# MODELOS LINEARES GENERALIZADOS PARA DADOS DE CONTAGENS

## Introdução à modelagem de dados de contagem

Os modelos tradicionais de regressão linear, que assume uma distribuição gaussiana, não consideram a característica discreta, assimetria e heterocedástica com o rigor necessário, sendo por isso, ineficiente na maioria dos casos de contagens. Além disso, esses modelos fornecem erros padrão irrealistas ou previsões possivelmente enviesadas do número esperado de eventos (ZEVIANI *et al.*, 2014). Por outro lado, a regressão de Poisson é o modelo padrão para dados de contagem que seguem uma lei de formação da distribuição de probabilidade de Poisson que, conforme Bilder; Loughin (2015), caracterizam-se por:

1. contagens geradas por um mesmo processo e com mesmas vizinhanças;
2. período de observação constante para cada contagem;

Se (1) e (2) são verdade, então espera-se que uma variável aleatória de Poisson qualquer, representada por de um dado processo com sua vizinhança apresente resultados de contagem aleatórios em cada realização. Por exemplo, considerando uma amostra de água bruta de um açude que se deseja determinar a densidade de uma espécie de cianobactéria. Espera-se que o número de células por mL de duas alíquotas dessa amostra (mesma vizinhança) retiradas e analisadas ao microscópio da mesma forma e pelo mesmo analista (mesmo processo gerador) produzam resultados diferentes de densidade em cada análise.

Uma função distribuição de probabilidade ou função de massa de probabilidade (PMF - P*robability Mass Function*) especifica uma probabilidade para cada valor possível de uma variável aleatória . Dados de contagem são variáveis aleatórias discretas. Isso significa que dada as mesmas condições experimentais, *n* experimentos apresentarão valores limitados (e.g. apenas números Naturais, não qualquer valor Real) e aleatórios (qualquer valor Natural é possível), regidos pelas leis da probabilidade. Assim, considerando o exemplo anterior, seja uma va (variável aleatória), uma PMF determina a probabilidade de . Geralmente representa-se os valores de uma va por letras minúsculas (e.g. ), então a probabilidade de va Y assumir um dado y valor é representado por ).

Uma PMF pode ser compreendida como sendo uma equação matemática (i.e., modelo matemático) que relaciona valores de uma va com uma probabilidade de ocorrência. Assim, quando se analisa um conjunto de dados, para ser possível realizar alguma inferência, são necessárias suposições sobre esse mecanismo aleatório de geração de dados, ou seja, sobre a distribuição de probabilidade que supostamente originou os dados.

Para modelos de regressão com respostas contínuas, a distribuição normal desempenha o papel central. Já para dados categóricos ou discretos três distribuições são relevantes: binomial, multinomial e Poisson. Essa última, também é a mais relevante para dados de contagem de eventos (AGRESTI, 2012; BILDER; LOUGHIN, 2015).

A PMF é frequentemente o meio principal de definir uma distribuição de probabilidade discreta, e tais funções existem para variáveis ​​aleatórias escalares ou multivariadas cujo domínio é discreto, tais como dados de contagem. Uma função distribuição de probabilidade conjunta está para variáveis aleatórias discretas assim com uma função de densidade de probabilidade (PDF) está para variáveis ​​aleatórias contínuas.

## Distribuição de Poisson

Essa distribuição pode ser utilizada para contagem de eventos que ocorrem randomicamente ao longo do tempo (nº de organismos por minuto) ou espaço (nº células por mL ou por área, por exemplo). Esses eventos devem ser independentes, isto é, cada observação em um dado período ou espaço não interfere em outra.

Uma característica chave da distribuição de Poisson é que sua variância é igual a sua média. Sendo assim, dada uma amostra que segue a distribuição de Poisson, quando maior for sua média mais alta será a variação de sua contagem. Portanto, se a média do número de células por mL de um dado organismo for 1000 espera-se maior variabilidade em suas contagens do que de outro com média é igual a 100.

