Análise de Dados Categóricos
Analisando uma resposta de contagem

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Novembro, 2023

- **4.3 Regressão da taxa Poisson**
- **4.4 Inflação de zeros**

“A maioria dos acidentes automobilísticos ocorre a menos de oito quilômetros de casa”.

Se for verdadeira, implica que as estradas mais perigosas estão perto de sua casa - e perto de nossas casas e perto da casa de todos! Uma explicação que se pode inventar para essa afirmação é que nos tornamos tão familiarizados com as estradas perto de nossas próprias casas que relaxamos demais quando dirigimos perto de casa e, assim, nos expomos a graves perigos.

Uma explicação muito mais simples para esse resultado é que talvez a maior parte do trânsito ocorra a 8 km de casa! Se assim for, então a taxa com que colidimos com outras pessoas pode não ser diferente ou até menor, perto de casa do que em qualquer outro lugar, mas a exposição a acidentes em potencial é muito maior.

Nesse contexto, contar o número de acidentes em diferentes distâncias de casa pode ser uma medida de risco menos significativa do que estimar a taxa de acidentes por milha percorrida.

Isso destaca um ponto importante levantado no início deste capítulo: uma suposição por trás do uso do modelo de Poisson para contagens é que tanto a intensidade ou taxa de ocorrência do evento quanto a oportunidade ou exposição para contagem são constantes para todas as observações disponíveis.

Na Seção 4.2, investigamos modelos que permitem que uma contagem média mude. Implicitamente, estávamos assumindo que a exposição era constante para todas as observações.

Por exemplo, ao contar as bebidas consumidas em um sábado, a duração da observação “sábado” era constante para todos os participantes, de modo que variava apenas a taxa ou a intensidade do consumo de bebida.

Em seguida, permitimos que as contagens médias, ou seja, taxa de consumo variassem usando um modelo de regressão de Poisson para relacioná-las com variáveis explicativas.

Em problemas onde a exposição não é constante entre todas as observações, modelar as contagens diretamente deixa as interpretações confusas.

Por exemplo, se medimos alguns participantes no estudo de álcool por uma hora e outros por um dia e outros por uma semana ou um mês, deveríamos esperar que o número de bebidas consumidas entre os participantes fosse diferente simplesmente por causa da maneira como conduzimos nosso estudo.

Isso pode obscurecer quaisquer efeitos que as variáveis explicativas possam ter. Teríamos, portanto, de desconsiderar o efeito da exposição, para podermos chegar ao cerne da questão que relaciona as variáveis explicativas à taxa de consumo.

Geralmente, “taxa” é definida como contagem média por unidade de exposição, por exemplo, bebidas por dia, animais presos por visita, acidentes por milha percorrida, árvores por hectare, etc..

Ou seja, $R = \mu/t$, onde R é a taxa, t é a exposição e μ é a contagem média ao longo de uma duração de exposição de t. Quando todas as nossas contagens observadas têm a mesma exposição, modelando a contagem média em função das variáveis explicativas $x_1, \cdots, x_p$ é o mesmo que modelar a taxa.

Quando as exposições variam, ainda podemos usar um modelo de regressão de Poisson para as médias, mas precisamos levar em conta a exposição no modelo. Isso é feito usando um modelo de regressão da taxa Poisson escrevendo o modelo como:

$$\log(r_i) = \log(\mu_i/t_i) + \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

ou

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi},$$

$i = 1, \cdots, n$.

Observe que o termo $\log(t_i)$ é conhecido e não é multiplicado por um parâmetro. Ele simplesmente ajusta a média para cada observação diferente para explicar sua exposição.

O termo $\log(t)$ é chamado de offset (deslocamento), portanto, esse modelo às vezes é chamado de regressão de Poisson com deslocamentos.

Uma consequência direta da equação acima é que

$$\mu = t \times \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right),$$

que mostra a influência direta da exposição na contagem média.

