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.

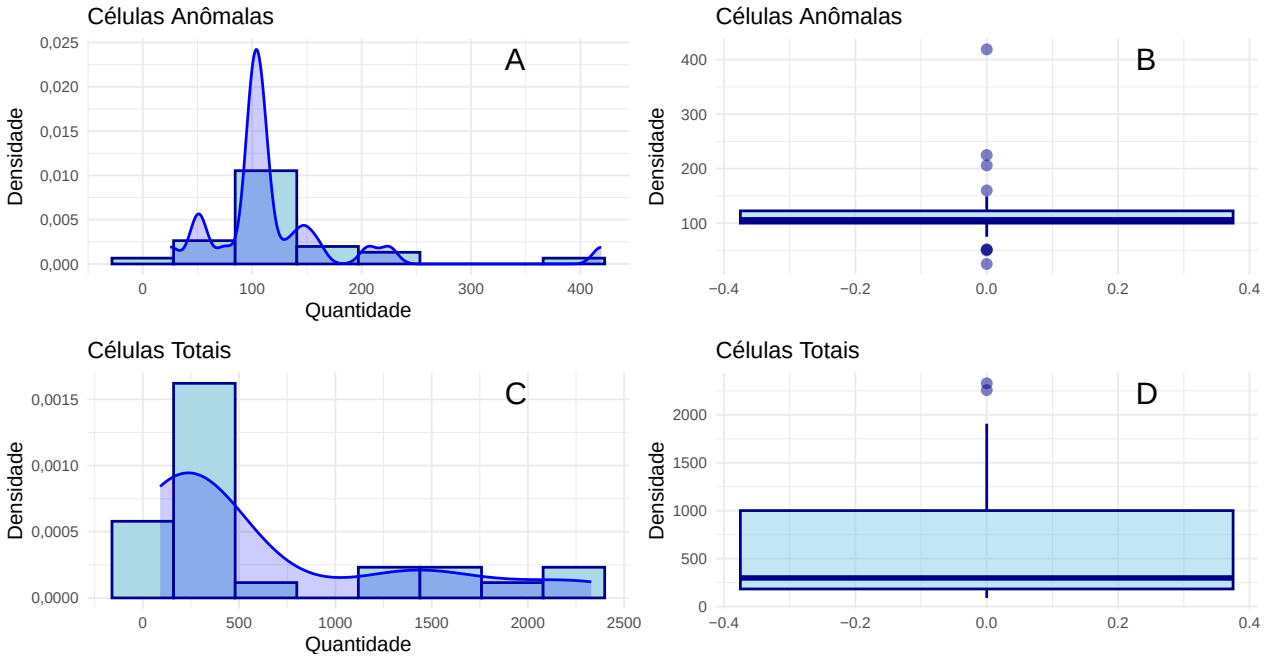
Tabela 1: Medidas resumo dos dados

	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, já que a mediana e a média foram bastante distintas. O coeficiente de variação da quantidade total de células é alto, desse modo, há um 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.

