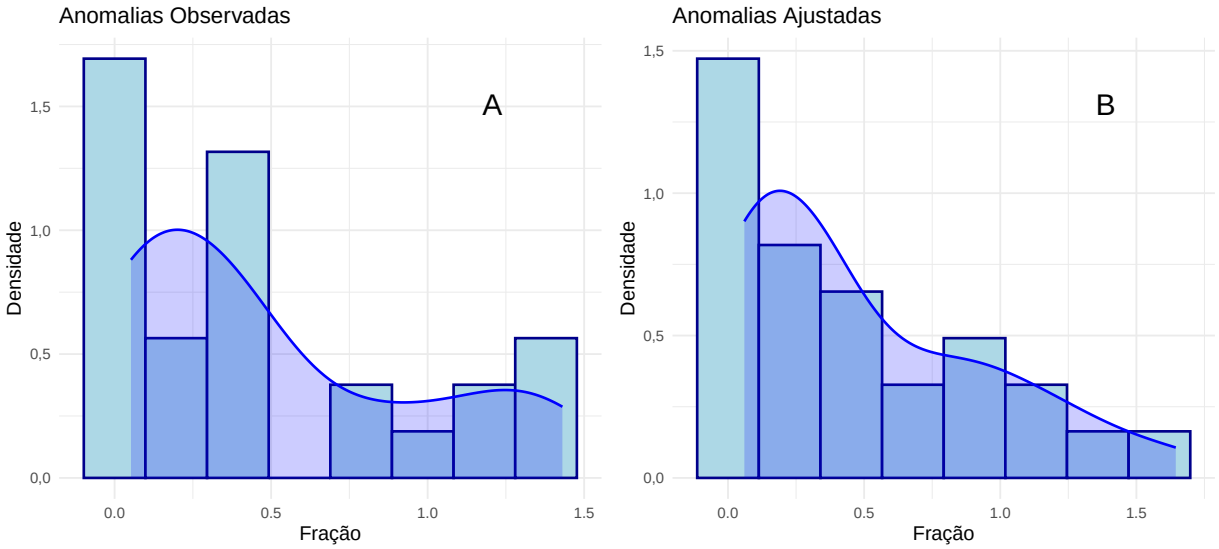
Quanto ao deviance dos resíduos deste modelo, juntamente com seus graus de liberdade (73,629 e 21, respectivamente), observamos que embora o valor do deviance dos resíduos seja substancialmente inferior ao valor nulo (4753,004), ele é aproximadamente três vezes maior do que o número de graus de liberdade, indicando um ajuste não totalmente satisfatório.

Tabela 3: Ajuste segundo o Modelo de Poisson aplicado, em comparação com o teste qui-quadrado.

	GL	Deviance	GL resídual	Resid. Deviance	Valor-p
NULL	NA	NA	26	4.753,004	NA
doserate	1	231,321	25	4.521,684	0
factor(doseamt)	2	4.426,890	23	94,794	0
doserate:factor(doseamt)	2	21,165	21	73,629	0

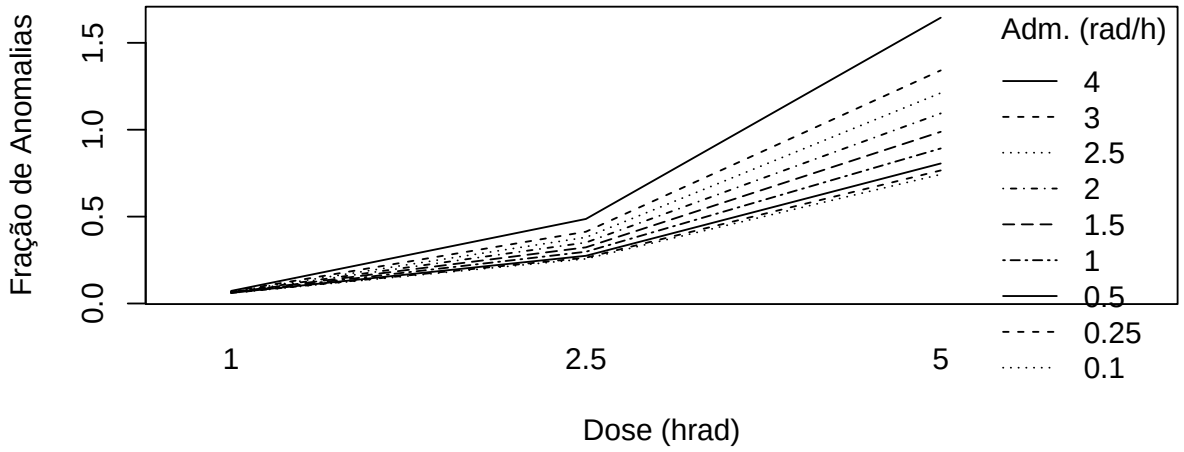
Observa-se novamente que todas as variáveis são necessárias para um bom ajuste do modelo. Entretanto, verificou-se que o variável variável dose total (doseamt) teve um grande impacto em diminuir o valor do *deviance*, indicando que é uma variável importantíssima para a melhor bondade de ajuste.

