UNIVERSIDADE FEDERAL DO MARANHÃO
Fundação Instituída nos termos da Lei 5.152 de 21/10/1966 - São Luís - MA
Centro de Ciências Exatas e Tecnologia
Curso de Matemática – Bacharelado

Ygor Carvalho Penha

Aplicações do Modelo de Regressão Poisson

São Luís - MA
2023

Ygor Carvalho Penha

Aplicações do Modelo de Regressão Poisson

Monografia (Trabalho de Conclusão de Curso) apresentada à Coordenadoria dos cursos de Matemática, da Universidade Federal do Maranhão, como requisito parcial para obtenção do grau de Bacharel em Matemática.

Curso de Matemática – Bacharelado

Universidade Federal do Maranhão

Orientador: Prof. Dr. Josenildo de Souza Chaves

São Luís - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Carvalho Penha, Ygor.
Aplicações do Modelo de Regressão Poisson / Ygor Carvalho Penha. - 2023.
45 f.

Orientador(a): Josenildo de Souza Chaves.
Monografia (Graduação) - Curso de Matemática, Universidade Federal do Maranhão, São Luís - MA, 2023.

1. Dados de Contagem. 2. Estimadores de Máxima Verossimilhança. 3. Regressão Poisson. I. de Souza Chaves, Josenildo. II. Título.

Ygor Carvalho Penha

Aplicações do Modelo de Regressão Poisson

Monografia (Trabalho de Conclusão de Curso) apresentada à Coordenadoria dos cursos de Matemática, da Universidade Federal do Maranhão, como requisito parcial para obtenção do grau de Bacharel em Matemática.

Trabalho **APROVADO**. São Luís - MA, 14/12/2023

Prof. Dr. Josenildo de Souza Chaves
Orientador
DEMAT/UFMA

Prof. Dr. Marcos Antonio Ferreira Araújo
Primeiro Examinador
DEMAT/UFMA

Prof.ª Dr.ª Valeska Martins de Souza
Segunda Examinadora
DEMAT/UFMA

Ao Deus de Abraão, Isaque e Jacó.

Agradecimentos

Primeiramente, agradeço a Deus e Nossa Senhora, que me concederam força, sabedoria e proteção ao longo de todo este percurso. Sem sua orientação divina, nada disso seria possível.

À minha mãe, Maria José França Carvalho, e ao meu pai, José Ribamar Caldas Penha, expresso minha eterna gratidão. O amor, apoio e incentivo foram fundamentais. Vocês são minha inspiração.

À minha querida tia Vera. Obrigado pelo apoio e carinho.

À minha esposa, Adrianna Krissya, agradeço por seu amor, paciência e compreensão.

Aos meus amigos João Mario, Beneilson Neves, Willian Barros, Larissa Barros e Alex Moreira, agradeço pela amizade sincera, pelas palavras de incentivo e pelo apoio constante. Suas presenças em minha vida são verdadeiros tesouros.

Aos amigos que conheci durante nossa luta e evolução dentro da universidade, Davi Komura, Carla Beatriz, Rafael Vieira, Renata França, Irlan Maycon e Ronaldo Pinheiro, foi uma honra compartilhar esta jornada com vocês.

Por fim, não posso deixar de agradecer ao meu orientador, prof. Dr. Josenildo de Souza Chaves, e todos os professores por suas orientações cuidadosas, paciência e conhecimentos compartilhados. Suas orientações foram valiosas para o sucesso deste trabalho.

Agradeço a todos por fazerem parte da minha jornada acadêmica e por terem um papel significativo na realização desta monografia. Sou profundamente grato a cada um. Levo comigo as lições aprendidas e as memórias compartilhadas ao longo desta caminhada.

"não temas, porque eu sou contigo; não te assombres, porque eu sou o teu Deus; eu te esforço, e te ajudo, e te sustento com a destra da minha justiça."

Isaías 41:10

Resumo

Este trabalho apresenta aplicações da Regressão Poisson, em três conjuntos de dados de contagem. Apresenta ainda uma introdução à probabilidade, a construção da distribuição de Poisson e uma introdução aos estimadores de Máxima Verossimilhança. Utilizando o software R, realizamos uma análise de dados simulados e reais da literatura ilustrando a eficácia do modelo. O trabalho ressalta a importância da modelagem estatística na compreensão de fenômenos complexos da área de saúde pública e epidemiologia.

Palavras-chave: Regressão Poisson, Dados de Contagem, Estimadores de Máxima Verossimilhança.

Abstract

This work presents applications of Poisson Regression in three count data sets. It also provides an introduction to probability, the construction of the Poisson distribution, and an introduction to Maximum Likelihood estimators. Using R software, we conducted an analysis of simulated and real data from the literature, illustrating the effectiveness of the model. The paper highlights the importance of statistical modeling in understanding complex phenomena in the field of public health and epidemiology.

Keywords: Poisson Regression, Counting Data, Maximum Likelihood Estimators.

Lista de ilustrações

Sumário

Introdução

A estatística e a probabilidade desempenham um papel fundamental no entendimento e interpretação de sistemas físicos e biológicos. Por meio do método estatístico, é possível obter informações relevantes para tomadas de decisões em diversas áreas. A estatística descritiva, por meio de resumos numéricos e gráficos, permite uma visão ampla da distribuição dos dados, facilitando a identificação de padrões. Modelos matemáticos, são eficazes na descrição de uma vasta gama de fenômenos. Em medicina, por exemplo, a estatística desempenha um papel crucial na avaliação da ocorrência e progressão de doenças.

No caso determinístico, alguns modelos matemáticos relativamente simples parecem ser capazes de descrever uma classe bastante grande de fenômenos (MEYER, 1983). Por exemplo, podemos determinar a taxa média de ocorrência de células cancerígenas em um determinado órgão do corpo humano.

Por meio dos métodos de estimação podemos realizar estimativas dos parâmetros com base em uma amostra. Em geral, realizar inferências sobre a presença da doença em um paciente específico e escolher as melhores opções de tratamento.

A distribuição de probabilidade de Poisson é a principal candidata para modelar dados de contagem, quando o interesse uma variável aleatória X para representar o número de ocorrências de um determinado evento por unidade de medida, por exemplo, tempo, área e volume.

Esta distribuição foi introduzida por Siméon Denis Poisson em um livro que escreveu a respeito da aplicação da teoria da probabilidade a processos, julgamentos criminais e similares. O livro, publicado em 1837, *Recherches sur la probabilité de jugements en matière criminelle et en matière civile(Investigação sobre a probabilidade de vereditos em matérias criminal e civil)* (ROSS, 2010).

Este trabalho explora a regressão de Poisson e suas aplicações utilizando o software R (R Core Team, 2022). As aplicações incluem dados simulados que permitem controlar variáveis e cenários distintos para compreender o comportamento e a eficácia do modelo. Exploramos dois conjuntos de dados reais também discutidos por (DALGAARD, 2019). Além disso, apresentamos os fundamentos teóricos relativos ao modelo de regressão Poisson, abrangendo sua formulação matemática, pressupostos, procedimentos de inferência, vantagens e limitações.

Organização do Trabalho

Este trabalho está estruturado por uma Introdução e mais 4 capítulos, cada um dedicado a explorar aspectos fundamentais e aplicáveis da regressão Poisson.

O Capítulo 1 concentra-se nos fundamentos teóricos de probabilidade e do processo de Poisson. Nele, são abordados conceitos como espaço amostral, eventos e variáveis aleatórias, além de explorar as características específicas da distribuição de Poisson.

O Capítulo 2 apresenta o Estimador de Máxima Verossimilhança (EMV) e a regressão Poisson.

No Capítulo 3, trata de aplicações do modelo de regressão Poisson. O modelo é aplicado a dados reais e simulados.

O Capítulo 4, abriga as Considerações Finais.

1 Fundamentos Teóricos de Probabilidade e o Processo de Poisson

Utilizando como referências principais, (MEYER, 1983), (R., 1978) e (MORETTIN; BUSSAB, 2017) exploramos neste capítulo os conceitos fundamentais para o desenvolvimento das aplicações do trabalho. Compreender esses conceitos é fundamental para o desenvolvimento e aprofundamento das aplicações do Capítulo 3.

1.1 Espaço Amostral

Definimos espaço amostral como o conjunto de todos os resultados possíveis de um experimento aleatório. Formalmente, ele é denotado como Ω.

Por exemplo, considere o lançamento de um dado honesto e o registro da face voltada para cima. Neste caso, o espaço amostral será definido por: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

O espaço amostral pode ser finito, infinito enumerável ou infinito não enumerável. Em geral supõe-se que:

i. a todo resultado possível corresponde um, e somente um, ponto $\omega \in \Omega$; e

ii. resultados distintos correspondem a pontos distintos em Ω, i.e. ω não pode representar mais de um resultado.

1.2 Eventos

Em probabilidade, um evento é qualquer subconjunto do espaço amostral. Formalmente, dado um espaço amostral Ω, um evento A é um subconjunto de Ω, ou seja, $A \subseteq \Omega$. Um evento A que possui apenas um único elemento de Ω é chamado de evento simples. Além disso, Ω o espaço amostral e $\varnothing$ o conjunto vazio são denominados, respectivamente, de evento certo e de evento impossível.

Os eventos podem ser combinados de várias maneiras para dar origem a outos mais complexos. Por exemplo, considere os eventos A e B, temos,

- União de eventos $(A \cup B)$ é o evento que ocorre se A, ou B, ou ambos ocorrerem.
- Interseção de eventos $(A \cap B)$ é o evento que ocorre se A e B ocorrerem.
- Complementar de um evento $\bar{A}$ é o evento que ocorre quando A não ocorre.

- Diferença entre eventos $(A - B)$ é o evento que ocorre quando ocorre A e B não ocorre.

De acordo com (CASELLA; BERGER, 2006) temos a seguinte definição de eventos disjuntos.

Definição 1.1. *Dois eventos A e B são disjuntos (ou mutuamente exclusivos) se $A \cap B = \varnothing$. Os eventos $A_1, A_2, ...$ são disjuntos dois a dois (ou mutuamente exclusivos) se $A_i \cap A_j = \varnothing$ para todo $i \neq j$.*

1.3 Espaço de Probabilidade

Um espaço de probabilidade é um modelo matemático que descreve um experimento aleatório. De acordo com (MAGALHÃES, 2010), uma função $\mathcal{P}$, definida na $\sigma - álgebra$ $\mathcal{A}$ de subconjuntos de Ω e com valores no intervalo $[0, 1]$, é uma probabilidade se satisfaz os Axiomas de Kolmogorov:

Axioma 1.2. *$P(\Omega) = 1$. A probabilidade de um evento certo é igual a 1.*

Axioma 1.3. *Para todo subconjunto $A \in \mathcal{A}$, $P(A) \geq 0$.*

Axioma 1.4. *Para toda sequência $A_1, A_2, ... \in \mathcal{A}$, mutuamente exclusivos, temos*

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

A trinca $(\Omega, \mathcal{A}, \mathcal{P})$ é denominada espaço de probabilidade. Os subconjuntos que estão em $\mathcal{A}$ são chamados de eventos e é somente a eles que se atribui probabilidade.