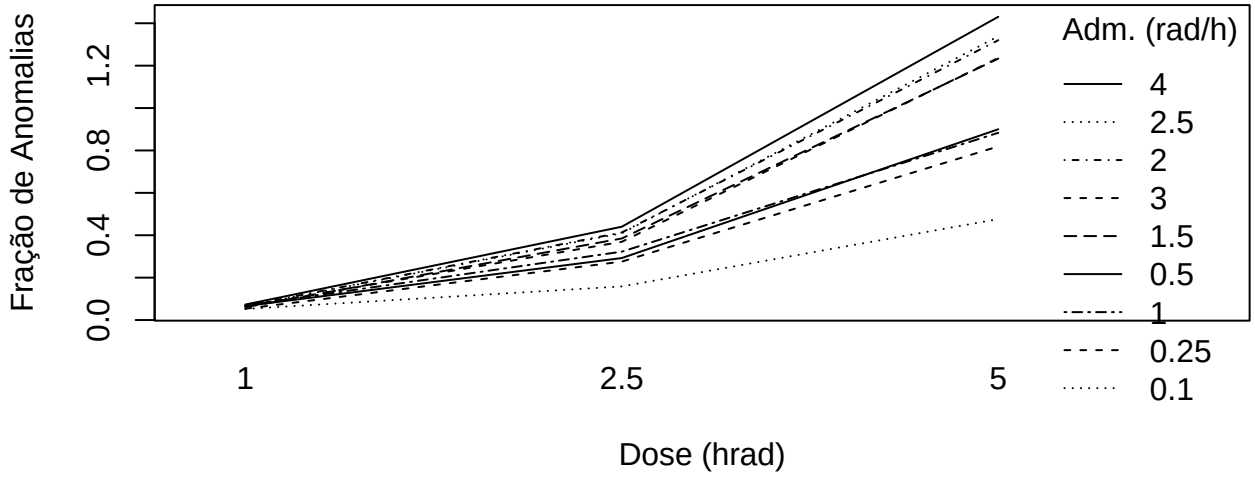
Figura 1: Histogramas e Boxplots das variáveis em análise.



Os gráficos confirmam as análises de simetria e variabilidade feitas anterioremente, acrescentando que em ambas as distribuições pode se verificar outliers, que por sua vez são mais

frequentes na variável dos cromossomos anômalos. Tanto no gráfico do histograma como no gráfico do box-plot podemos verificar que há uma assimetria à direita na distribuição das células totais, apresentando uma variabilidade maior acima da mediana.

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



A figura 2 indica que o aumento tanto da dosagem total quanto da taxa de administração da radiação tem efeito positivo no aumento da fração de anomalias observado.

Ajuste inferencial

Função de ligação canônica

Utilizando a função de ligação canônica para a qual $\eta = log(\mu)$ e procedendo um ajuste dos dados segundo um modelo poisson baseado em MLG temos o sequinte resutado descrito em uma tabela ANOVA.

Tabela 2: Ajuste segundo o Modelo de Poisson aplicado.

```
Call:
glm(formula = ca ~ doserate + factor(doseamt) + doserate:factor(doseamt) +
    offset(log(cells)), family = poisson, data = dicentric)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.81751    0.05632 -50.028  < 2e-16 ***
doserate                   0.04816    0.02809   1.715  0.08641 .
factor(doseamt)2.5         1.44259    0.08289  17.403  < 2e-16 ***
```

```
factor(doseamt)5             2.49955    0.07377  33.883  < 2e-16 ***
doserate:factor(doseamt)2.5  0.11545    0.03888   2.969  0.00299 **
doserate:factor(doseamt)5    0.15572    0.03428   4.542 5.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.004  on 26  degrees of freedom
Residual deviance:   73.629  on 21  degrees of freedom
AIC: 261.04

Number of Fisher Scoring iterations: 4
```

Todas os coeficientes, com exceção do coeficiente da variável taxa de administração (doserate) , foram significativas ao nível de 5%. Apesar da variável doserate não ter sido significativa, a interação dessa variável com os níveis 2.5 e 5 da variável dose total (doseamt) foram significativas. Desse modo, há interação entre essas duas variáveis e, portanto, a variável doserate não pode ser retirada do modelo. Chama a atenção nessa análise, que o valor da *deviance* dos resíduos, embora bem abaixo do valor nulo é cerca do triplo do valor dos graus de liberdade, indicando que o ajuste ainda apresenta certa imprecisão.