Figura 3: Histogramas das frações de anomalias.



A Figura 3 reforça a conclusão obtida na Tabela 2, pois o gráfico à direita na Figura 3B claramente ilustra as limitações do modelo Poisson para se adequar aos dados, uma vez que não apresenta um declínio visível.

Figura 4: Interação entre a fração de anomalias estimada e a dosimetria aplicada



Nota-se que a interação para a fração de anomalias ajustada capta a tendência vista na Figura 2 para a interação com a fração de anomalias observada: com o aumento da dosagem total e da taxa de administração da radiação há um aumento da fração de anomalias. o padrão de crescimento mais previsível que aquela mostrada na Figura 2, referente aos dados observados.

Tabela 2: Tabela 4: Medidas resumo dos resíduos

	Min	Q1	Median	Mean	Q3	Max	Std.Dev	CV	Skewness	Kurtosis
Resíduos	-5,86	-1,4	0,2	-0,12	1,28	2,36	1,98	-17,1	-1,13	1,01

A Tabela 4 mostra que há uma clara distinção entre a média e a mediana dos resíduos, indicando falta de simetria (visto também pelo *skewness* distante de zero) o que, por sua vez, implica que os resíduos não estão seguindo uma distribuição normal padrão.

Figura 5: Análise do resíduo componente do desvio pelos valores ajustados.

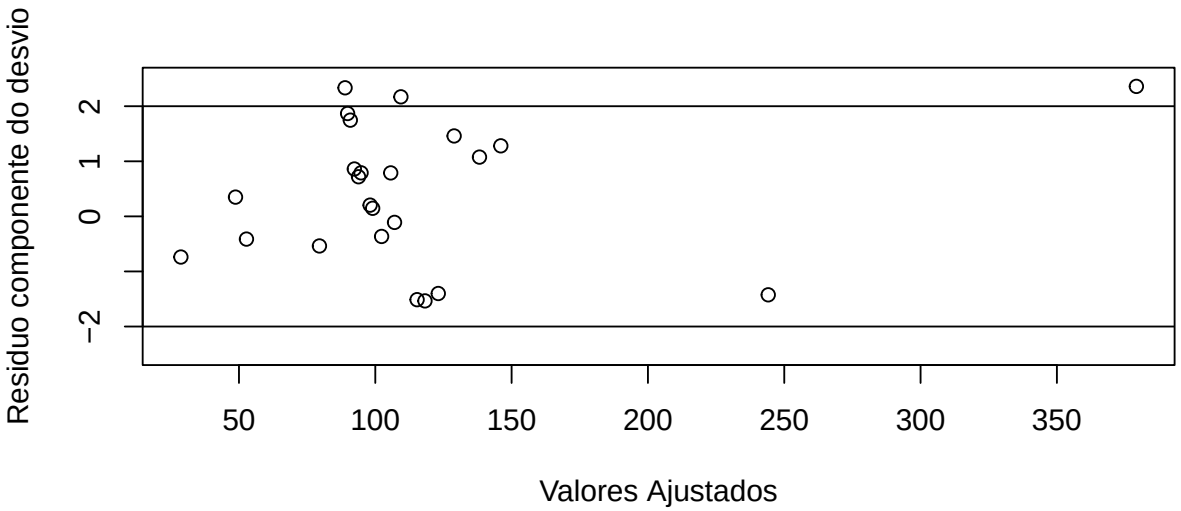


Figura 6: Análise do resíduo componente do desvio pelos valores observados.