1.4 Variáveis Aleatórias

As variáveis aleatórias são ferramentas importantes em qualquer pesquisa científica ou análise estatística, e podem ser classificadas como discretas ou contínuas.

Definição 1.5. *Seja Ω um espaço amostral associado a um experimento aleatório ε. Uma função X, que associa a cada elemento $\omega \in \Omega$ um número real, $X(\omega)$, é denominada variável aleatória.*

1.5 Variáveis Aleatórias Discretas

Definição 1.6. *Uma variável aleatória X é discreta se o número de valores possíveis de X for um número finito ou infinito enumerável.*

Exemplo 1.7. *São exemplos de variáveis aleatórias discretas:*

- *Número de acidentes de trânsito em um cruzamento durante o dia.*
- *Número de chamadas recebidas em uma central telefônica no período de uma hora.*
- *Número de itens vendidos em uma loja durante o dia.*
- *Números de células cancerígenas num determinado órgão do corpo humano.*

A distinção entre variáveis discretas e contínuas influencia a escolha do método estatístico apropriado para analisar os dados. Por exemplo, a distribuição de Poisson é frequentemente utilizada para modelar variáveis discretas, enquanto a distribuição normal é apropriada para variáveis contínuas.

No contexto deste trabalho, nosso objetivo é utilizar o modelo de Poisson para descrever e analisar variáveis discretas em um determinado conjunto de dados. Além disso, estimar os parâmetros do modelo, testar hipóteses e calcular probabilidades.

1.6 Distribuição de Poisson

Essa distribuição é caracterizada por um único parâmetro λ, que representa a taxa de ocorrência dos eventos. A função de probabilidade da distribuição de Poisson permite calcular a probabilidade de ocorrer um número específico de eventos por unidade de medida.

Definição 1.8. *Seja X uma variável aleatória discreta, como função de probabilidade*

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},\ k = 0,\ 1,\ 2,\ \ldots \tag{1.1}$$

então, X tem distribuição de Poisson com parâmetro $\lambda > 0$.

Podemos observar que a expressão (1.1) define uma distribuição de probabilidade. Com efeito,

$$\sum_{k=0}^{\infty} P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\ , \tag{1.2}$$

em que, $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$ é conhecida como série de Taylor da função exponencial e^{λ}.

Portanto,

$$\sum_{k=0}^{\infty} P(X = k) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1. \tag{1.3}$$

Sabendo que X tem distribuição de Poisson e parâmetro λ, temos que $E(X) = Var(X) = \lambda$.

$$E(X) = \sum_{k=0}^{\infty} kp(k). \tag{1.4}$$

Então,

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k(k-1)!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!}. \end{aligned}$$

Aplicando, $n = k-1$ e $k = n+1$, nesta última expressão,

$$E(X) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^{n+1}}{n!} = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n\lambda}{n!} = \lambda\sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} = \lambda. \tag{1.5}$$

Sabendo que a variância $Var(X) = E(X^2) - [E(X)]^2$, faz-se necessário calcular $E(X^2)$.

$$E(X^2) = \sum_{k=0}^{\infty} k^2\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=0}^{\infty} kk\frac{e^{-\lambda}\lambda^k}{k(k-1)!} = \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{(k-1)!}. \tag{1.6}$$

Para, $n = k-1$ e $k = n+1$, no último somatório da expressão (1.6), segue-se que

$$\begin{aligned} E(X^2) &= \sum_{n=0}^{\infty}(n+1)\frac{e^{-\lambda}\lambda^{n+1}}{n!} = \lambda\sum_{n=0}^{\infty}(n+1)\frac{e^{-\lambda}\lambda^n\lambda}{n!} \\ &= \lambda\sum_{n=0}^{\infty} n\frac{e^{-\lambda}\lambda^n}{n!} + \lambda\sum_{n=0}^{\infty}\frac{e^{-\lambda}\lambda^n}{n!} = \sum_{n=1}^{\infty}\frac{e^{-\lambda}\lambda^n}{(n-1)!} + \lambda. \end{aligned}$$

Novamente, seja $j = n-1$ e $n = j+1$,

$$\begin{aligned} E(X^2) &= \lambda\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^{j+1}}{j!} + \lambda = \lambda\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^j\lambda}{j!} + \lambda \\ &= \lambda\lambda\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^j}{j!} + \lambda = \lambda\lambda + \lambda = \lambda^2 + \lambda. \end{aligned} \tag{1.7}$$

Das expressões (1.5) e (1.7), obtermos

$$Var(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

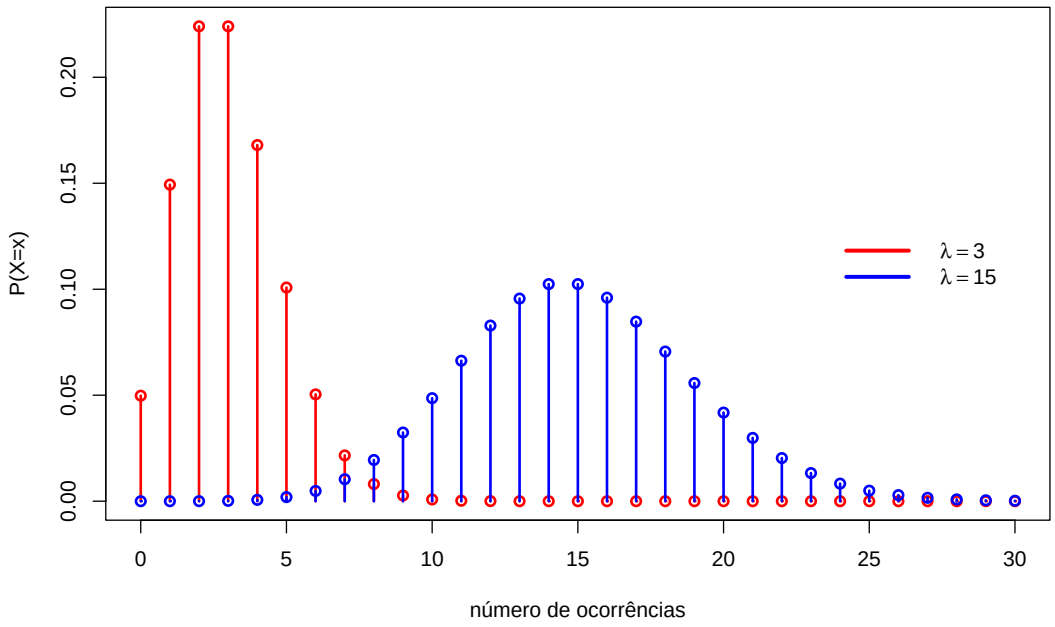
Portanto, a igualdade $Var(X) = E(X) = \lambda$ ressalta uma característica fundamental da distribuição de Poisson, na qual a média e a variância da v.a. X são determinadas pelo

mesmo valor, representada pela taxa de ocorrência. Essa relação estatística entre a média e a variância é utilizada na análise de eventos modelados pela distribuição de Poisson.

Podemos definir um evento raro como um evento A com uma baixa probabilidade de ocorrência em relação ao tamanho do intervalo considerado. Por exemplo, podemos considerar a ocorrência de 10 acidentes de aviões em um mês numa determinada região ou a quantidade de 5 gols em uma determinada partida de futebol do campeonato brasileiro da série "A".

Na Figura 1.1, o eixo das abscissas representa os valores possíveis de uma variável Poisson, enquanto eixo das ordenadas denota sua probabilidade de ocorrência.

Figura 1.1 – Funções de probabilidade de uma v.a. de Poisson.



Fonte: Próprio autor (2023).

Ao lidar com eventos raros, a distribuição de Poisson surge como uma escolha adequada para modelar sua ocorrência. Essa distribuição assume que a probabilidade de um evento ocorrer é proporcional ao tamanho do intervalo considerado. Em outras palavras, quando a taxa de ocorrência, representada por λ, é pequena em relação a uma pequena unidade de medida, $P(X = x)$, é pequena para valores grandes de X.

Além disso, a distribuição de Poisson também é útil na modelagem de eventos que ocorrem em uma sequência de intervalos independentes e com a mesma taxa média de ocorrência. Isso significa que a distribuição de Poisson pode ser aplicada a eventos que ocorrem de forma independente no tempo, como a contagem de células de câncer em diferentes regiões do corpo humano ao longo de um determinado período.

1.7 Distribuição de Poisson como Aproximação da Binomial

Teorema 1.9. *A variável aleatória de Poisson também pode ser usada como uma aproximação para a variável aleatória binomial com parâmetros (n, p) no caso particular em que n grande e p suficientemente pequeno para que np tenham tamanho moderado (ROSS, 2010).*

Seja X uma variável aleatória distribuída binomialmente com parâmetros n e p e função de probabilidade

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}. \quad k = 0, 1, \cdots, n. \tag{1.8}$$

Admita-se que quando $n \to \infty$, $np = \lambda$ fique constante, ou equivalentemente, quando $n \to \infty$, $p \to 0$, de modo que $np \to \lambda$. Nessas condições temos

$$\lim_{n\to\infty} P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

que é a distribuição de Poisson com parâmetro λ.

Seja a expressão (1.8) geral binomial,

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)(n-2)...(n-k+1)}{k!} p^k (1-p)^{n-k}. \end{aligned}$$

Seja $np = \lambda$. Logo, $p = \frac{\lambda}{n}$, e $1 - p = 1 - \frac{\lambda}{n} = \frac{(n-\lambda)}{n}$. Substituindo os termos que possuem p pela expressão equivalente em termos de λ,

$$\begin{aligned} P(X = k) &= \frac{n(n-1)(n-2)...(n-k+1)}{k!} (\frac{\lambda}{n})^k (\frac{n-\lambda}{n})^{n-k} \\ &= \frac{\lambda^k}{k!}[(1)(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})][1-\frac{\lambda}{n}]^{n-k} \\ &= \frac{\lambda^k}{k!}[(1)(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})](1-\frac{\lambda}{n})^n (1-\frac{\lambda}{n})^{-k} \end{aligned}$$

Fazendo $n \to \infty$, de modo que $np = \lambda$ continue constante. Portanto, $p \to 0$ e $n \to \infty$. Outra equivalência seria $n \to \infty$ e $p \to 0$, de modo que $np \to \lambda$.