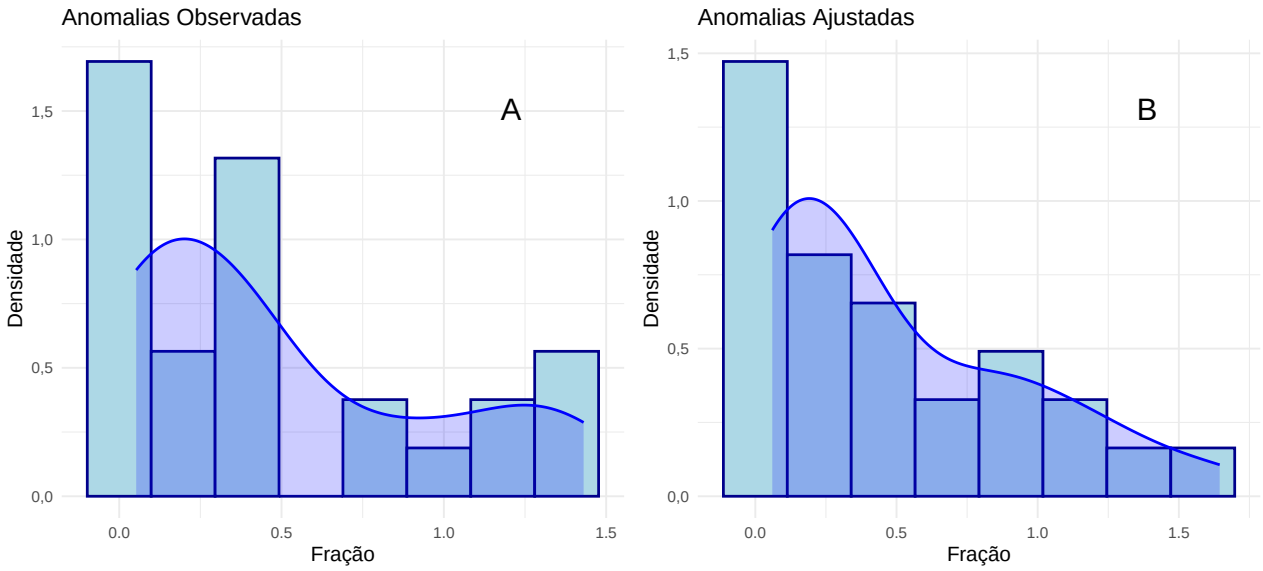
Tabela 3: Ajuste segundo o Modelo de Poisson aplicado, em comparação com o teste qui-quadrado.

```
Analysis of Deviance Table

Model: poisson, link: log

Response: ca

Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                        26     4753.0
doserate                  1    231.3        25     4521.7 < 2.2e-16 ***
factor(doseamt)           2   4426.9        23       94.8 < 2.2e-16 ***
doserate:factor(doseamt)  2     21.2        21       73.6 2.535e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observa-se novamente que todas as variáveis são necessárias para um bom ajuste do modelo.

