Análise de Dados Categóricos
Analisando uma resposta de contagem
Seção 4.1

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Novembro, 2023

- **4.1 Modelo de Poisson para dados de contagem**
- **4.1.1 Distribuição de Poisson**
- **4.1.2 Verossimilhança e inferência de Poisson**
- **4.2 Modelos de regressão de Poisson para respostas de contagem**
- **4.2.1 Modelo para média: Ligação logarítmica**
- **4.2.2 Estimação e inferência de parâmetros**

Considere uma situação em que algum tipo de evento ocorre durante um período fixo de tempo, como carros passando por uma ponte. É fácil imaginar que a velocidade ou intensidade com que os carros cruzam a ponte pode variar de acordo com a hora do dia; por exemplo, mais carros na hora do rush, menos às 3 da manhã ou dia da semana; mais em um dia de trabalho, menos em fins de semana e feriados.

Suponha contemos carros por exatamente uma hora no mesmo horário todas as semanas, digamos quarta-feira, das 7h às 8h. Mesmo que não haja diferenças práticas na intensidade populacional de semana para semana, não esperaríamos que exatamente o mesmo número de carros atravessasse a ponte a cada semana.

Algumas pessoas que normalmente cruzam essa ponte durante esse horário podem pegar carona ou pegar o trânsito; podem estar doentes, de férias, em reunião ou em teletrabalho; ou eles podem simplesmente chegar mais cedo ou mais tarde do que o normal e perder os cortes para a hora da contagem. Assim, esperamos que a contagem real de carros durante esse período varie aleatoriamente, mesmo que a taxa populacional subjacente ao processo não esteja mudando.

Essa é a natureza da distribuição de Poisson para contagens. Se pudermos assumir que:

(1) todas as contagens realizadas em algum processo têm a mesma intensidade subjacente e

(2) o período de observação é constante para cada contagem,

então as contagens seguem uma distribuição de Poisson.

Se Y é uma variável aleatória de Poisson, então a função de probabilidade (PMF) de Y é

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \qquad y = 0, 1, 2, \cdots,$$

onde $\mu > 0$ é um parâmetro. Podemos abreviar essa distribuição escrevendo $Y \sim Po(\mu)$.

Adaptações deste modelo serão dadas posteriormente para casos onde a intensidade não é constante para todas as contagens, Seção 4.2 e para quando o período de observação não é constante para todas as contagens, Seção 4.3.

Propriedades da distribuição de Poisson

A distribuição de Poisson tem várias propriedades atraentes que a tornam um modelo conveniente para se trabalhar. Em primeiro lugar, a forma da distribuição é bastante simples. Existe apenas um parâmetro μ, que representa a média e a variância da distribuição:

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu.$$

Observe que, ao contrário da distribuição normal, a variância da distribuição de Poisson muda conforme a média muda. Isso faz sentido porque as contagens são limitadas abaixo por 0. Por exemplo, suponha que as médias para dois grupos diferentes de contagens sejam 5 e 50. Então as contagens no primeiro grupo podem estar principalmente na faixa de 0 a 10, enquanto as do segundo grupo pode variar naturalmente de, digamos, 20-80.

Outra propriedade útil é que somas de variáveis aleatórias de Poisson também são variáveis aleatórias de Poisson:

Se $Y_1, Y_2, \cdots, Y_m$ são independentes com $Y_k \sim Po(\mu_k)$, então

$$\sum_{k=1}^{m} Y_k \sim Po\Big(\sum_{k=1}^{m} \mu_k\Big).$$

Assim, os totais de várias contagens podem ser modelados com distribuições de Poisson, desde que as contagens constituintes possam.

A forma da distribuição de Poisson se aproxima da distribuição normal à medida que μ cresce. De fato, a distribuição normal às vezes é usada como modelo para contagens, embora não seja uma distribuição discreta.

Exemplo 4.1: carros em um cruzamento.