Às vezes é útil converter a exposição para outras unidades para criar uma taxa mais interpretável. Por exemplo, se estivermos analisando acidentes e a exposição for dada como milhas percorridas, o número médio de acidentes em 1 milha deve ser extremamente pequeno. Se preferirmos, podemos optar por reescrever a taxa como uma quantidade mais interpretável, como "acidentes por 1.000 milhas", dividindo as milhas de exposição por 1.000 antes de aplicar o modelo.

Dado um conjunto de estimativas $\widehat{\beta}_0, \cdots, \widehat{\beta}_p$, taxas previstas para valores dados $x_1, \cdots, x_p$ são encontrados a partir de

$$\widehat{R} = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p\right).$$

Uma contagem prevista para uma determinada exposição t é encontrada simplesmente multiplicando a taxa pela exposição, $\widehat{\mu} = t \times \widehat{R}$.

Exemplo 4.12: Resposta de postura de besouros à aglomeração.

Muitas populações consistem em subgrupos de diferentes tipos de unidades. Uma situação particularmente comum é que uma população consiste em duas classes: indivíduos “suscetíveis” e “imunes”. Por exemplo, máquinas de ressonância magnética (MRI) são equipamentos médicos grandes e muito caros.

Alguns hospitais, mas não todos, têm uma máquina de ressonância magnética. Se um hospital tem uma, certamente não fica parada por longos períodos de tempo.

Suponha que os administradores do hospital sejam solicitados em uma pesquisa a relatar o número de ressonâncias magnéticas que eles realizam em um mês.

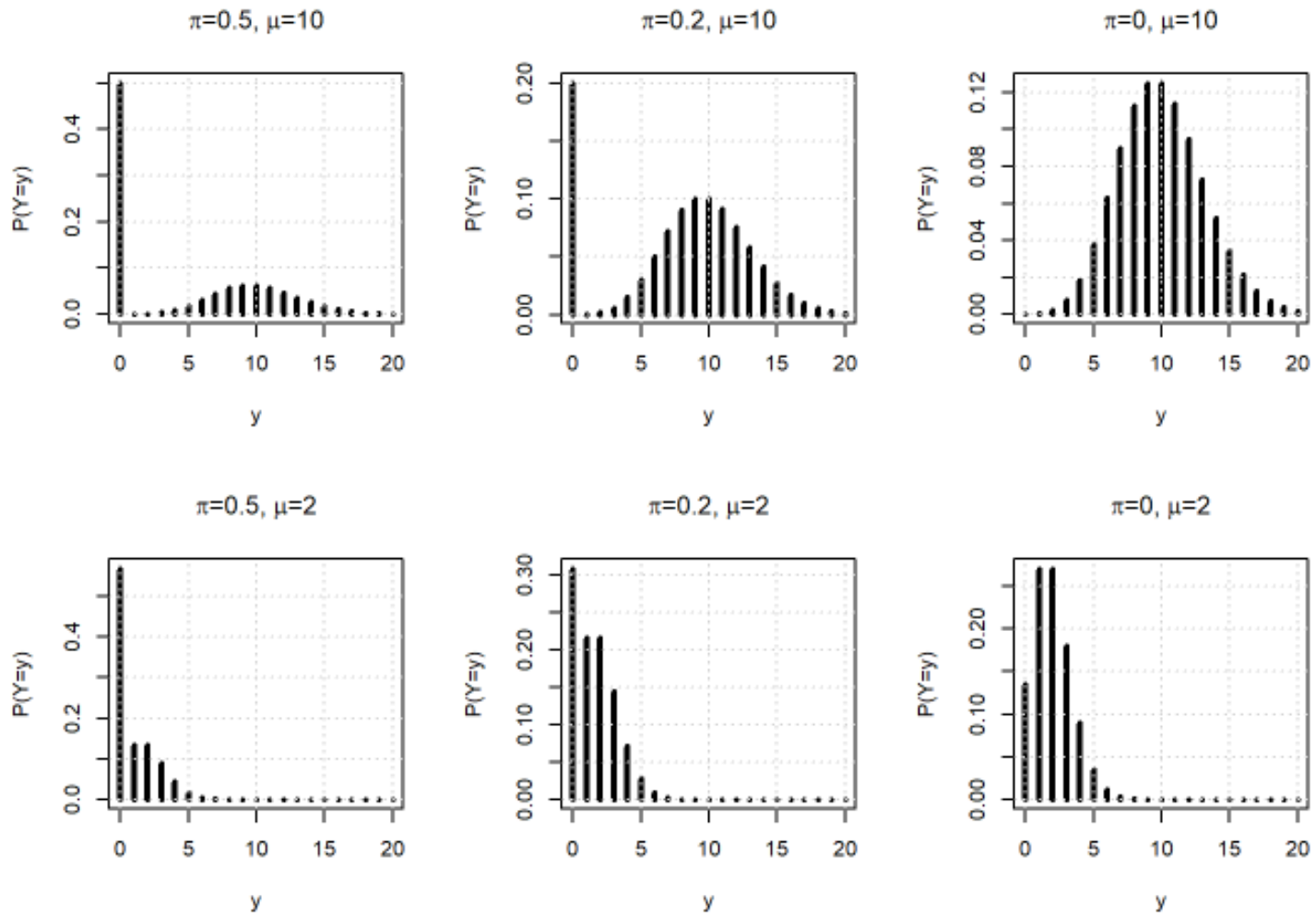
Hospitais sem máquina de ressonância magnética não estão realizando nenhuma ressonância magnética, enquanto o restante relatará alguma contagem diferente de 0. Como um exemplo diferente, suponha que registremos o número de peixes capturados em vários lagos em viagens de pesca de 4 horas em Minnesota.