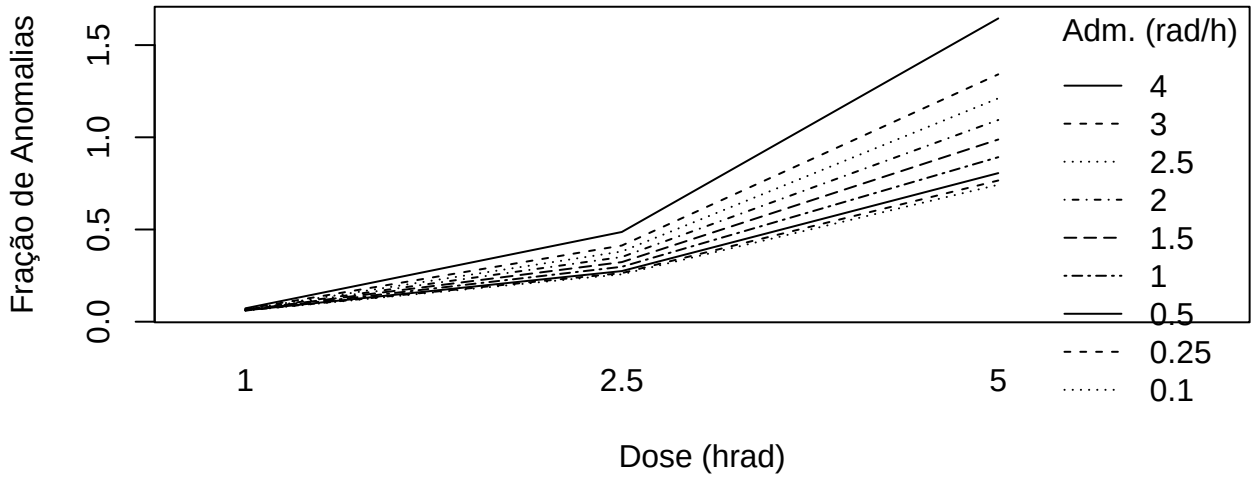
Entretanto, neste caso, verificou-se que o valor da *deviance* indica uma melhor bondade de ajuste.

Figura 3: Histogramas das frações de anomalias.



A Figura 3 confirma a análise da Tabela 2, uma vez que o lado direito da Figura 3B evidencia a dificuldade de ajuste pelo modelo Poisson, já que nesses dados não há um decaimento observado.

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



Nota-se que a interação ajustada é 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 uma distribuição dos resíduos aproximada de uma distribuição normal.

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