Note que, nos termos na forma $(1 - \frac{1}{n})$, $(1 - \frac{2}{n})$, ..., a medida que $n \to \infty$ os termos aproximam-se de 1. Daí, teremos $(1 - \frac{\lambda}{n})^{-k}$.

Pela definição de exponencial, $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$ quando $n \to \infty$. Portanto, aplicando limite em $P(X = k)$, obtemos,

$$\lim_{n\to\infty} P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}. \tag{1.9}$$

Logo, no limite, teremos a distribuição de Poisson com parâmetro λ.

1.8 Processo de Poisson

Definição 1.10. *Um processo de contagem $\{N(t), t \geq 0\}$ é chamado de Processo de Poisson com taxa $\lambda > 0$, quando satisfaz as seguintes condições:*

(i) (Incrementos independentes) Para quaisquer instantes de tempo não negativo $t_1 < t_2 < ... < t_n$ as variáveis aleatórias $N(t_2) - N(t_1), N(t_3) - N(t_2), ..., N(t_n) - N(t_{n-1})$ são independentes. Em outras palavras, o número de ocorrências em intervalos de tempo disjuntos são independentes entre si.

(ii) (Incrementos estacionários)Para todo $s > 0$, as v.a.'s $N(t)$ e $M(t)$ que representam o número de ocorrências no intervalo $[0, t)$ e $[s, s + t)$, respectivamente, têm a mesma distribuição.

(iii) Seja t suficientemente pequeno, a probabilidade $P(N(t) = 1) \approx \lambda h$.

(iv) Para t suficientemente pequeno em (iii), a probabilidade $P(N(t) \geq 2) \approx 0$. Em outras palavras, a probabilidade de haver duas ou mais ocorrências em um intervalo suficientemente pequeno é desprezível.

(v) $P(N(0) = 0) = 1$ (não há ocorrência no tempo inicial);

Utilizando os itens acima (i) a (v), podemos deduzir uma expressão para $P(N(t) = n) = p_n(t)$.

Observe os itens (i) e (ii), onde são $N(t)$ e $N(t + h) - N(t)$ v.a.'s independentes.

Considerando $n = 0$, $P(N(t + h)) = 0$ podemos denotar as seguintes observações:

$$N(t + h) = 0;\ N(t) = 0;\ N(t + h - t) = N(h) = 0.$$

Logo, $P(N(t + h) = 0) = P(N(t) = 0, N(t + h - t) = 0)$, já que os intervalos são disjuntos,

$$P[N(t) = 0; N(t + h - t) = 0] = P(N(t) = 0)P(N(h) = 0). \tag{1.10}$$

No entanto, $P(N(h)=0)$, podendo ser denotado por $p_0(h)$, daí, usando a noção de complementar,

$$\begin{aligned} P(N(h)=0) &= 1-P(N(h)>0) \\ &= 1-[P(N(t)=1)+P(N(h)>1)]. \end{aligned}$$

Portanto, por (iii) e (iv),

$$p_0(h) = 1-p_1(h)-\sum_{k=2}^{\infty} p_k(h) \sim 1-\lambda t+p(h) \tag{1.11}$$

quando $t \to 0$.

Então, na forma (1.10)

$$\begin{aligned} p_0(t+h) &= P[N(t+h)=0] \\ &= P[N(t)=0; N(t+h)-N(t)=0] = P[N(t)=0; N(t+h-t)=0] \\ &= P[N(t)=0; N(h)=0] \\ &= p_0(t)p_0(h) \sim p_0(t)[1-\lambda h] \end{aligned} \tag{1.12}$$

Deduzindo (1.12),

$$\begin{aligned} p_0(t+h) &\sim p_0(t)[1-\lambda h] \\ &\sim p_0(t)-p_0(t)\lambda h \\ p_0(t+h)-p_0(t) &\sim -p_0(t)\lambda h \\ \frac{p_0(t+h)-p_0(t)}{h} &\sim -p_0(t)\lambda. \end{aligned}$$

Fazendo $h \to 0$,

$$\lim_{h\to 0} \frac{p_0(t+h)-p_0(t)}{h} = \lim_{h\to 0} -p_0(t)\lambda. \tag{1.13}$$

Pela definição de derivada em (1.13),

$$p_0'(t) = -\lambda p_0(t) \tag{1.14}$$

Note que (1.14) é uma equação diferencial ordinária de primeira ordem. Então, resolvendo-a,

$$p_0'(t) = -\lambda p_0(t) \Rightarrow \frac{p_0'(t)}{p_0(t)} = -\lambda dt.$$

Integrando,

$$\int \frac{p_0'(t)}{p_0(t)} = \int -\lambda dt \Rightarrow \ln p_0(t) = -\lambda t + k \ (constante)$$

Por (iv) sendo $t = 0$, temos que $c = 0$. Então,

$$p_0(t) = e^{-\lambda t}. \tag{1.15}$$

Portanto, obtivemos uma expressão para $P(N(t) = 0)$. Deste modo, podemos determinar, para $n > 0$, $p_n(t)$.

Considere os intervalos: $N(t) = n - k$; $N(h) = k$; $N(t + h) = n$. Além de, $p_n(t + h) = P[N(t + h) = n]$. Portanto,

$$p_n(t+h) = \sum_{i=0}^{k} p_i(t)p_{n-i}(h) = p_n(t)p_0(h) + p_{n-1}(t)p_1(h) + \sum_{i=0}^{k-2} p_i(t)p_{n-i}(h). \tag{1.16}$$

Já que $p_0(h) = \lambda h$, pelos itens (iii), (iv) e (Equação 1.15),

$$\begin{aligned}
p_n(t+h) &\sim p_n(t)[1-\lambda h] + p_{n-1}(t)\lambda h \\
&\sim p_n(t) - \lambda h p_n(t) + \lambda h p_{n-1}(t) \\
p_n(t+h) - p_n(t) &\sim -\lambda h p_n(t) + \lambda h p_{n-1}(t) = h[-\lambda p_n(t) + \lambda p_{n-1}(t)] \\
\frac{p_n(t+h) - p_n(t)}{h} &\sim -\lambda p_n(t) + \lambda p_{n-1}(t).
\end{aligned}$$

Fazendo $h \to 0$,

$$\begin{aligned}
\lim_{h\to 0} \frac{p_n(t+h) - p_n(t)}{h} &= \lim_{h\to 0}(-\lambda p_n(t) + \lambda p_{n-1}(t)). \\
p_n'(t) &= -\lambda p_n(t) + \lambda p_{n-1}(t) \ \text{(Eq. de recorrência).}
\end{aligned} \tag{1.17}$$

Sendo o fator de integração $\mu = e^{\lambda t}$, multiplicando na (Eq.1.17)

$$\begin{aligned}
e^{\lambda t}p_n'(t) &= -e^{\lambda t}\lambda p_n(t) + \lambda e^{\lambda t}p_{n-1}(\lambda) \\
e^{\lambda t}p_n'(t) + e^{\lambda t}\lambda p_n(t) &= \lambda e^{\lambda t}p_{n-1}(\lambda) \\
(e^{\lambda t}p_n(t))' &= \lambda e^{\lambda t}p_{n-1}(t).
\end{aligned} \tag{1.18}$$

Para $n = 1$ e, sendo $p_0(t) = e^{-\lambda t}$,

$$\begin{aligned}
(e^{\lambda t}p_n(t))' &= \lambda e^{\lambda t}p_{n-1}(t) \\
(e^{\lambda t}p_1(t))' &= \lambda e^{\lambda t}p_{1-1}(t) \\
&= \lambda e^{\lambda t}p_0(t) = \lambda e^{\lambda t}e^{-\lambda t} = \lambda.
\end{aligned}$$

Integrando,

$$\int (e^{\lambda t} p_1(t))' = \int \lambda dt \Rightarrow e^{\lambda t} p_1(t) = \lambda t + k \ \ (constante)$$

Observe que $p_1(0) = 0$, então, $c = 0$. Portanto,

$$p_1(t) = \lambda t e^{-\lambda t}.$$

Para $n = 2$, utilizando (Eq. 1.17),

$$\begin{aligned} (e^{\lambda t} p_2(t))' &= \lambda e^{\lambda t} p_{2-1}(t) \\ &= \lambda e^{\lambda t} p_1(t) = \lambda e^{\lambda t} \lambda t e^{-\lambda t} \\ (e^{\lambda t} p_2(t))' &= \lambda^2 t. \end{aligned}$$

Integrando,

$$\begin{aligned} \int (e^{\lambda t} p_2(t))' &= \int \lambda^2 t dt \Rightarrow e^{\lambda t} p_2(t) = \frac{\lambda^2 t^2}{2} + k \\ p_2(t) &= e^{-\lambda t} [\frac{\lambda^2 t^2}{2} + k]. \end{aligned}$$

Seja $p_2(0) = 0$, então, $c = 0$. Portanto,

$$p_2(t) = \frac{e^{-\lambda t} (\lambda t)^2}{2}$$

Utilizando o método de indução:

Considere $p_n(t) = \frac{e^{-\lambda} (\lambda t)^n}{n!}$.

Para $n + 1$, temos

$$\begin{aligned} (e^{\lambda t} p_{n+1}(t))' &= \lambda e^{\lambda t} p_n(t) \\ &= \lambda e^{\lambda t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \end{aligned}$$

Integrando,

$$\begin{aligned} \int (e^{\lambda t} p_{n+1}(t))' &= \int \lambda e^{\lambda t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} dt \\ e^{\lambda t} p_{n+1}(t) &= \int \lambda \frac{(\lambda t)^n}{n!} dt \\ &= \int \frac{\lambda^{n+1} t^n}{n!} dt = \frac{\lambda^{n+1} t^{n+1}}{(n+1) n!} + k. \end{aligned}$$

Veja, $p_{n+1}(0) = 0$ e $c = 0$. Portanto,

$$p_{n+1}(t) = \frac{e^{-\lambda t} (\lambda t)^{n+1}}{(n+1)!}.$$

Consequentemente, mostramos que a probabilidade de ocorrência durante o intervalo de tempo $[0, t)$, é uma v.a, $P(N(t) = n)$ tem distribuição de Poisson, com parâmetro (λt).