Alguns lagos em Minnesota são muito rasos para que os peixes sobrevivam ao inverno, então pescar nesses lagos não produzirá capturas. Por outro lado, mesmo em um lago onde os peixes são abundantes, podemos ou não pegar algum peixe devido às condições ou à nossa própria competência. Assim, o número de peixes capturados será zero se o lago não comportar peixes e será zero, um ou mais se comportar.

Essas situações são exemplos em que as respostas vêm de misturas de distribuições e, em particular, são casos em que uma das distribuições colocam toda a sua massa em zero. Ou seja, os "suscetíveis" na população retornam uma contagem de acordo com alguma distribuição, enquanto todos os "imunes" retornam um zero.

Os dados resultantes desse tipo de mistura têm a propriedade de haver mais contagens de zero do que seria previsto por um único modelo distributivo. Isso é conhecido como inflação zero. Em geral, seja π a proporção de imunes na população e as contagens de suscetíveis tenham uma distribuição $Po(\mu)$.

Essa mistura é chamada de modelo de Poisson inflado de zero.


π=0.5, μ=10
π=0.2, μ=10
π=0, μ=10
π=0.5, μ=2
π=0.2, μ=2
π=0, μ=2
P(Y=y)
y

A inflação de zeros é óbvia quando $\mu \gg 0$ como é o caso na linha superior. Dificilmente alguém seria tentado a usar uma única distribuição de Poisson como modelo para dados com um histograma como os mostrados para $\pi > 0$. No entanto, quando a contagem média entre os suscetíveis pode colocar uma massa apreciável em zero ou um, linha inferior, então o a presença de inflação zero pode passar despercebida se a natureza do problema não o sugerir antecipadamente.

A presença de variáveis explanatórias também pode obscurecer a distinção entre as classes, porque a contagem média de susceptíveis e as probabilidades de imunidade podem variar muito entre os membros da população. Na prática, a inflação zero é muitas vezes descoberta apenas ao ajustar um modelo e encontrando o ajuste ruim usando as técnicas descritas no Capítulo 5.

Modelos corretivos

Quando se sabe ou acredita-se que uma população consiste em dois subgrupos, conforme descrito acima, ou quando um excesso de contagens zero aparece inesperadamente, os modelos de Poisson comuns provavelmente produzirão ajustes ruins aos dados. Vários modelos foram propostos para explicar os zeros em excesso. O modelo principal é o modelo de Poisson inflado de zero (ZIP) de Lambert (1992).

O modelo ZIP especifica que a imunidade é uma variável aleatória binária que pode depender de variáveis explanatórias denotadas coletivamente por z, enquanto as contagens entre os suscetíveis seguem uma distribuição de Poisson cuja média pode depender de variáveis explanatórias coletivamente denotadas por x.