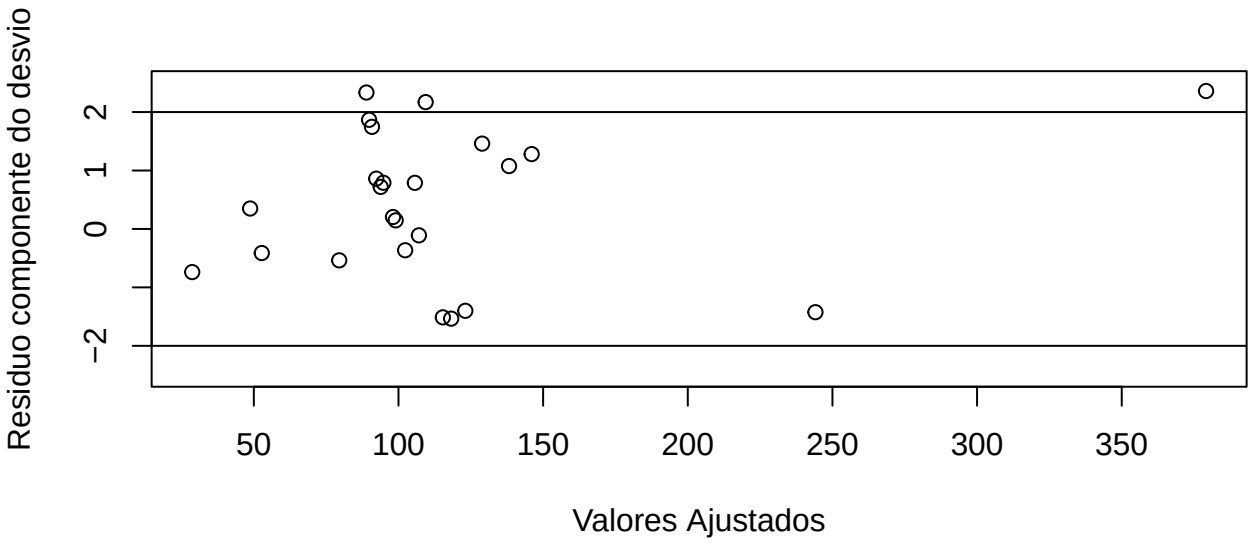


Figura 6: Análise do resíduo componente do desvio pelos valores observados.

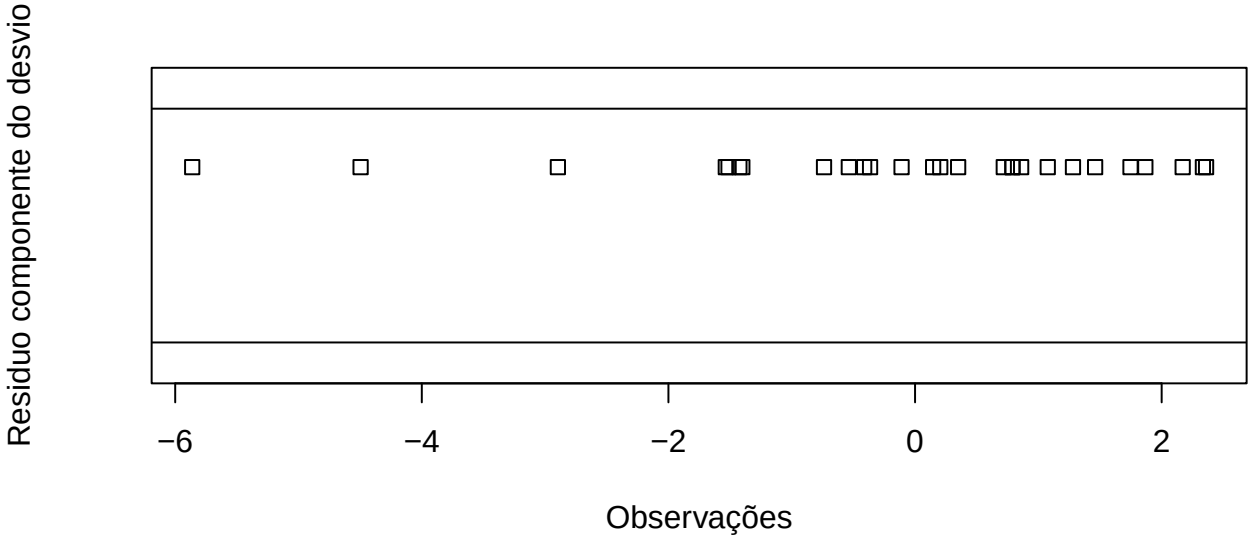
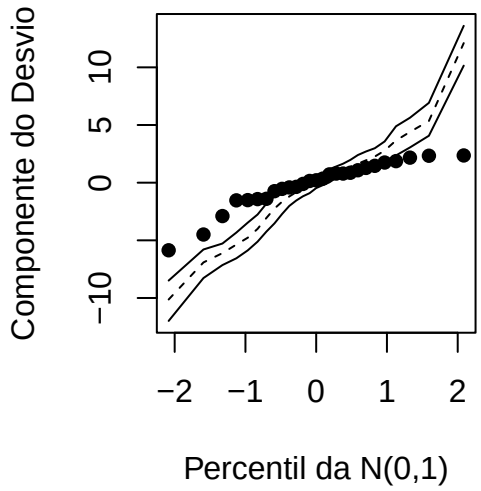


Figura 7: Ajuste do modelo sob avaliação em relação ao modelo poisson.



Das figuras acima verifica-se 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

Utilizando a função de ligação canônica para a qual $\eta = \sqrt{(\mu)}$ e procedendo um ajuste dos dados segundo um modelo poisson baseado em MLG temos o sequinte resutado descrito em uma tabela ANOVA.

Tabela 5: ANOVA do ajuste segundo o Modelo de Poisson aplicado.