2 Estimador de Máxima Verossimilhança e Regressão Poisson

Neste capítulo, discutiremos o estimador de máxima verossimilhança, enfatizando sua definição e propriedades. Em seguida, abordamos a regressão Poisson e suas características, seguindo as referências (MORETTIN; BUSSAB, 2017) e (MEYER, 1983).

2.1 Definições

Seja $(x_1, x_2, ..., x_n)$ uma amostra aleatória $(a.a)$ de tamanho n de uma variável aleatória X. O parâmetro que estamos interessados estimar é representado por θ.

Definição 2.1. *Dado X com uma distribuição de probabilidade dependente de um parâmetro ou vetor de parâmetros desconhecido θ e uma amostra $X_1, ..., X_n$ com valores correspondentes $x_1, ..., x_n$, usamos $g(X_1, ..., X_n)$ para estimar θ, representando a estimativa como $\hat{\theta} = g(x_1, ..., x_n)$.*

Definição 2.2. *Se $\hat{\theta}$ estima θ em X, então $\hat{\theta}$ é não-tendencioso se $E(\hat{\theta}) = \theta$ para todo θ.*

Definição 2.3. *Seja $\hat{\theta}$ uma estimativa não-tendenciosa de θ. Diremos que $\hat{\theta}$ é uma estimativa não-tendenciosa, de variância mínima de θ, se para todas as estimativas θ^* tais que $E(\theta^*) = \theta$, tivermos $Var(\hat{\theta}) \leq Var(\theta^*)$ para todo θ. Isto é, dentre todas as estimativas não-tendenciosos de θ, $\hat{\theta}$ tem a variância menor de todas.*

Definição 2.4. *Seja $\hat{\theta}$ um estimador do parâmetro θ. Diremos que $\hat{\theta}$ é um estimador consistente se,*

$$\lim_{n\to\infty} P[|\hat{\theta} - \theta| > \varepsilon] = 0, \forall \varepsilon > 0$$

ou, equivalentemente, se

$$lim_{n\to\infty} P[|\hat{\theta} - \theta| \leq \varepsilon] = 1, \forall \varepsilon > 0$$

Teorema 2.5. *Seja $\hat{\theta}$ um estimador de θ baseado em uma amostra aleatória de tamanho n. Se $lim_{n\to\infty} E(\hat{\theta}) = \theta$, e se $lim_{n\to\infty} Var(\hat{\theta}) = 0$, então $\hat{\theta}$ será um estimador consistente de θ.*

Demonstração: Considere a desigualdade de Chebyschev:

$$P[|X - c| \geq \varepsilon] \leq \tfrac{1}{\varepsilon^2} E(X - c)^2$$

Portanto, reescrevendo, temos

$$P[|\hat{\theta} - \theta| \geq \varepsilon] \leq \frac{1}{\varepsilon^2} E[\hat{\theta} - \theta]^2$$

$$\begin{aligned} &= \frac{1}{\varepsilon^2} E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= \frac{1}{\varepsilon^2} E\{[\hat{\theta} - E(\hat{\theta})]^2 + 2[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] + [E(\hat{\theta}) - \theta]^2\} \\ &= \frac{1}{\varepsilon^2} \{Var\hat{\theta} + 0 + [E(\hat{\theta}) - \theta]^2\}. \end{aligned}$$

Portanto, fazendo $n \to \infty$ e empregando as hipóteses do teorema, encontraremos

$$\lim_{n \to \infty} P[|\hat{\theta} - \theta| \leq \varepsilon] \geq 0$$

e, por isso, igual a 0 (Zero). ■

Teorema 2.6. *Seja X uma variável aleatória com esperança finita μ e variância σ^2. Seja $\bar{X}$ a média amostral, baseada em uma amostra aleatória de tamanho n. Nesse caso $\bar{X}$ será uma estimativa não-tendenciosa e consistente de μ.*

2.2 Estimadores de Máxima Verossimilhança

Nesta seção, vamos explorar o método de máxima verossimilhança (MV), uma abordagem essencial na área da estatística que se concentra na obtenção de estimativas para um parâmetro específico associado a uma distribuição de probabilidade.

Considere X uma variável aleatória, cujo função de probabilidade (fdp) pode ser expressa como $f(x;\theta)$.

Dada uma amostra aleatória $X_1, ..., X_n$ proveniente da variável aleatória X, na qual os valores amostrais são representados por $x_1, ..., x_n$, introduzimos a função de verossimilhança L. Essa função é definida como uma expressão que depende tanto da amostra quanto do parâmetro θ, sendo:

$$L(X_1, ..., X_n; \theta) = f(X_1; \theta) f(X_2; \theta) ... f(X_n; \theta) \tag{2.1}$$

Se X for discreta, $L(x_1, ..., x_n; \theta)$ representará $P[X_1 = x_1, ..., X_n = x_n]$, enquanto se X for contínua, $L(x_1, ..., x_n; \theta)$ representará a função de probabilidade conjunta de $(X_1, ..., X_n)$.

Definição 2.7. *Seja X uma v.a com fdp $f(x;\theta)$, onde θ é o parâmetro desconhecido. Considere uma amostra aleatória $X_1, ..., X_n$ de X. Portanto, (2.1) a função de verossimilhança, sendo:*

$$\begin{aligned} L(X_1, ..., X_n; \theta) &= f(X_1; \theta) f(X_2; \theta) ... f(X_n; \theta) \\ &= \prod_{i=1}^{n} f(X; \theta) \end{aligned}$$

Para determinar o estimador de máxima verossimilhança (EMV), faz-se necessário determinar-se o valor máximo da função. Portanto, é conveniente aplicar-se o logaritmo em L, uma vez que $\log x$ é crescente e bijetiva, sendo:

$$\log L(X_1, ..., X_n) = \ln L(X_1, ..., X_n) \tag{2.2}$$

Definição 2.8. *Sob condições gerais, assumindo que $\theta \in \mathbb{R}$ e que $L(X_1, .., X_n; \theta)$ seja uma função derivável em relação a θ, podemos obter uma estimativa de máxima verossimilhança (MV) para θ a partir da solução da equação que maximiza $L(X_1, ..., X_n; \theta)$, sendo,*

$$\frac{\partial}{\partial\theta} L(X_1, ..., X_n; \theta) = 0 \tag{2.3}$$

Tal equação é conhecida como equação de verossimilhança.

2.2.1 Propriedades das estimativas de MV

(a) A estimativa de MV pode ser tendenciosa em algumas circunstâncias, e essa tendenciosidade pode ser corrigida por meio da multiplicação por uma constante apropriada.

(b) As estimativas de MV são consistentes. Isso significa que, quando o tamanho da amostra é grande, a estimativa de MV se aproxima do valor verdadeiro do parâmetro a ser estimado.

(c) (Invariância) Considere $\hat{\theta}$ como a estimativa de MV de θ, então podemos mostrar que a estimativa de MV de $g(\theta)$, sendo g contínua e monótona, é $g(\hat{\theta})$. Logo, se o estatístico A faz sua mensuração em metros quadrados (m^2), o estatístico B mede em metros (m), e se a estimativa de MV de A for θ, então, a estimativa de B será $\sqrt{\hat{\theta}}$.

(d) (Assintótica das estimativas de MV). Caso $\hat{\theta}$ seja uma estimativa de MV para o parâmetro θ, baseada em uma amostra aleatória $X_1, ..., X_n$ de uma v.a X, logo, para um n suficientemente grande, a v.a terá aproximadamente a distribuição

$$\hat{\theta} \approx N(\theta, \frac{1}{I_F}), \tag{2.4}$$

onde

$$I_F = nE[\frac{\partial}{\partial\theta} \ln f(X; \theta)]^2 \tag{2.5}$$

aqui f é a distribuição de probabilidade por pontos ou a fdp de X, dependendo de ser X discreta ou contínua, e se supõe que $\theta \in \mathbb{R}$.

2.3 Regressão Poisson

O estimador de máxima verossimilhança é usado para ajustar um modelo linear generalizado como a regressão de Poisson. Como no Capítulo 1, utilizamos uma v.a. X com distribuição de Poisson com média μ.

Além disso, a média depende de variáveis independentes $x_1, x_2, ..., x_n$. Logo, a relação entre μ e as variáveis independentes é modelada através da função linear, onde $\log(\mu)$ é uma combinação linear das variáveis aleatórias e dos coeficientes $\beta_0, \beta_1, ..., \beta_n$:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{2.6}$$

Para cada x_i, existem n observações que são independentes da variável aleatória de Poisson $(X_1, ..., X_n)$. A média μ de X, dada por, $\mu = \lambda N$.

A média μ de X não é determinada apenas por variáveis independentes, mas também pode ser influenciada pelo grau de exposição, denotado por N (offset).

Portanto, temos o processo de Poisson que ocorre em níveis variados de "exposição". A exposição pode indicar quanto tempo ou espaço está disponível para que haja ocorrência.

Logo,

$$\log(\mu) = \log(\lambda) + \log(N). \tag{2.7}$$

E, $\log(\lambda)$ é uma função linear dada por:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

Na regressão Poisson, a exposição N é uma variável crucial, mas não é um componente ativo da própria equação de regressão. Em vez disso, é visto como um fator de ajuste que representa o risco total ou o tempo até o risco. Neste caso, N é conhecida, desempenha um papel importante na modelagem, mas não é uma variável preditora.

Sua inclusão é fundamental para interpretação completa dos resultados e revisões das estimativas de outras variáveis explicativas. Portanto, consideraremos cuidadosamente o efeito de N, mas não é incluído diretamente na equação de regressão.

Note que, caso a variável aleatória X seja única e binária (0 e 1). Temos,

$$\begin{aligned} \log(\lambda) &= \beta_0 = \lambda_0, \ x = 0 \\ \log(\lambda) &= \beta_0 + \beta_1 = \lambda_1, \ x = 1 \end{aligned}$$

Portanto, a diferença,

$$\log(\lambda_1) - \log(\lambda_0) = \beta_1$$

Aplicando a exponencial,

$$\begin{aligned} e^{\log(\lambda_1)-\log(\lambda_0)} &= e^{\beta_1} \Rightarrow e^{\log \frac{\lambda_1}{\lambda_0}} = e^{\beta_1} \\ e^{\beta_1} &= \frac{\lambda_1}{\lambda_0}. \end{aligned} \tag{2.8}$$

3 Aplicações do Modelo de Regressão Poisson

Neste capítulo, iniciaremos uma análise essencial para compreensão do estudo. Primeiramente, vamos explorar um exemplo da Regressão de Poisson aplicada à dados simulados. A simulação de dados é uma ferramenta poderosa em estatística, permitindo-nos explorar padrões, testar hipóteses e aprimorar nossos métodos analíticos. Neste contexto, realizamos uma simulação de dados para exemplificar e analisar um modelo de regressão Poisson.