Observa-se que o resultado p(yi) de cada realização i de um experimento de uma va de Poisson é dependente um único valor específico. Esse valor é denominado de parâmetro da distribuição de Poisson e é representado por ou por , como alguns autores preferem. Assim, considerando um dado parâmetro , todos os valores p(yi) possíveis definem uma distribuição de probabilidade de Poisson. Isso pode ser representado por ), ou de forma abreviada, A coleção de todas as distribuições de probabilidade dos diferentes parâmetros possíveis é conhecida como sendo uma família de distribuições de probabilidade. Matematicamente, a PMF da distribuição de probabilidade de Poisson que depende apenas de da média é definida como:

Em que esperança de Y, isto é, E(Y) = var(Y) = (ou ), ou seja, a média é igual a variância. Algumas características:

* Unimodal, como valor de moda igual a parte inteira de
* A assimetria (*skewness*) descrita por
* A distribuição aproxima-se da normalidade quando cresce;
* Existe uma aproximação entre a distribuição binomial e Poisson quando o parâmetro é muito grande e π muito pequeno. Nesses casos o parâmetro de Poisson aproxima-se de (AGRESTI, 2012; BILDER; LOUGHIN, 2015; FRIENDLY, 2016).

No entanto, nesse manuscrito a variância dos dados de contagem excederam às previstas pelas distribuições de Poisson ou binomial, conforme visto na Tabela 1. Esse fenômeno é denominado de superdispersão, ocasionado por fatores não medidos e que impactam no valor médio da contagem, tais como tempo de análise, fatores morfológicos, fisiológicos, erros analíticos, erros na identificação dos organismos, dentre outros.

Tabela 1 - Média de variância para cada filo. ( A ausência foi considerada como 0)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Bacillariophyta | Charophyta | Chlorophyta | Cyanobacteria | Euglenozoa | Ochrophyta |
| Média | 13948.60 | 6853.49 | 16080.79 | 2282179.75 | 710.55 | 35.36 |
| Variância | 1.33 109 | 1.97 108 | 1.42 109 | 1.73 1013 | 3,46 106 | 1.87 104 |

Nesses casos, uma distribuição binomial negativa ou uma extensão da distribuição Poisson, com um parâmetro extra, chamada de quasi-Poisson podem fornecer melhor modelagem. Nessa última, assume-se que a variância é uma função linear da média e as estimativas são feitas por meio da inferência de quasi-verossimilhança.(AGRESTI, 2012).

## Modelos lineares generalizados (GLM)

Trata-se de uma classe de modelos que estendem os clássicos gaussianos, tais como os lineares, para os casos em que as respostas não apresentam distribuição normal ou se deseja modelar em função do valor esperado, ou seja, a média. Nessa classe de modelos encontram-se a regressão logística e a Poisson com suas derivações, dentre outros.

De acordo com Agresti (2012), três partes principais compões um GLM, a saber:

* Componente randômico: identificado pela variável resposta Y e sua respectiva distribuição de probabilidade. Para dados de contagem, o componente randômico Y pode ajusta-se a distribuição de Poisson ou Binomial Negativa, por exemplo.
* Componente sistemática: diz respeito ao conjunto de variáveis explicativas ou independentes. Essa componente é conhecida com função preditora linear.
* Função de ligação: função relaciona o componente randômico e o componente sistemático.

As etapas de modelagem usando uma GLM são muito semelhantes às realizadas em qualquer outro processo de modelagem de regressão. Essas etapas incluem:

1. Especificação do modelo: nessa etapa algumas perguntas a serem respondidas são:

* Que tipo de va é Y? discreta ou contínua?
* Qual a distribuição de Y?
* O que será modelado? Uma probabilidade? Um valor médio de contagem? Uma proporção?
* Quais as relações com as variáveis as variáveis ​​explicativas?

1. Selecionar as variáveis ​​explicativas: aqui é importante saber:

* Quais os melhores preditores lineares?
* Qual o tipo de cada preditor? Categórico? Quantitativo?
* Existe dados faltantes ou outliers?
* Existe interação entre as variáveis?