```
Call:
glm(formula = ca ~ doserate + factor(doseamt) + doserate:factor(doseamt) +
    offset(log(cells)), family = poisson(link = "sqrt"), data = dicentric)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.64267    0.27442   9.630  < 2e-16 ***
doserate                    -0.09581    0.13213  -0.725    0.468
factor(doseamt)2.5           0.19297    0.38809   0.497    0.619
factor(doseamt)5             2.52888    0.38809   6.516 7.21e-11 ***
doserate:factor(doseamt)2.5  0.79187    0.18686   4.238 2.26e-05 ***
doserate:factor(doseamt)5    1.74958    0.18686   9.363  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 1103.92  on 26  degrees of freedom
Residual deviance:  316.13  on 21  degrees of freedom
AIC: 503.53

Number of Fisher Scoring iterations: 4
```

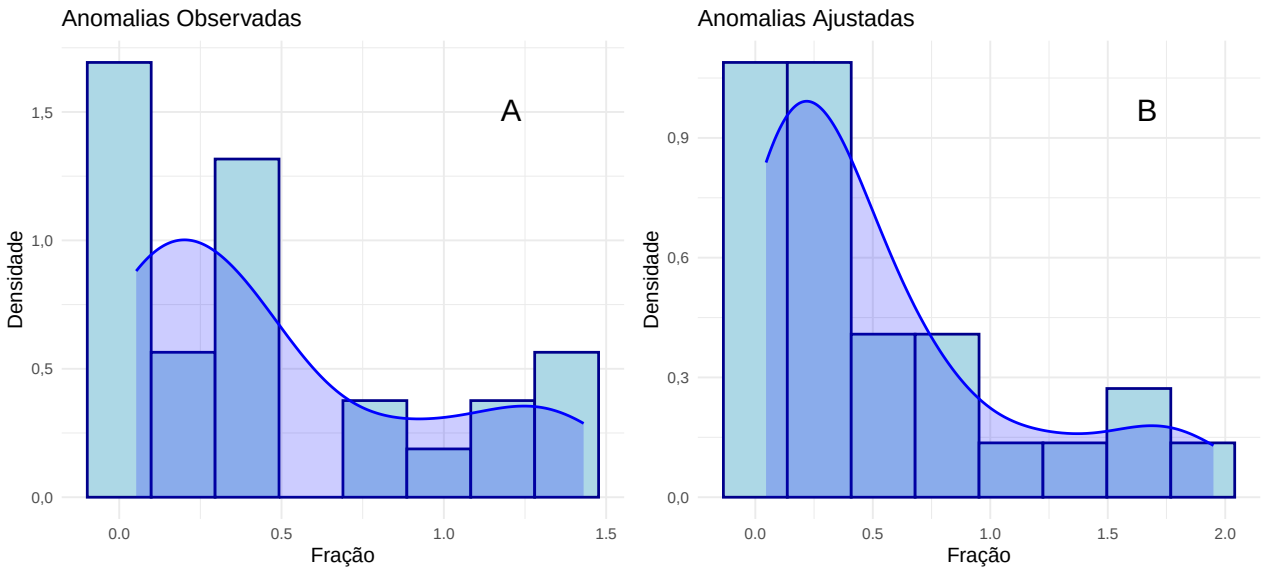
Nota-se que todos as variáveis do modelo foram significativas para o ajuste do modelo. Embora a variável taxa de administração (doserate) isoladamente apresentar uma significância abaixo dos níveis usuais, verifica-se que sua interação com a variável dose total (doseamt) é significativa no nível 5 hrad, sendo assim não é uma variável prescindível. 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: ANOVA do ajuste segundo o Modelo de Poisson aplicado, em comparação com o teste qui-quadrado.