Posteriormente, apresentamos dois exemplos com conjuntos de dados reais sobre incidência do câncer de pulmão em quatro cidades Dinamarquesas e mortalidade por câncer de mama.

Os dados reais estão disponíveis no pacote *ISwR* (Introduction to the Theory of Statistics with R) (DALGAARD, 2019), especificamente nas seções relacionada ao conjunto de dados *eba1977* e *bcmort*.

A linguagem R (R Core Team, 2022) foi utilizada para realizar as análises.

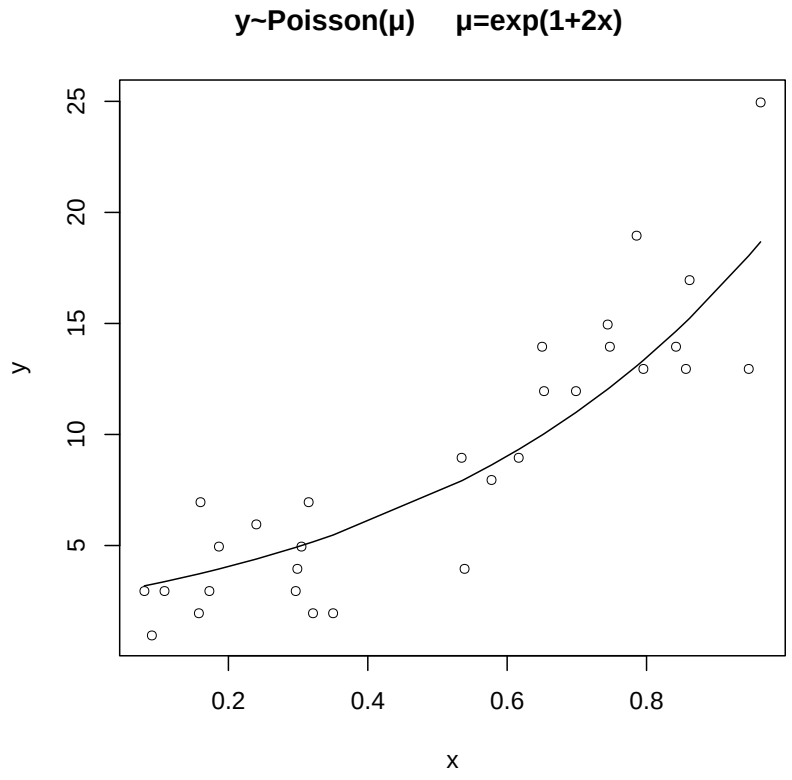
3.1 Dados Simulados do Modelo de Regressão Poisson

Simulamos os dados de contagem (y) utilizando o modelo de regressão Poisson. Sendo definido com uma média verdadeira (μ) expressa como e^{1+2x}, onde x é gerado aleatoriamente de uma distribuição uniforme no intervalo $[0, 1]$. Os dados seguintes representam $n = 30$ observações simuladas da variável Y.

$$y = 3, 1, 3, 2, 7, 3, 5, 6, 3, 4, 5, 7, 2, 2, 9, 4, 8, 9, 14, 12, 12, 15, 14, 19, 13, 14, 13, 17, 13, 25.$$

Cada valor y corresponde a uma observação única do preditor x.

Figura 3.1 – Contagens simuladas por observação da variável preditora X.



Fonte: Próprio autor (2023)

A linha representa a verdadeira média em execução $\mu = e^{1+2x}$ que foi utilizada para gerar os dados simulados. Essa curva é a relação subjacente entre a variável preditora X e a média de Y.

Note que os resultados obtidos, são denotados como os valores de β_0 e β_1. Portanto, temos um vetor unidimensional,

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 0,8531837 \\ 2,637467 \end{bmatrix}.$$

Podemos interpretar os valores da seguinte maneira:

- $\beta_0 = 0,8531837$: Denota o logaritmo natural da média de Poisson quando $x = 0$. Ou seja, utilizando a equação (2.6), obtemos:

$$\begin{aligned} \log(\mu) &= \beta_0 + \beta_1 x_1 \Rightarrow \log(\mu) = 0,8531837 + 2,637467.0 \\ \log(\mu) &= 0,8531837. \end{aligned}$$

- $\beta_1 = 2,637467$: Representa como o logaritmo natural da média de Poisson varia com uma unidade de mudança em x. Ao exponenciar β_1 obteremos a razão de mudança

na média para uma unidade de aumento em x. Então, pela equação (2.8),

$$e^{\beta_1} = e^{2,637467} \approx 13,97775.$$

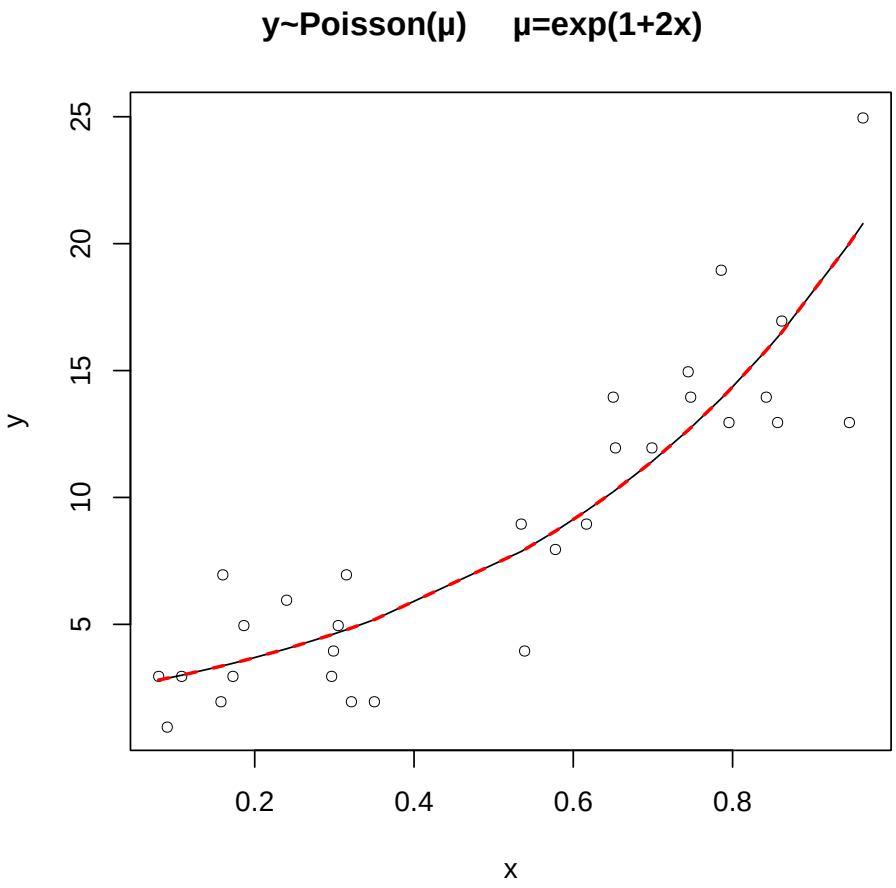
A exponenciação do β_1 prediz que a cada unidade de aumento em x, a média estimada da distribuição de Poisson é, aproximadamente, multiplicada por $13,97775$.

Portanto, para os valores específicos gerados na simulação. Obtermos uma média (μ), dada por,

$$\mu = e^{0,8531837+2,2637467.x} \tag{3.1}$$

Note que a média da Poisson (3.1) é estimada com base nos coeficientes do modelo ajustado. A partir daí, temos a seguinte representação gráfica.

Figura 3.2 – Modelo Ajustado.



Fonte: Próprio autor (2023)

A linha tracejada em vermelho representa a estimativa do modelo ajustado para média de Poisson (μ). Portanto, é obtida a partir do ajuste do modelo aos dados simulados, utilizando os parâmetros estimados, aqui representados por $\beta_0 = 0,8531837$ e $\beta_1 =$

$2,2637467$ que prevê a média de Poisson de cada valor de x, já previsto pelo processo iterativo de mínimos quadrados ponderados (IRLS). IRLS é uma técnica utilizada para estimar os parâmetros em modelos de regressão.

Faz-se necessário calcularmos a matriz de informação de Fisher, onde seu papel é de extrema importância para estimação dos erros-padrão, aqui denotada por I, utilizando a equação,

$$I = (X^T W X)^{-1}, \tag{3.2}$$

sendo X a matriz de variáveis independentes dos preditores,

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}.$$

W é uma matriz diagonal com os valores estimados de μ na diagonal,

$$W = diag(\mu_1, \mu_2, ..., \mu_n)$$

Logo, seja uma matriz 2 x 2, a matriz de informação de Fisher pode ser representada na forma:

$$I = \begin{bmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{bmatrix}$$

E seus elementos são calculados pela equação (3.2), então,

$$i_{11} = (X^T W X)^{-1}_{11},\ ...,\ i_{22} = (X^T W X)^{-1}_{22}$$

Os elementos de I, representam as variâncias e covariâncias dos coeficientes estimados no modelo de regressão de Poisson.

```
>  sqrt(diag(I))
x
0.1742004  0.2452351
```

Os valores obtidos a partir da extração da $\sqrt{}(diag(I))$ apresenta os erros-padrão para os coeficientes estimados.

- Um erro-padrão de aproximadamente $0,1742004$ para o coeficiente associado à variável preditora x.

- Um erro-padrão de aproximadamente $0,2452351$ para o intercepto do modelo de regressão.

```
> summary(glm(y~1+x,family='poisson'))
Call:
glm(formula = y ~ 1 + x, family = "poisson")

Deviance Residuals:
 Min      1Q       Median   3Q       Max
-1.6737  -0.7103   0.0568   0.6114   1.7241
Coefficients:
           Estimate Std. Error  z value Pr(>|z|)
(Intercept)   0.8532     0.1742   4.898 9.69e-07 ***
x             2.2637     0.2452   9.231  < 2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 120.568  on 29  degrees of freedom
Residual deviance:  24.072  on 28  degrees of freedom
AIC: 141.17

Number of Fisher Scoring iterations: 4
```

Primeiramente discutiremos os resultados obtidos em Deviance Residuals. Os Deviance Residuals (Desvio Residual) são importantes para avaliarmos o quão ajustado o modelo é com relação aos dados. Eles denotam a diferença entre os valores observados e os valores preditos pelo modelo. Em termos gerais, quanto menor a magnitude dos resíduos dada por:

$$\begin{aligned} Mag(Min) &= |Deviance\ R.\ Min| \\ Mag(Max) &= |Deviance\ R.\ Max|, \end{aligned}$$

melhor será o ajuste do modelo aos dados. Caso os resíduos sejam bem próximos, temos que o modelo é capaz de explicar mais precisamente os dados observados.