1. Estimativa dos parâmetros do modelo: utilização de mínimos quadrados ou verossimilhança?
2. Avaliação do ajuste do modelo
3. Realizar inferência sobre os parâmetros e outras quantidades de interesse

### Componente randômico dos GLMs

Consiste em uma variável de resposta Y com observações independentes (y1,. . ., yN) de uma distribuição da família exponencial. Esta família tem função de densidade de probabilidade (se va contínua) ou função massa da forma (se va discreta):

Em que e dependem da distribuição de probabilidade de Y e é chamado de parâmetro natural. Na Tabela 2 são exemplificados os casos de Poisson e Bernoulli com , e n = 1.

Tabela 2 - Exemplos da função de densidade de probabilidade

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Distribuição de Y |  |  |  |  |  |
| Poisson |  |  |  |  |  |
| Bernoulli |  |  |  |  |  |

### Componente sistemático de um GLM

Trata-se da combinação linear das variáveis exploratórias. Esse componente relaciona um vetor por meio dessa combinação, conforme a relação abaixo:

Para , para todo i, representa o intercepto do modelo.

### Função de ligação

Responsável por conectar as componentes randômicas e sistemáticas, a função de ligação é monotônica, diferenciável e se , para as observações i = 1, 2, ...N, então o modelo deve relacionar a por meio de um função , ou seja, , conforme visto a seguir:

Ressalta-se:

* O link é chamado de ligação identidade ou link identidade;
* Chama-se de link canônico ou função de ligação canônica aqueles que transforma a média no parâmetro natural Na coluna da Tabela 2 encontram-se os exemplos de links canônicos para a distribuição de Poisson e binomial

Portanto, para a regressão binomial e Poisson as funções de ligação canônicas são o *logito* e o *log*, respectivamente, por isso o modelo de regressão de Poisson é denominado de loglinear às vezes. Existem outros links para binomial além canônico, tais como *probito, Cauchit*, cloglog. Já para Poisson os principais são raiz quadrada e identidade.

### Modelos de regressão de Poisson (\*base para M & M)

A Tabela 3 mostra os resultados de processos de contagem do total de células por mL para 7 filos encontrados em alíquotas dos Mesocosmos 1 a 6. A priori, considerando que a probabilidade de ocorrência de um dado valor de contagem em um determinado intervalo de tempo seja proporcional à sua duração e independente da ocorrência de outros eventos, e considerando também que tanto os processos geradores da contagens e suas vizinhanças são constante ao longo de todo o experimento, os resultados da Tabela 3 segue aproximadamente uma distribuição de probabilidade de Poisson (FARAWAY, 2016). De ante desses resultados, duas modelagens podem ser realizadas:

* **ABORDAGEM 1**: cada Mesocosmos é um sistema independente, com seu próprio processo, vizinhança e período de observação constate para cada contagem;
* **ABORDAGEM 2**: os Mesocosmos 1 a 3 e de 4 a 6 são considerados realizações de dois processos distinto. Um processo é caracterizado pela presença de reatores de fotocatálise e o outro não. Cada processo tem sua vizinhança semelhante e período de observação constate para cada contagem;

Independente da abordagem, tem-se n = 72 observações da va (número de organismos de um determinado filo por mL). Os filos identificados foram *Bacillariophyta, Charophyta, Chlorophyta, Cryptophyta, Cyanobacteria, Euglenozoa, Ochrophyta*.