```
Analysis of Deviance Table

Model: poisson, link: sqrt

Response: ca

Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                        26    1103.92
doserate                  1   129.25        25     974.67 < 2.2e-16 ***
factor(doseamt)           2   574.43        23     400.24 < 2.2e-16 ***
doserate:factor(doseamt)  2    84.12        21     316.13 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

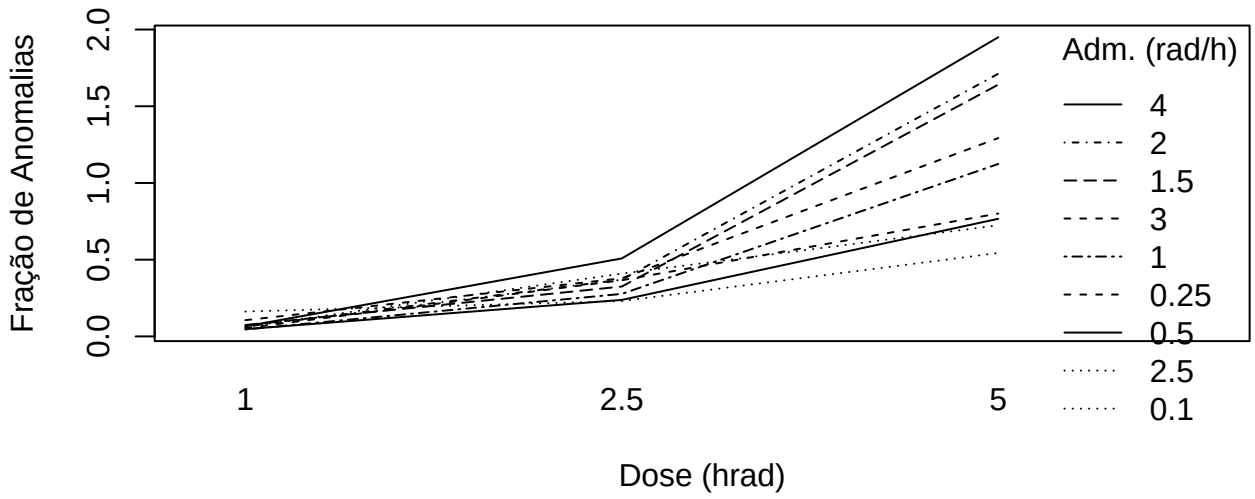
Observa-se novamente que todas as variáveis são necessárias para um bom ajuste do modelo. Entretanto, neste caso, verificou-se que o valor da *deviance* indica uma melhor bondade de ajuste.

Figura 8: Histogramas das frações de anomalias.



A figura 8 corfirma 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 observado.

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



Nota-se que a interação ajustada é mais previsível que aquela mostrada na Figura 2, referente aos dados observados.

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

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

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

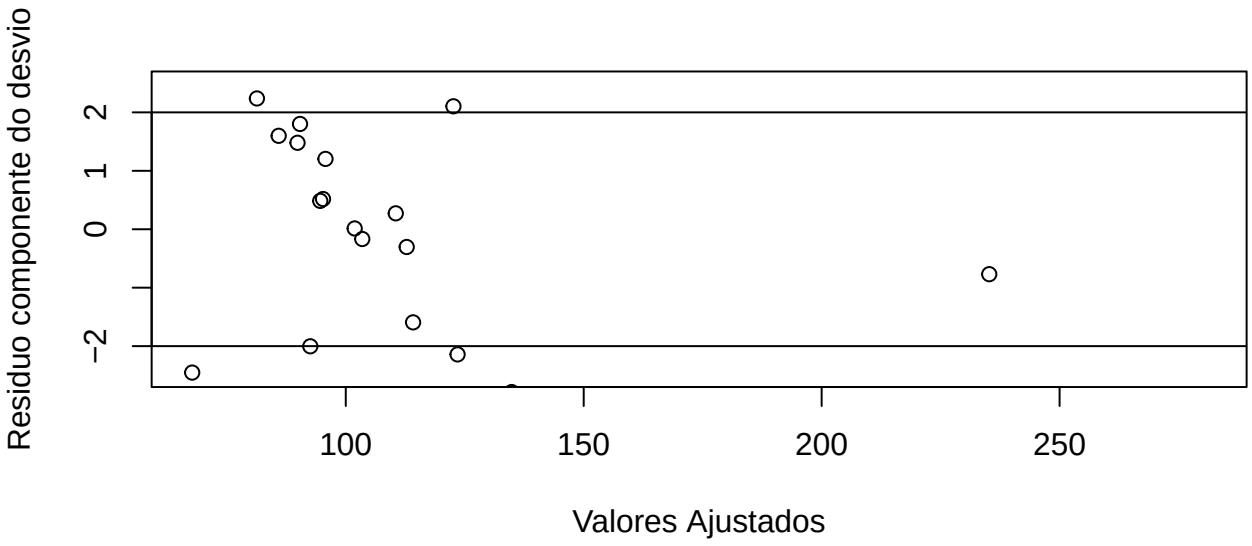


Figura 11: Análise do resíduo componente do desvio pelos valores observados.