No caso específico, os valores são: $Min = -1,6737$ e $Max = 1,7241$.

$$\begin{aligned} Mag(Min) &= |-1,6737| = 1,6737 \\ Mag(Max) &= |1,7241| = 1,7241. \end{aligned}$$

Sugerem que, em média, os valores observados estão relativamente próximos dos previstos pelo modelo. Portanto, obtivemos o ajuste razoável, com alguma variabilidade nas predições.

O valor estimado para o intercepto é de aproximadamente $0,8532$. A qual representa a estimativa do $\log(\mu)$ para $x = 0$. Além disto, o erro padrão associado a essa estimativa é aproximadamente $0,1742$. Ela indica a incerteza em torno da estimativa do intercepto.

O termo z ajuda avaliar a significância de um coeficiente estatístico. Para o intercepto da simulação, $z \approx 4,898$, indicando que este coeficiente é $4,898$ vezes o erro padrão distante de zero. Portanto, podemos inferir com um grande nível de confiança que o intercepto é estatisticamente significativo para a modelagem.

Ou seja, o coeficiente ser estatisticamente significativo, implica que há uma relação entre a variável preditora e a variável de resposta. Em outras palavras, a variável preditora não está desempenhando um papel aleatório nos resultados, mas contribuindo de maneira confiável para a explicação da variação na variável de interesse.

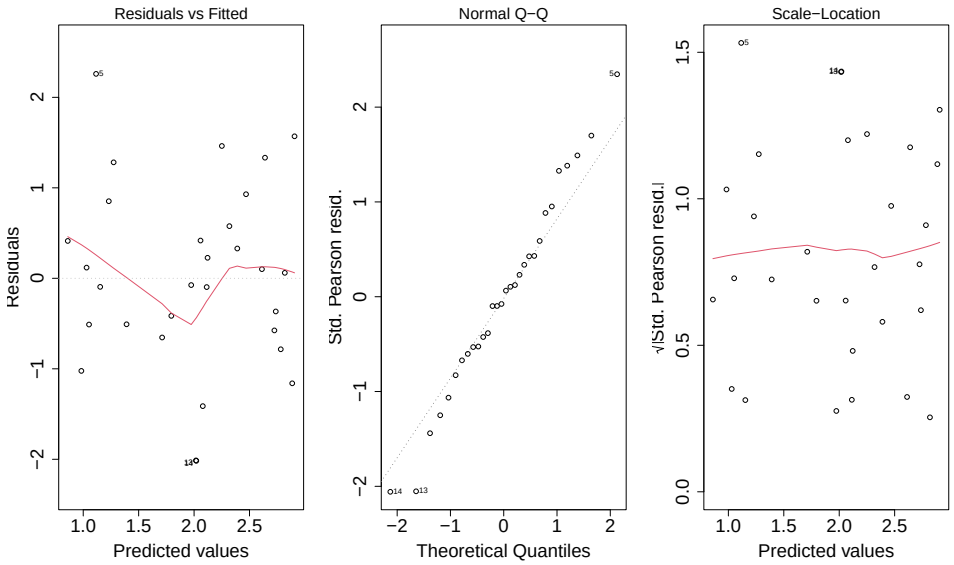
Note, também, que a estimativa associada à variável preditora x é aproximadamente $2,2637$, já descrito anteriormente. Além de seu erro padrão de aproximadamente $0,2452$. A qual representa incerteza na estimativa do coeficiente para x.

Para o valor de $z \approx 9,231$, temos, também alta significância estatística para x. Portanto, a variável preditora tem influência estatisticamente nas contagens observadas, indicando uma relação substancial entre x e y.

O Null Deviance (Desvio Nulo) é de $120,568$ para 29 graus de liberdade, enquanto o Residual Deviance (Desvio Residual) é de $24,072$ com 28 graus de liberdade. Um desvio residual menor em comparação ao desvio nulo indica que o modelo ajustado fornece uma melhor explicação para os dados do que um modelo sem variáveis preditoras.

E o por fim, o critério de informação AIC (Akaike Information Criterion) é $141,17$. Em geral, quanto menor o valor AIC, melhor o modelo em relação aos outros modelos considerados. Neste contexto, a qualidade do ajuste é satisfatória para a análise dentro do modelo de regressão Poisson aos dados observados.

Figura 3.3 – Relação entre resíduos e valores preditos, quantis dos resíduos com uma normal, variância constante dos resíduos.



Fonte: Próprio autor (2023)

3.2 Modelo de Regressão Poisson em Dados de Câncer de Pulmão

O objetivo desta análise é ajustar um modelo de Regressão de Poisson na análise dos dados contidos no pacote "ISwR"seção "eba1977", disponível no ambiente R. Os dados são dispostos em números de casos de câncer de pulmão em quatro cidades dinamarquesas de 1968 a 1971, sendo elas: Fredericia, Horsens, Kolding e Vejle. Vale ressaltar que o modelo de regressão de Poisson desta análise que foi ajustado, a categoria de referência para variável cidade (city) é Fredericia. Para modelos categóricos como este, a categoria de referência é aquela contra a qual as outras categorias são comparadas, e ela geralmente não aparece explicitamente na saída do modelo, como será observado ao decorrer da investigação. Assumiremos que o número de incidência de casos é diretamente proporcional ao tamanho da população.

```
> summary(eba1977)
city           age          pop              cases            lpop
Fredericia:6   40-54:4   Min.   : 509.0   Min.   : 2.000   Min.   :6.232
Horsens    :6   55-59:4   1st Qu.: 628.0   1st Qu.: 7.000   1st Qu.:6.443
Kolding    :6   60-64:4   Median : 791.0   Median :10.000   Median :6.673
Vejle      :6   65-69:4   Mean   :1100.3   Mean   : 9.333   Mean   :6.814
                70-74:4   3rd Qu.: 954.8   3rd Qu.:11.000   3rd Qu.:6.860
                75+  :4   Max.   :3142.0   Max.   :15.000   Max.   :8.053
```

```
age.mid         e.rate
Min.   :47.00   Min.   :0.001273
1st Qu.:57.00   1st Qu.:0.007099
Median :64.50   Median :0.012870
Mean   :63.33   Mean   :0.011900
3rd Qu.:72.00   3rd Qu.:0.016602
Max.   :75.00   Max.   :0.022187
```

Em cada cidade analisada, há registros de 6 observações ou pontos de dados coletados. Além disto, a população varia de $Min.: 509,0$ e $Max.: 3142,0$ indivíduos, com uma média $mean: 791,0$. Quanto ao número de casos registados, estes variam de $Min.: 2,0$ a $Max.: 15,0$.

```
> mean(eba1977$cases)
[1] 9.33333
>var(eba1977$cases)
[1] 9.971014
```

Analisando o modelo, vamos observar a equidispersão de formas distintas, condição na qual a média é, aproximadamente, igual à variância, uma vez que tal característica é crucial pra a validade do ajuste.

Note que média de aproximadamente $mean: 9,333333$ e a variância $var: 9,971014$, são próximos, denotando que o modelo proposto possui ajuste satisfatório. Ou seja, sendo os valores de média e a variância, de casos, próximos, sugere que a modelagem de Poisson é apropriada.

Esta proximidade indica que o número de casos poder ser adequadamente modelado por um processo de Poisson, onde a variação é principalmente devido à média.

Observando os quartis, na categoria casos (cases), nota-se que em 1st Qu.(5%): $\leq 7,0$ de casos, e 3rd Qu. (75%): $\leq 11,0$. Logo, esta distribuição sugere uma variabilidade moderada no número de casos entre diferentes grupos.

```
Call:
glm(formula = cases ~ city + age, family = poisson, data = eba1977,
offset = lpop)

Deviance Residuals:
Min        1Q        Median        3Q       Max
-2.63573  -0.67296  -0.03436   0.37258   1.85267
```

```
Coefficients:
           Estimate Std. Error  z value Pr(>|z|)
(Intercept)  -5.6321     0.2003 -28.125  < 2e-16 ***
cityHorsens  -0.3301     0.1815  -1.818   0.0690 .
cityKolding  -0.3715     0.1878  -1.978   0.0479 *
cityVejle    -0.2723     0.1879  -1.450   0.1472
age55-59      1.1010     0.2483   4.434 9.23e-06 ***
age60-64      1.5186     0.2316   6.556 5.53e-11 ***
age65-69      1.7677     0.2294   7.704 1.31e-14 ***
age70-74      1.8569     0.2353   7.891 3.00e-15 ***
age75+        1.4197     0.2503   5.672 1.41e-08 ***
---

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 129.908  on 23  degrees of freedom
Residual deviance:  23.447  on 15  degrees of freedom
AIC: 137.84

Number of Fisher Scoring iterations: 5
```

Observe o valor Residual Deviance: $23,447$ e Degrees of Freedon: 15, os valores estão, relativamente próximos um do outro. Realizando a razão entre os valor:

$$\frac{23,447}{15} \approx 1,5631$$

Temos um resultado próximo de 1 que é um indicativo de que o modelo é um bom ajuste, pois sugere que não há sobredispersão significativa ou subdispersão nos dados.

Uma vez que X_i é a contagem dos incidentes de câncer de pulmão e, seja m_i o tamanho da amostra (população), o modelo de regressão é dado pela equação (2.7):

$$\begin{aligned} \log(\mu) &= \log(\lambda) + \log(N) \Rightarrow \log(\mu) - \log(N) = \log(\lambda) \Rightarrow \log(\frac{\mu}{N}) = \log(\lambda) \\ \frac{\mu}{N} &= e^{\beta_1} \end{aligned} \tag{3.3}$$

Onde, $\log(\lambda) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$.