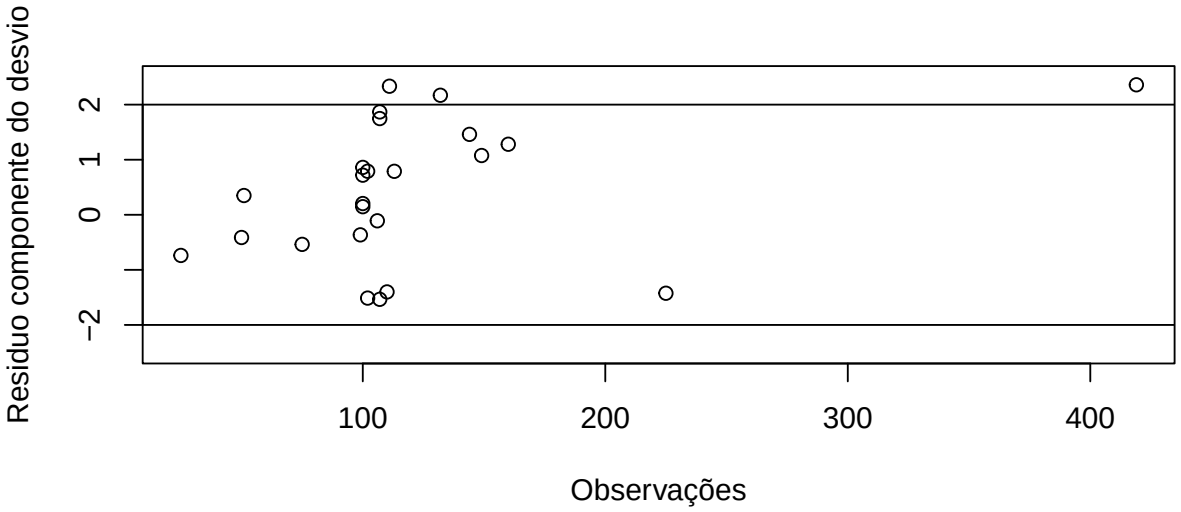
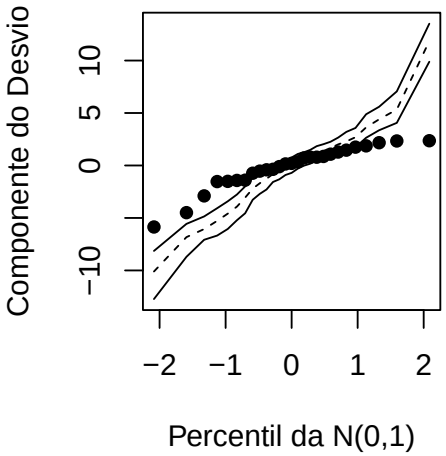


Figura 7: Ajuste do modelo sob avaliação em relação ao modelo poisson.



O gráfico da Figura 5 indica o mesmo que o gráfico da Figura 6: que devido a quantidade de observações fora dos limites de -2 a 2, os resíduos do modelo Poisson podem estar fugindo da normalidade. Essa afirmação é reforçada no gráfico da Figura 7, pois os quantis do componente do desvio não está seguindo os quantis da normal padrão, tendo diversos pontos fora do envolope. Portanto, é possível concluir que o modelo de Poisson ajustado pela função de ligação canônica não foi capaz de representar a variabilidade dos dados observados.

Função de ligação alternativa

Os dados foram submetidos novamente ao ajuste do modelo linear generalizado com distribuição Poisson. Nesse contexto, a função de ligação utilizada foi $\eta = \sqrt(\mu)$. Os coeficientes obtidos a partir desse ajuste estão disponíveis na Tabela 5.

Tabela 5: Ajuste segundo o Modelo de Poisson com função de ligação $\eta = \sqrt(\mu)$.

	Coeficientes	EP	Estatística z	Valor-p
Intercepto	2,643	0,274	9,630	0,000
doserate	-0,096	0,132	-0,725	0,468
doseamt-2.5	0,193	0,388	0,497	0,619
doseamt-5	2,529	0,388	6,516	0,000
doserate:doseamt-2.5	0,792	0,187	4,238	0,000
doserate:doseamt-5	1,750	0,187	9,363	0,000

Novamente, a variável taxa de administração (doserate) e o nível 2.5 da variável dose total não foram significativas ao nível de 5%. Desse modo. Embora a variável taxa de administração (doserate) isoladamente não ser significativa, verifica-se que sua interação com a variável dose total (doseamt) é significativa, já que foi significativa no nível de 5% para o níveis 2.5 e 5 da variável dose total, sendo assim é uma variável imprescindível no modelo.