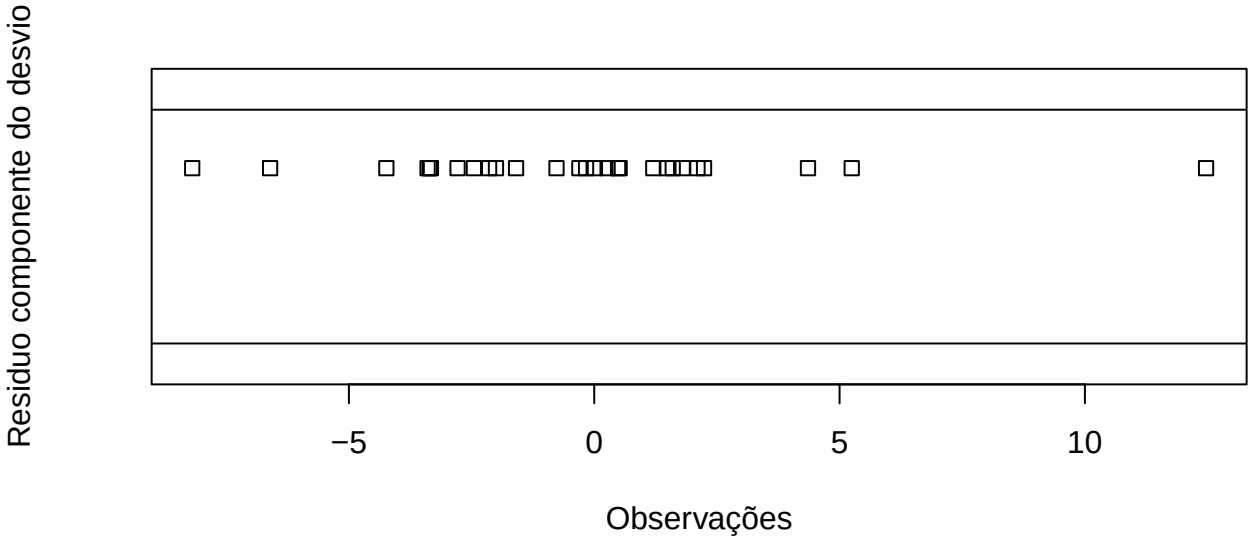
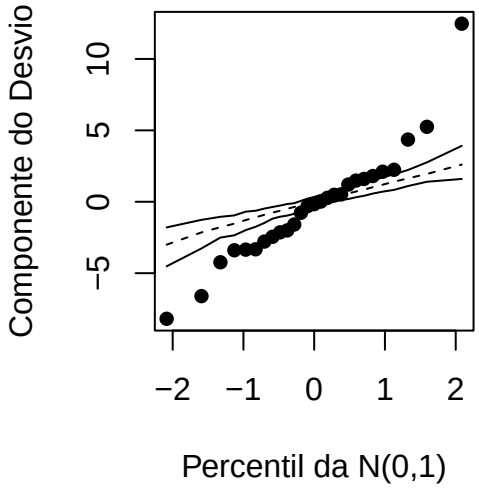


Figura 12: Ajuste do modelo sob avaliação em relação ao modelo poisson.



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.

Conclusões

Verificou-se que a proposta de identificar o processo de contagens de células com anomalias após o tratamento de radiação com o modelo de Poisson, não obteve um bom ajuste mesmo com a utilização da função de ligação canônica (logarítimica) ou de uma função de ligação alternativa (raiz-quatrática). Uma fonte desta dificuldade deste ajuste é a aleatoriedade da quantidade de células totais amostradas que introduz uma variabilidade que o modelo não é capaz de explicar. Um possível modelo com melhor ajuste seria baseado na distribuição Beta, que por sua vez ajustaria a fração de células comprometidas e não a contagem delas.