Vamos considerar as cidades Horsens, Kolding e Vejle sendo $M_1,\ M_2,\ M_3$ e, $I_1,\ I_2,\ ...,\ I_5$ os marcadores para os grupos de faixa etária.

$$\log(\frac{\mu}{N}) = -5,6321 - 0,3301M_1 - 0,3715M_2 - 0,2723M_3 + 1,1010I_1... + 1,4197I_5$$

Assim, podemos calcular a taxa média estimada de câncer de pulmão. Por exemplo, para a faixa etária 40 a 54 anos na cidade de Kolding, temos pela equação (3.3):

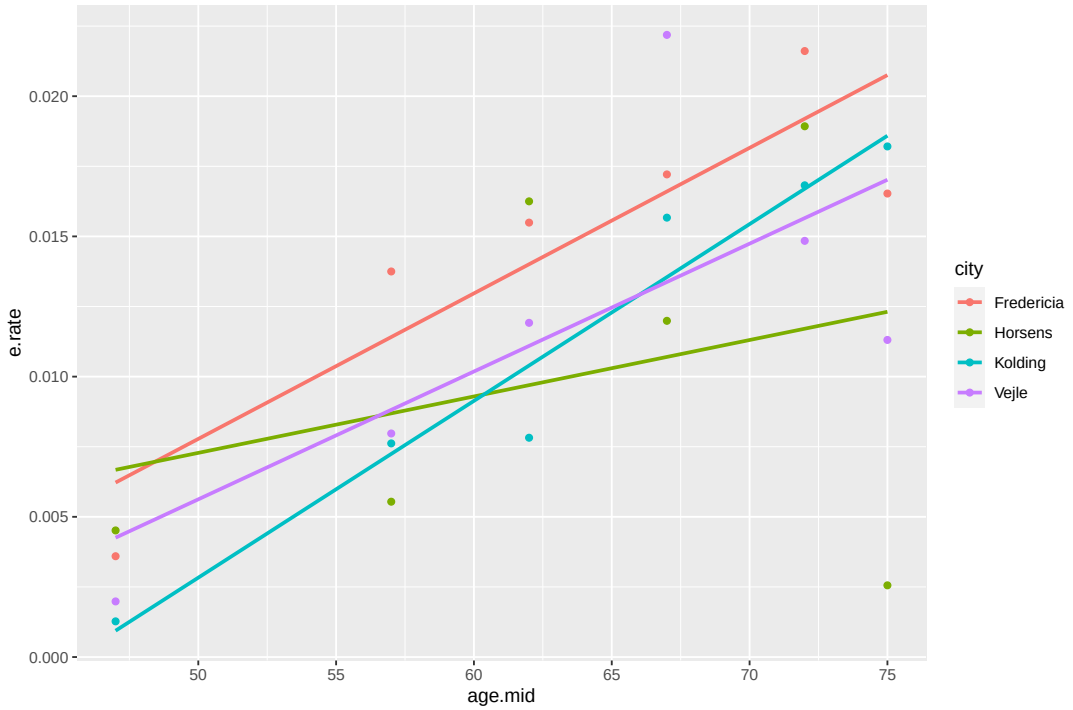
$$e^{-0,3715} \approx 0,6896.$$

Portanto, a taxa média de câncer de pulmão é $0,6896$ vezes menor que a cidade de referência Fredericia. Logo, ela apresenta a redução de aproximadamente $31,04\%$ na incidência de câncer de pulmão.

Levando em consideração as idades como quantitativas, uma vez que foram registradas em seis grupos. Selecionamos os pontos médios de cada faixa etária, assumindo-a como variável quantitativa atribuindo um valor numérico.

O gráfico a seguir, mostra o resultado deste processo de criar uma variável quantitativa para as faixas a partir dos seus pontos médio.

Figura 3.4 – Incidência de Câncer de Pulmão.



Fonte: Próprio autor (2023).

O gráfico, que representa a relação entre a média entre as idades (age.mid) e as taxas de câncer de pulmão (e.rate) para diferentes cidades dinamarquesas, indica uma tendência crescente. Conforme a idade aumenta, há um aparente aumento nas taxas de câncer de pulmão em todas as cidades estudadas. Esta tendência sugere que a idade é um

fator significativo no risco de câncer de pulmão, com riscos maiores observados em faixas etárias mais avançadas.

Além disso, o gráfico parece mostrar variações entre as cidades, indicando um possível "efeito cidade"na incidência de câncer de pulmão. Algumas cidades podem ter taxas mais altas do que outras, o que pode ser devido a uma variedade de fatores ambientais, genéticos, ou de estilo de vida específicos para essas localidades.

Interessantemente, a análise sugere a possibilidade de uma interação entre a cidade e a idade. Esta interação poderia indicar que o aumento da taxa de câncer de pulmão com a idade não é uniforme em todas as cidades, mas pode variar de uma cidade para outra. Em outras palavras, o efeito da idade sobre a taxa de câncer de pulmão pode ser modulado pelo ambiente específico ou características demográficas de cada cidade.

Contudo, é importante ressaltar que a validade e a força dessas tendências e interações são questionáveis, devido ao tamanho relativamente pequeno do conjunto de dados. Com apenas 24 observações (6 para cada uma das 4 cidades), é difícil fazer afirmações categóricas sobre esses efeitos sem uma análise mais robusta e um conjunto de dados mais amplo. Em amostras pequenas, os resultados podem ser facilmente influenciados por variações aleatórias ou por alguns poucos pontos de dados atípicos, e isso limita a capacidade de generalizar as conclusões para uma população maior.

3.3 Modelo de Regressão Poisson em Dados de Câncer de Mama

Para realizar a análise, utilizaremos os dados contidos no conjunto "bcmort", disponível no pacote "ISwR"do ambiente estatístico R. Este conjunto de dados específico aborda as taxas de mortalidade por câncer de mama na Dinamarca e é composto por quatro coortes, sendo uma coorte de estudo e três grupos de Referência.

Os indivíduos foram categorizados com base em seus grupos etários. A escolha desses dados visa explorar e compreender as dinâmicas associadas à mortalidade por câncer de mama, considerando diferentes coortes e faixas etárias.

```
> summary(bcmort3)
age       cohort            bc.deaths       p.yr              lp.yr
50-54:4   Study gr.    :6   Min.   :  9.0   Min.   :  25600   Min.   :10.15
55-59:4   Nat.ctr.     :6   1st Qu.: 51.0   1st Qu.:  86434   1st Qu.:11.37
60-64:4   Hist.ctr.    :6   Median :104.0   Median : 169261   Median :12.04
65-69:4   Hist.nat.ctr.:6   Mean   :213.2   Mean   : 396520   Mean   :12.31
70-74:4                     3rd Qu.:436.2   3rd Qu.: 781897   3rd Qu.:13.57
75-79:4                     Max.   :545.0   Max.   :1067778   Max.   :13.88

age.mid        e.rate
Min.   :52.0   Min.   :0.0001160
1st Qu.:57.0   1st Qu.:0.0004043
Median :64.5   Median :0.0005876
Mean   :64.5   Mean   :0.0006273
3rd Qu.:72.0   3rd Qu.:0.0007569
Max.   :77.0   Max.   :0.0014181
```

Veja que em cada grupo de estudado (Study gr., Nat. ctr., Hist. ctr. Hist. nat. ctr.), temos 6 observações. Com 4 observações por faixa etária.

Preliminarmente, a categoria (bc.deaths), representa o quantitativo de mortes por câncer de mama sendo nossa variável resposta. Observe que 'bc.deaths' varia de $Min.: 9,0$ a $Max.: 545,0$.

A contagem é realiza para cada combinação entre o (age) ou grupo etário e (cohort) ou coorte, grupo que compartilha as mesmas características. Ou seja, (bc.deaths) é uma variável de contagem que mede a frequência de um ocorrência, morte por câncer de mama, para cada grupo distinto definido pela combinação de idade e coorte.

A coluna (p.yr) denota o número de pessoas-ano da população em cada coorte e grupo etário durante a análise. Anos-pessoa representa a uma medida de exposição que

combina o tamanho da população com o tempo. Em nossa análise, será um termo de (offset) no modelo.

Note os valores dos quartis, na categoria (bc.deaths), onde $1stQu. :\leq 51,0$ e $3rdQu. :\leq 436,2$ de casos (mortes). Logo, a distribuição sugere uma grande variabilidade no quantitativo de casos entre os grupos.

```
Call:
glm(formula = bc.deaths ~ cohort + age + offset(lp.yr), family = poisson,
data = bcmort3)

Deviance Residuals:
Min       1Q       Median    3Q       Max
-2.8806  -0.7386  -0.0176   0.7624   3.1646

Coefficients:

                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         -8.85378    0.09374 -94.451  < 2e-16 ***
cohortNat.ctr.       0.08260    0.07013   1.178  0.23887
cohortHist.ctr.      0.22246    0.08229   2.703  0.00686 **
cohortHist.nat.ctr.  0.03625    0.07042   0.515  0.60678
age55-59             1.04262    0.07424  14.044  < 2e-16 ***
age60-64             1.27942    0.07321  17.477  < 2e-16 ***
age65-69             1.39687    0.07294  19.152  < 2e-16 ***
age70-74             1.63950    0.07313  22.420  < 2e-16 ***
age75-79             1.95870    0.08097  24.191  < 2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1039.855  on 23  degrees of freedom
Residual deviance:   46.579  on 15  degrees of freedom
AIC: 223.53

Number of Fisher Scoring iterations: 4
```

O grupo etário "age 50-54"e "Study gr."não foram listada dentro da tabela acima, logo, são estabelecidos como níveis de base para as variáveis categóricas.

O parâmetro de interceptação do modelo (Intercept), com um valor estimado

$-8,85378$ é significativo, uma vez . Este intercepto é o logaritmo da taxa estimada de mortalidade por câncer de mama para o grupo etário de base "age 50-54"na coorte de base "Study gr.".Aplicando a exponenciação ao valor logarítmico,

$$e^{-8,85378} \approx 0,000143$$

Portanto, está é nossa taxa de mortalidade real para faixa etária (50-54) do Study gr.. Ou seja, temos uma pequena proporção de pessoas, na faixa etária, afetadas pela condição (câncer de mama). Vamos aplicar a exponencial em todos os valores:

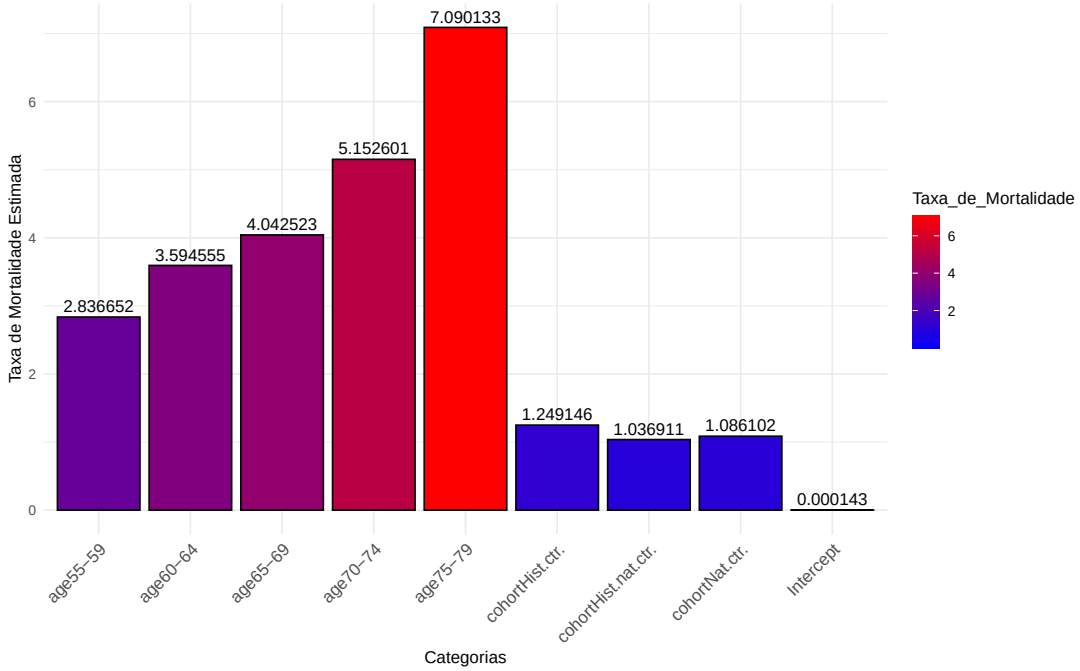
```
> round(exp(coef(model)),6)

(Intercept)      cohortNat.ctr.     cohortHist.ctr. cohortHist.nat.ctr.
0.000143           1.086102           1.249146           1.036911

age55-59           age60-64           age65-69           age70-74
2.836652           3.594555           4.042523           5.152601

age75-79
7.090133
```

Figura 3.5 – Taxa de Mortalidade Estimada por Categoria.



Fonte: Próprio autor (2023).

Com os valores das exponenciais, obtemos as porcentagens de similaridade com o grupo de estudo.

cohort	Porcentagem (%)
CohortNat.ctr.	8.610216
CohortHist.ctr.	24.914624
CohortHist.nat.ctr.	3.691074
age55-59	183.665232
age60-64	259.455468
age65-69	304.252307
age70-74	415.260104
age75-79	609.013299

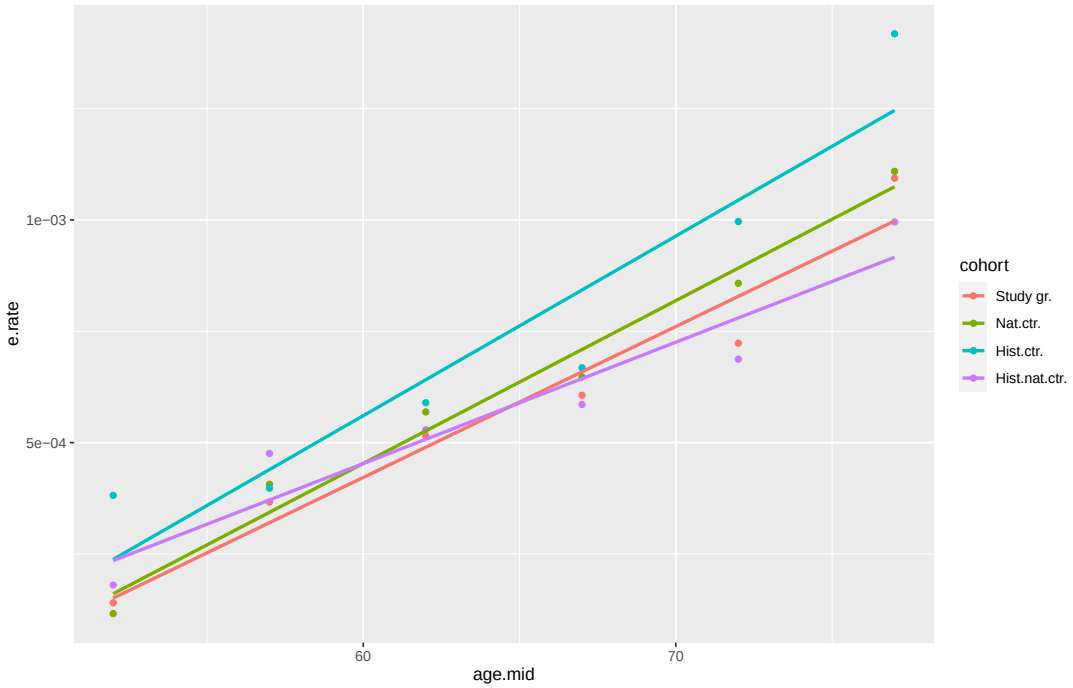
Cada porcentagem representa o aumento da taxa de incidência (câncer de mama) comparado ao grupo de estudo. No caso, "Hist.ctr."possui a taxa de mortalidade é significativamente mais alta, representando o aumento de $\approx 24,91\%$ em relação ao grupo de estudo. Já "Nat.ctr.", a diferença é menor, com um aumento de $\approx 8,61\%$. Por fim, temos "Hist.nat.ctr.", com diferença mínima, relacionada as outras anteriores, com apenas aumento de $\approx 3,69\%$.

Já as faixas etárias são fatores críticos que influenciam diretamente à taxa de mortalidade. Em outras palavras, à medida que as pessoas envelhecem, a taxa de ocorrência aumentam significativamente. Isso implica, as faixas etárias e o risco são diretamente proporcionais.

Vamos destacar um exemplo ilustrativo desta tendência no grupo de estudo. A taxa de mortalidade para a faixa (75-79) é de aproximadamente $7,0901$ vezes maior do que a taxa observada na faixa de controle (50-54). Isso implica, entre as faixas, aumento substancial à medida que os indivíduos envelhecem.

A tendência crescente na taxa de ocorrência é uma descoberta comum em estudos epidemiológicos. Ela reflete o fato de que à medida que envelhecemos, o corpo se torna mais suscetível a doenças.

Figura 3.6 – Incidência de Câncer de Mama.



Fonte: Próprio autor (2023).

O gráfico de dispersão mostra a relação entre a idade média (idade.mid) e a taxa de mortalidade esperada (e.rate), com diferentes cores representando as diferentes categorias. Observamos que as taxas de mortalidade aumentam à medida que a idade aumenta, o que é evidenciado pelo padrão geral de aumento das taxas à medida que nos deslocamos para a direita no gráfico.

Depois de levarmos em consideração todos os fatores discutidos acima, podemos inferir que a regressão Poisson é adequado para a análise. Por conta da modelagem versátil das taxas de mortalidade, ao mesmo tempo em que considera variáveis categóricas e ajustes temporais. Portanto, as estimativas dos coeficientes mostram que a taxa de mortalidade é influenciada pela idade e por cada categoria.

4 Considerações Finais

A análise estatística e probabilística desempenha um papel crucial na interpretação e compreensão de fenômenos complexos, particularmente na medicina e em estudos epidemiológicos. Dados de contagem permitem compreender melhor a extensão de uma doença, avaliar sua agressividade e determinar o melhor curso de tratamento. Além disso, os métodos de estimação permitem realizar inferência sobre a presença da doença e escolher as melhores opções de tratamento.

O estudo detalhou aplicações da distribuição de Poisson na modelagem de dados de contagem, especialmente no contexto do câncer de pulmão e mama. Esta distribuição é ideal para descrever eventos raros ou discretos, sendo amplamente utilizada em diferentes tipos de câncer, cada um com características específicas de incidência e progressão.

A análise de dados reais e simulados demonstrou a eficácia da regressão de Poisson na área da saúde. Exemplos com os dados reais (*eba1977* e *bcmort*), respectivamente, sobre incidência de câncer de pulmão e mortalidade por câncer de mama revelam como este modelo estatístico pode ser empregado para entender melhor as taxas de ocorrência e mortalidade, considerando diferentes variáveis como idade, localização geográfica e coortes de estudo.

O estudo de simulação permitiu a criação de um ambiente controlado para testar a eficácia do modelo de regressão de Poisson. Os dados foram gerados com parâmetros pré-definidos, permitindo uma compreensão clara da relação entre as variáveis independentes e a variável de resposta. Neste cenário controlado, o modelo de regressão Poisson se adequou bem aos dados, com os coeficientes estimados refletindo com precisão as taxas de incidência simuladas. Este resultado reforça a utilidade do modelo de Poisson para analisar dados de contagem, especialmente quando a relação entre as variáveis é conhecida ou pode ser teoricamente prevista.

Os dados reais, por outro lado, apresentaram um cenário mais complexo e imprevisível. Os conjuntos de dados relativos ao câncer de pulmão e câncer de mama na Dinamarca revelaram variações significativas que são típicas de dados epidemiológicos reais. Ao aplicar o modelo de regressão de Poisson a estes dados, observou-se que, embora o modelo pudesse identificar tendências gerais e efeitos significativos de determinadas variáveis (como a idade), a complexidade inerente aos dados reais introduziu um grau de incerteza maior do que na simulação. No entanto, dada a natureza dos dados reais, surge a consideração de outros modelos estatísticos que possam ser adequados, como a distribuição Binomial Negativa, especialmente em situações onde os dados apresentam superdispersão. Ou seja, quando a variância é significativamente maior que a média.

Um ponto crucial na análise de dados reais foi a consideração de fatores como a variabilidade na incidência da doença entre diferentes cidades e grupos etários. Isso demonstrou que, em situações do mundo real, os modelos estatísticos precisam ser robustos o suficiente para acomodar variações e complexidades que podem não ser previstas em simulações.

Referências

CASELLA, G.; BERGER, R. L. *Inferência Estatística*. São Paulo: Editora edups, 2006. v. 2. Citado na página 14.

DALGAARD, P. Introductory statistics with r. CRAN, 2019. Citado 2 vezes nas páginas 11 e 28.

MAGALHÃES, M. N. *Probabilidade e Variáveis Aleatórias*. São Paulo: Cengace Learning, 2010. v. 2. Citado na página 14.

MEYER, P. L. *Probabilidade: aplicações à estatística*. [S.l.]: Livros Técnicos e Científicos Rio de Janeiro, 1983. Citado 3 vezes nas páginas 11, 13 e 23.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. São Paulo: Saraiva Educação SA, 2017. Citado 2 vezes nas páginas 13 e 23.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado 2 vezes nas páginas 11 e 28.

R., M. S. *Probabilidade e estatística*. [S.l.]: McGraw - Hill do Brasil, 1978. Citado na página 13.

ROSS, S. *Probabilidade - Um curso moderno com aplicações*. [S.l.]: Pearson Education, 2010. Citado 2 vezes nas páginas 11 e 18.