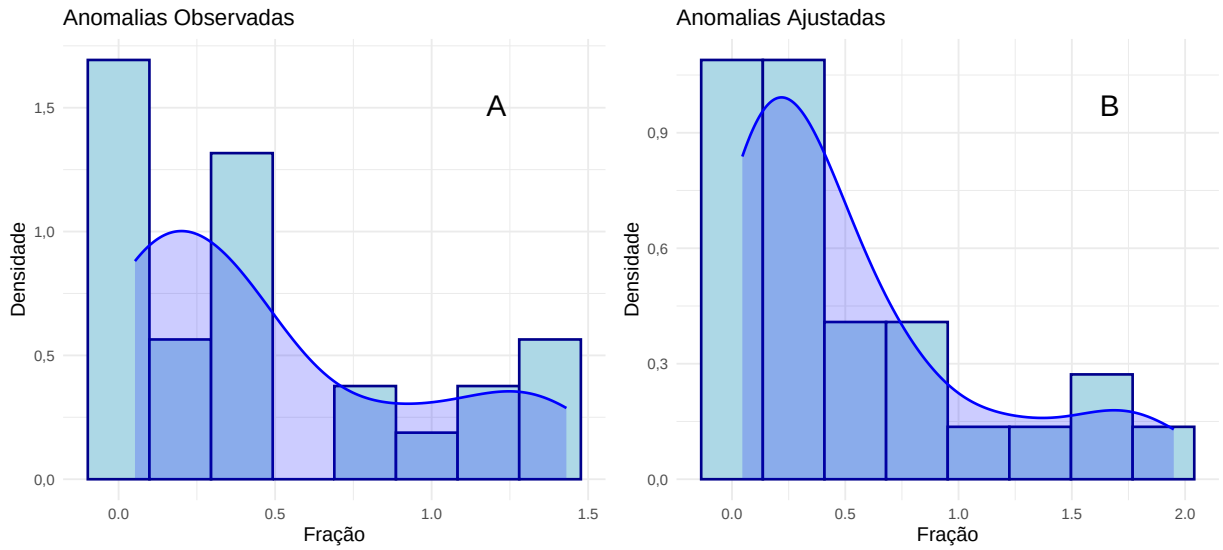
Chama a atenção nessa análise, que o valor da *deviance* dos resíduos, embora bem abaixo do valor nulo é muito superior ao valor dos graus de liberdade, indicando que o ajuste ainda apresenta certa imprecisão.

Tabela 6: Ajuste segundo o Modelo de Poisson aplicado, em comparação com o teste qui-quadrado.

	GL	Deviance	GL resídual	Resid. Deviance	Valor-p
NULL	NA	NA	26	1.103,921	NA
doserate	1	129,249	25	974,671	0
factor(doseamt)	2	574,428	23	400,244	0
doserate:factor(doseamt)	2	84,117	21	316,126	0

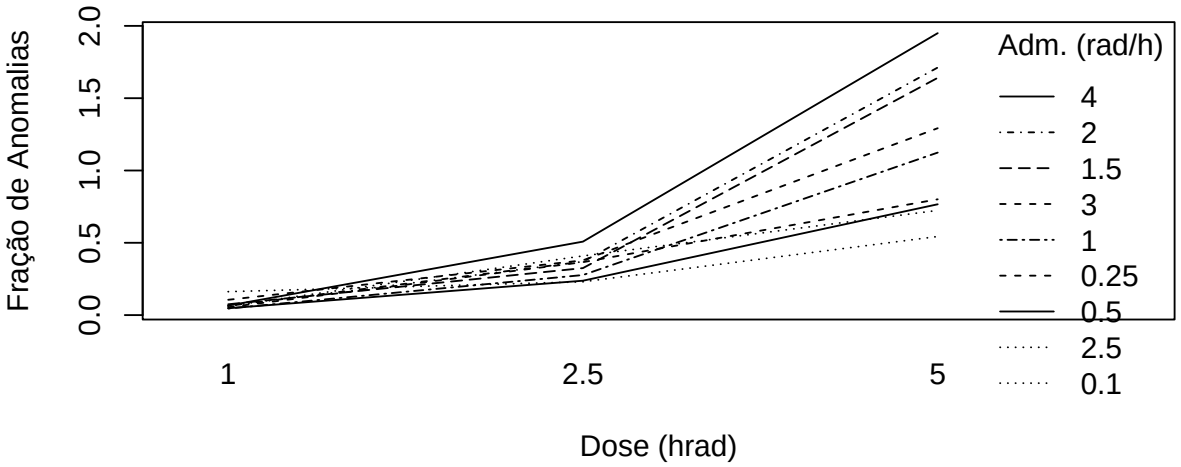
Observa-se novamente que todas as variáveis são necessárias para um bom ajuste do modelo. Entretanto, neste caso, a variável dose total teve menos impacto no *deviance* neste modelo Poisson com função de ligação raiz quadrada que no modelo Poisson com função de ligação canônica.