Desta forma,

$$\begin{array}{rcll} Y & = & 0 & \text{com probabilidade} \quad \pi(z) \\ Y & \sim & Po(\mu(x)) & \text{com probabilidade} \quad 1-\pi(x) \end{array}$$

Normalmente, um modelo logístico é assumido para $\pi(z)$

$$logit\big(\pi(z)\big) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_r z_r,$$

enquanto o modelo com ligação logaritmo usual é usado para $\mu(x)$

$$\log(\big(\mu(x)\big) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

onde $\gamma_0, \gamma_1, \cdots, \gamma_r$ e $\beta_0, \beta_1, \cdots, \beta_p$ são parâmetros de regressão desconhecidos.

É admissível que $x = z$ de forma que as mesmas variáveis explicativas sejam utilizadas tanto para distinguir as classes quanto para modelar a média dos suscetíveis, mas isso não é obrigatório. Geralmente, x e z podem ser conjuntos de variáveis arbitrariamente semelhantes ou diferentes.

Pode-se mostrar que

1. $P(Y = 0 \mid x, z) = \pi(z) + \big(1 - \pi(z)\big)e^{-\mu(x)}$,
2. $\mathrm{E}(Y \mid x, z) = \mu(x)\big(1 - \pi(z)\big)$, e
3. $\mathrm{Var}(Y \mid x, z) = \mu(x)\big(1 - \pi(z)\big)\big(1 + \mu(x)\pi(z)\big)$.

Assim, a média de Y quando $\pi(z) > 0$ é sempre menor que $\mu(x)$, mas a variância de Y é maior que sua média por um fator de $1 + \mu(x)\pi(z)$.

Um parente do modelo ZIP, chamado de modelo hurdle (Mullahy, 1986), difere em que uma contagem é sempre maior que zero sempre que algum obstáculo real ou conceitual é ultrapassado. O exemplo de ressonância magnética que iniciou esta seção é deste tipo, porque se um hospital gastou a instalação de um aparelho de ressonância magnética, o "obstáculo", certamente o usará.

Neste caso, todos os zeros correspondem a imunes, àquelas unidades que não ultrapassaram o obstáculo de instalação de uma máquina, enquanto o modelo de contagem para suscep-tíveis assume valores começando em 1 em vez de 0.

Isso é feito impondo probabilidade 0 em $Y = 0$ e, em seguida, aumentando todas as outras probabilidades proporcionalmente para que elas mais uma vez somem 1. A distribuição resultante é chamada de "Poisson truncado à esquerda", porque a cauda esquerda da distribuição ($Y = 0$) é cortado da distribuição normal de Poisson.

O modelo logístico para a probabilidade de imune e o logaritmo da média da distribuição de Poisson de contagens truncada à esquerda para susceptíveis são usados no modelo hurdle como no modelo ZIP.

Ambos os modelos podem ser estimados por máxima verossimilhança usando funções do pacote **pscl**. A sintaxe para o ajuste de ambos os modelos é muito semelhante e é descrita em detalhes por Zeileis et al. (2008).

O ajuste do modelo pode ser mais fácil com o modelo hurdle, porque fica imediatamente claro que todos os zeros são imunes. Então a modelagem da função média é baseada no subconjunto de dados com $Y > 0$ e é independente da probabilidade de imunes. No entanto, o modelo ZIP parece ser o preferido na maioria das áreas. Nossa preferência geralmente é determinada por qual modelo para os imunes concorda melhor com a fonte dos dados.

Se uma contagem zero define um imune, como no exemplo da ressonância magnética, um obstáculo faz mais sentido. Se os suscetíveis às vezes podem fornecer contagens zero, como no exemplo da pesca, isso aponta para um modelo ZIP. Se não estiver claro se os suscetíveis podem produzir contagens zero, então considerações empíricas, como medidas de ajuste, podem prevalecer. Cconsulte o Capítulo 5 para obter detalhes.

Exemplo 4.13: Resposta de postura de besouro à aglomeração.

Exemplo 4.14: Simulação comparando os modelos ZIP e Poisson.