Tabela 3 - Resultados dos experimentos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tempo**  **(h)** | **Percentual** | **Mesocosmos** | **Bacillariophyta**  **(Cel mL-1)** | **Charophyta**  **(Cel mL-1)** | **Chlorophyta**  **(Cel mL-1)** | **Cryptophyta**  **(Cel mL-1)** | **Cyanobacteria**  **(Cel mL-1)** | **Euglenozoa**  **(Cel mL-1)** | **Ochrophyta**  **(Cel mL-1)** |
| 0 | 0 | 1 | 1215 | 83 | 3502 | 125 | 515826 | 166 | 0 |
| 2 | 690 | 101 | 4184 | 913 | 495612 | 152 | 0 |
| 3 | 1359 | 0 | 2299 | 177 | 450710 | 44 | 0 |
| 0 | 4 | 1108 | 0 | 2909 | 349 | 425141 | 0 | 0 |
| 5 | 1228 | 44 | 3823 | 155 | 378631 | 44 | 0 |
| 6 | 813 | 0 | 2040 | 478 | 403022 | 87 | 0 |
| 72 | 100 | 1 | 3976 | 2609 | 3234 | 626 | 608674 | 52 | 0 |
| 2 | 221868 | 85334 | 221718 | 20211 | 25334760 | 11228 | 0 |
| 3 | 1575 | 674 | 853 | 112 | 186648 | 62 | 0 |
| 0 | 4 | 379 | 112 | 3703 | 393 | 389496 | 0 | 0 |
| 5 | 13912 | 3619 | 26035 | 1206 | 3230362 | 1508 | 0 |
| 6 | 8787 | 7065 | 13962 | 1544 | 2349226 | 712 | 0 |
| 168 | 100 | 1 | 5655 | 3162 | 5130 | 846 | 899905 | 223 | 0 |
| 2 | 4977 | 3505 | 5144 | 803 | 656123 | 137 | 0 |
| 3 | 6079 | 1952 | 3843 | 320 | 630314 | 331 | 0 |
| 0 | 4 | 1327 | 985 | 2112 | 343 | 438687 | 86 | 0 |
| 5 | 7088 | 4856 | 16618 | 1005 | 2247892 | 670 | 167 |
| 6 | 8226 | 6969 | 7341 | 1628 | 1780519 | 771 | 0 |
| 336 | <70 | 1 | 14122 | 1765 | 12178 | 2900 | 1706829 | 63 | 0 |
| 2 | 17059 | 4452 | 13021 | 1863 | 1441670 | 1043 | 801 |
| 3 | 18154 | 5265 | 14864 | 1645 | 1227544 | 1382 | 132 |
| 0 | 4 | 5653 | 891 | 1933 | 260 | 595864 | 192 | 0 |
| 5 | 46168 | 11463 | 79192 | 6659 | 7688710 | 1747 | 0 |
| 6 | 13637 | 3808 | 27799 | 1063 | 4350334 | 974 | 0 |
| 504 | <70 | 1 | 10257 | 5148 | 9971 | 1113 | 1751455 | 487 | 70 |
| 2 | 11763 | 6258 | 9346 | 1145 | 1691342 | 229 | 0 |
| 3 | 11861 | 7273 | 7735 | 1265 | 1921800 | 443 | 0 |
| 0 | 4 | 4895 | 8900 | 4259 | 668 | 2005644 | 223 | 0 |
| 5 | 3126 | 6455 | 7034 | 145 | 1567402 | 0 | 0 |
| 6 | 7289 | 4692 | 6348 | 276 | 2680160 | 0 | 0 |
| 720 | <50 | 1 | 5584 | 3799 | 6001 | 475 | 1224285 | 339 | 0 |
| 2 | 8084 | 5994 | 3731 | 1447 | 1433545 | 207 | 103 |
| 3 | 10815 | 12539 | 11919 | 2409 | 2187194 | 425 | 0 |
| 0 | 4 | 8159 | 15647 | 11686 | 1649 | 2333839 | 210 | 0 |
| 5 | 4239 | 8832 | 7170 | 1766 | 2477514 | 258 | 0 |
| 6 | 11020 | 12476 | 16273 | 3255 | 2451791 | 1085 | 0 |