Figura 8: Histogramas das frações de anomalias.



A Figura 8 confirma a análise da Tabela 5, uma vez que o lado esquerdo da Figura 7B evidencia a dificuldade de ajuste pelo modelo Poisson, já que nesses dados não há um decaimento relevante observado.

Figura 9: Interação entre a fração de anomalias e a dosimetria aplicada ajustadas.



É possível observar assim como no primerio modelo, a fração de anomalias estimadas indicou tendência de crescimento a partir do momento em há um crescimento da taxa de dose administrada e da quantidade de dose total aplciada.

Tabela 3: Tabela 7: Medidas resumo dos resíduos

	Min	Q1	Median	Mean	Q3	Max	Std.Dev	CV	Skewness	Kurtosis
Resíduos	-8,19	-2,79	-0,17	-0,28	1,6	12,47	3,96	-14,17	0,87	2,13

Podemos notar que na Tabela 7, apesar de que a diferença entre a média e mediana seja menor para esse modelo do que o modelo anterior, ele ainda apresenta um valor de *skewness* relativamente alto, o que indica também assimetria e fuga da normalidade.

A Tabela 7 mostra uma distribuição dos resíduos bastante aproximada de uma distribuição normal.

A Tabela 4 mostra que há uma clara distinção entre a média e a mediana dos resíduos, indicando falta de simetria o que, por sua vez, implica que os resíduos não estão seguindo uma distribuição normal padrão.

Figura 10: Análise do resíduo componente do desvio pelos valores ajustados.

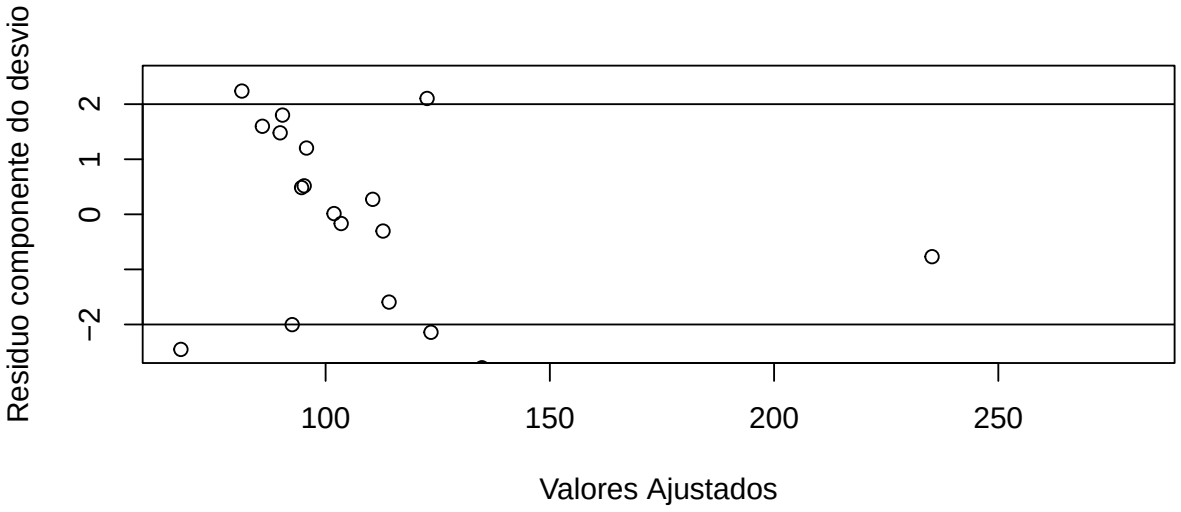


Figura 11: Análise do resíduo componente do desvio pelos valores observados.