Verossimilhança e inferência de Poisson

O parâmetro μ no modelo de Poisson é estimado por máxima verossimilhança. Dada uma amostra aleatória $y_1, \cdots, y_n$; de uma distribuição de Poisson com média μ, a função de verossimilhança é

$$L(\mu; y_1, \cdots, y_n) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}.$$

O estimador de máxima verossimilhança (MLE) é a média amostral, $\widehat{\mu} = \overline{y}$. Usando a segunda derivada da log-verossimilhança, a variância de $\widehat{\mu}$ é facilmente mostrada como estimada por

$$\widehat{\text{Var}}(\widehat{\mu}) = \widehat{\mu}/n.$$

Isso difere de nossa estimativa usual para a variância de $\overline{y}$, s^2/n, porque a média e a variância são as mesmas na distribuição de Poisson. Em grandes amostras de uma distribuição de Poisson, a média e a variância da amostra serão quase idênticas.

Assim como ocorre com o parâmetro da probabilidade de sucesso em uma distribuição binomial, há muitas maneiras de formar intervalos de confiança para o parâmetro de média da Poisson.

Discutimos aqui Wald, razão de verossimilhança, escore e métodos exatos para desenvolver um intervalo de confiança para μ.

Em particular, o intervalo de confiança Wald de $100(1-\alpha)\%$ é

$$\widehat{\mu} \pm Z_{1-\alpha/2}\sqrt{\widehat{\mu}/n}.$$

O intervalo escore é derivado dos testes de hipótese de $H_0 : \mu = \mu_0$, que podem ser conduzidos usando a estatística escore

$$Z_0 = \frac{\widehat{\mu} - \mu_0}{\sqrt{\widehat{\mu}/n}},$$

comparando Z_0 com quantis apropriados da distribuição normal padrão.

Invertendo os resultados deste teste encontramos o intervalo de confiança,

$$\left(\widehat{\mu} + \frac{Z^2_{1-\alpha/2}}{2n}\right) \pm Z_{1-\alpha/2}\sqrt{\frac{\widehat{\mu} + Z^2_{1-\alpha/2}/4n}{n}}.$$

O intervalo da razão de verossimilhanças (LR) é o conjunto de valores de μ_0 para o qual

$$-2\log\left(\frac{L(\mu_0 \mid y_1, \cdots, y_n)}{L(\widehat{\mu} \mid y_1, \cdots, y_n)}\right) \leq \chi^2_{1,1-\alpha}.$$

Não tem uma solução fechada e deve ser encontrado usando procedimentos numéricos iterativos. Portanto, geralmente não é usado para esse problema simples. O intervalo exato é desenvolvido usando uma lógica semelhante à que foi usada para o intervalo de Clopper-Pearson para a probabilidade binomial da Seção 1.1.2. Sua forma é

$$\chi^2_{2n\widehat{\mu},\alpha/2}/2n < \mu < \chi^2_{2(n\widehat{\mu}+1),1-\alpha/2}/2n.$$

Alguns desses métodos são muito semelhantes aos testes e intervalos de confiança usados para inferência sobre as médias da distribuição normal.

A principal vantagem desses métodos baseados na Poisson é o uso da função de verossimilhança de Poisson para obter a relação $\text{Var}(\overline{Y}) = \mu/n$, o que permite intervalos de confiança mais curtos e testes mais poderosos, especialmente em amostras menores.

A utilização da função de verossimilhança de Poisson também pode ser sua desvantagem, pois é comum que os processos que geram contagens tenham intensidade que não permaneça rigidamente constante.

Isso faz com que as contagens se comportem como se tivessem mais variabilidade do que o modelo de Poisson espera. Isso é chamado de superdispersão e é abordado na Seção 5.3.

Portanto, $\widehat{\text{Var}}(\widehat{\mu}) = \widehat{\mu}/n$ geralmente subestima a verdadeira variabilidade da contagem média. Embora s^2 seja uma estimativa menos precisa de Var(Y) quando a distribuição de Poisson é exata, ela estima a variabilidade presente devido a todas as fontes, incluindo quaisquer variações na intensidade do processo. Portanto, *t*-testes comuns e intervalos de confiança para uma média populacional são mais robustos a desvios da suposição de Poisson e às vezes são usados com dados de contagem extraídos de uma única população. No entanto, observe que os intervalos de confiança desenvolvidos especificamente para a distribuição de Poisson podem ser usados mesmo quando $n = 1$, enquanto o intervalo baseado na disribuição *t* não pode.