### Modelos binomiais negativos

Nos casos de superdispersão em que E(Y) < var(Y), isto é, a variância é maior do que a média e, portanto, excede a acomodada pelo distribuição de Poisson (Tabela 1), modelos binomiais negativos podem fornecer resultados mais realistas. Para a distribuição binomial negativa a equação abaixo apresenta a PMF, em que os dois parâmetros , que representa a média, e , que representa a extensão da superdispersão e geralmente é desconhecido, se relacionam:

, y = 0, 1, 2, ...,

Em que ou é chamada de função gama, definida:

Para essa distribuição tem-se:

* E(Y) = ;
* Var(Y) = ;

Denomina-se parâmetro de dispersão o índice . Quando o parâmetro de dispersão tende a zero ( a binomial negativa tende a distribuição de Poisson. Dessa forma, a distribuição de Poisson pode ser considerada como um caso particular da binomial negativa para

Ressalta-se que os valores dos parâmetros estimados pelos modelos binomial negativo e Poisson são iguais, porém os do erro padrão não. Assim, em relação a distribuição de Poisson, as estimativas dos erros padrões são maiores para a binomial negativa. Isso impacta no tamanho do intervalo de confiança, que são maiores para a binomial negativa, significando uma maior imprecisão, característica esperada quando existe superdispersão ().

### Estimativa quasi-verossimilhança para casos especiais de dados de contagem

Casos de superdispersão que a variância excede a média como os da Tabela 1 não são devidamente modelados por uma distribuição de Poisson. Uma das causas para isso é a heterogeneidade, além da já mencionada desconsideração de preditores. Isso sugere uma alternativa, uma relação de média-variância na forma:

Em que:

: é uma função de variância qualquer;

é uma constante, em que , para dados superdispersos e para casos subdispersos, o que significa uma maior flexibilidade dessa abordagem em relação à binomial negativa, que só acomoda superdispersão.

Para a estimativa de pode ser calculada:

Em que , é a soma dos resíduos padronizados ao quadrados, é o número de observações e é o número de parâmetros estimados do modelo.

Portanto, em linhas gerais, a abordagem de quase-verossimilhança para dados de contagem é simplesmente um ajuste do modelo de Poisson comum, em que as estimativas de parâmetros são as mesmas, mas as estimativas de erro padrão são multiplicadas por (AGRESTI, 2012).

### Offset: Influência do tempo de exposição

Na modelagem de dados de contagem geralmente os tempos de exposição a um dado fator são iguais. Todavia, quando isso não ocorre, o investigador deve considerar um termo de correção referente a essa variação da unidade de tempo. Assim, que para cada participante ou observação do estudo é possível considerar tempos diferentes. Esse termo é designado na literatura como “eventos por unidade de tempo”, “tempo de exposição” ou, simplesmente, “exposição”. Na regressão de Poisson e suas extensões isso é feito através de um *offset* alocado no modelo ajustado.

Considerando um modelo de Poisson com a função de ligação canônica log, a equação abaixo exemplifica um termo *offset:*

Em que é a forma matricial de e a “Exposição” representa as unidade de tempo variáveis para cada evento ou observação. Geralmente variável que mede tempo são boas candidatas a *offset*. Manipulando a equação acima, tem-se:

Portanto, cada função de ligação exige uma manipulação específica do termo de *offset.*

AGRESTI, A. **Categorical Data Analysis**. 2. ed. New Jersey: John Wley & Sons, 2012.

BILDER, C. R.; LOUGHIN, T. M. **Analysis of categorical data with R**. New York: CRC Press, 2015.

FARAWAY, J. J. **Extending the linear model with R generalized linear, mixed effects and nonparametric regression models**. New York: CRC Press, 2016.

FRIENDLY, M. **Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data**. New York: Chapman and Hall/CRC, 2016.

ZEVIANI, W. M.; RIBEIRO, P. J.; BONAT, W. H.; SHIMAKURA, S. E.; MUNIZ, J. A. The Gamma-count distribution in the analysis of experimental underdispersed data. **Journal of Applied Statistics**, v. 41, n. 12, p. 2616–2626, 2014. Disponível em: https://doi.org/10.1080/02664763.2014.922168