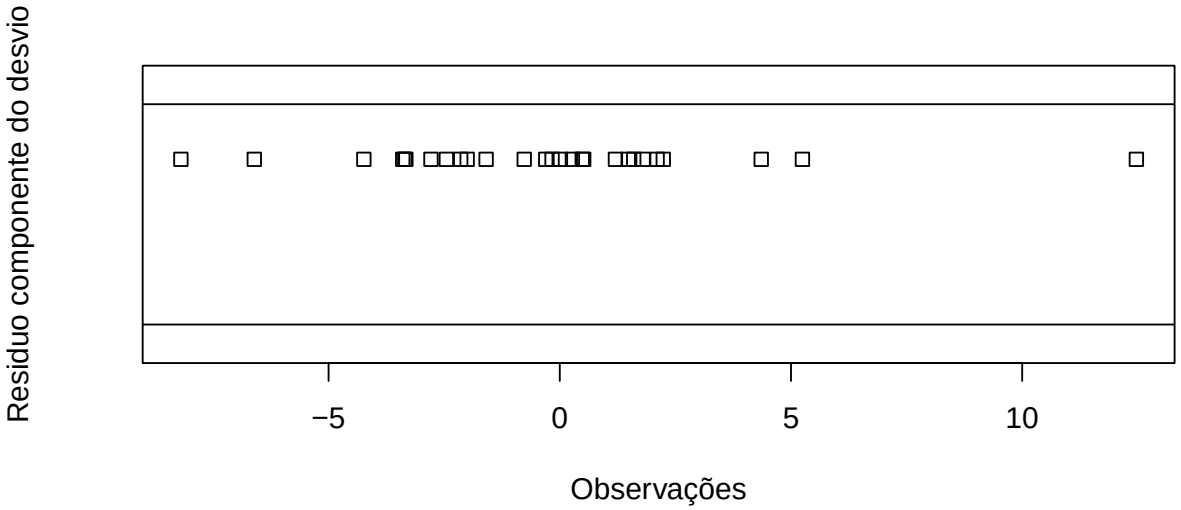
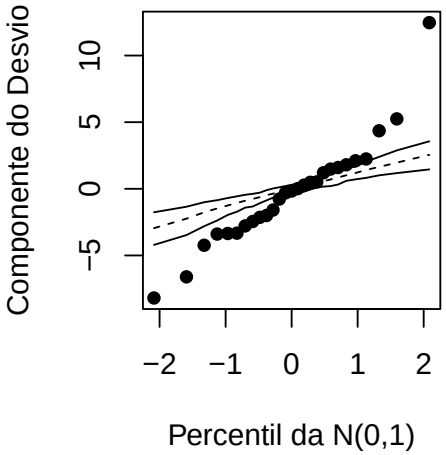


Figura 12: Ajuste do modelo sob avaliação em relação ao modelo poisson.



Das figuras acima verifica-se que o modelo de Poisson ajustado pela função de ligação alternativa também não foi capaz de representar a variabilidade dos dados observados.

Nos gráficos da Figura 10 e 11, é possível ver que há diversos pontos fora dos limites de -2 a 2, desse modo, os resíduos do modelo Poisson podem estar fugindo da normalidade. No gráfico da Figura 12, é notável que os quantis dos resíduos não estão seguindo os quantis da normal padrão, tendo a maior parte dos pontos fugindo do envelope. Portato, conclui-se que

o modelo Poissson ajustado pela função de ligação raiz quadrada não foi capaz de representar a variabilidade dos dados observados

indica o mesmo que o gráfico da Figura 6: que devido a quantidade de observações fora dos limites de -2 a 2, os resíduos do modelo Poisson podem estar fugindo da normalidade. Essa afirmação é reforçada no gráfico da Figura 7, pois os quantis do componente do desvio não está seguindo os quantis da normal padrão, tendo diversos pontos fora do envolope. Portanto, é possível concluir que o modelo de Poisson ajustado pela função de ligação canônica não foi capaz de representar a variabilidade dos dados observados.

Conclusões

Verificou-se que a proposta de identificar o processo de contagens de cromossomos com anomalias após o tratamento de radiação com o modelo de Poisson, não obteve um bom ajuste nem com a utilização da função de ligação canônica (logarítimica) e nem com a função de ligação alternativa (raiz-quatrática). Desse modo, o modelo Poisson não foi capaz de explicar a variabilidade da variável quantidade de cromossomos anômalos. Isso pode ter se dado em razão da grande variabilidade, sendo muito maior que a média, violando a necessidade que o modelo Poisson tem de que a sua média seja igual a variância.