Exemplo 4.2: carros em um cruzamento.

Nossas recomendações sobre quais intervalos usar para o parâmetro médio da distribuição de Poisson refletem àquelas para o parâmetro de probabilidade de sucesso do binomial.

O intervalo de Wald não é bom, porque seu verdadeiro nível de confiança pode ser muito baixo e raramente atinge seu nível declarado. Por outro lado, o intervalo exato pode ser excessivamente conservador e, portanto, mais amplo do que o necessário. O intervalo escore é um bom compromisso, pois seu verdadeiro nível de confiança é geralmente melhor do que o Wald e geralmente é mais curto do que o intervalo exato.

As diferenças entre os vários intervalos geralmente são pequenas, no entanto, como mostra o próximo exemplo. Para mais detalhes sobre diferentes intervalos de confiança possíveis e suas propriedades, consulte Swift (2009).

Exemplo 4.3: Comparação dos intervalos de confiança para a média de Poisson.

O intervalo escore pode ser recomendado para uso geral, desde que se espere que a distribuição de Poisson se mantenha.

Modelos de regressão para dados de contagem também foram desenvolvidos com base na distribuição de Poisson.

Um tipo especial de modelo de regressão de Poisson, chamado de modelo loglinear, não apenas replica a análise clássica de tabelas de contingência descritas nas Seções 1.2.3, 2.2.6 e 3.2, mas também pode estender essas análises para qualquer número de variáveis e pode acomodar tanto variáveis contínuas e ordinais.

Os modelos de regressão de Poisson podem até ser usados para replicar análises de regressão logística, embora seu uso dessa maneira seja um tanto complicado.

As etapas de modelagem com uma regressão de Poisson são praticamente as mesmas realizadas em qualquer outro processo de modelagem de regressão. Essas etapas incluem

(1) especificar o modelo, ou seja, a distribuição, os parâmetros a serem modelados; por exemplo, probabilidades ou médias e sua relação com as variáveis explicativas;

(2) selecionar variáveis explicativas;

(3) estimar os parâmetros do modelo;

(4) avaliar o ajuste do modelo e

(5) realizar inferências nos parâmetros do modelo e outras quantidades de interesse.

4.2.1 Modelo para média: Ligação logarítmica

Temos n observações de uma variável aleatória de resposta Y e $p \geq 1$ variáveis explanatórias fixas, $x_1, \cdots, x_p$. Assumimos que para cada observação $i = 1, \cdots, n$,

$$Y_i \sim Po(\mu_i),$$

onde

$$\mu_i = exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right).$$

Assim, nosso modelo linear generalizado tem um componente aleatório de Poisson, um componente sistemático linear $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ e uma ligação logarítmica,

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Cada uma dessas três especificações é uma suposição que pode ser questionada. Por exemplo, a distribuição de Poisson pode fornecer um ajuste ruim aos dados. Existem diferentes distribuições que podem servir como componentes aleatórios para dados de contagem neste caso.

Além disso, o preditor linear poderia ser substituído por algo que relacionasse as variáveis explicativas aos parâmetros de maneira não linear, ao custo da complexidade interpretativa e computacional.

Por fim, a função de ligação pode assumir uma forma alternativa. Em particular, a ligação identidade $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, pode parecer uma forma mais simples de usar. No entanto, a distribuição de Poisson exige que $\mu > 0$ e a ligação identidade pode levar a um valor não positivo de μ para determinados valores das variáveis explicativas.

A função de ligação logaritmo garante $\mu > 0$, por isso é quase universalmente usada.

Uma consequência do uso da função de ligação logaritmica com um preditor linear é que as variáveis explicativas afetam a média da resposta multiplicativamente.

Considere um modelo de regressão de Poisson com uma variável explicativa: $\mu(x) = \exp\left(\beta_0 + \beta_1 x\right)$, onde nossa notação enfatiza que μ muda em função de x. Quando aumentamos a variável explicativa em c unidades, o resultado é

$$\mu(x + c) = \exp\left(\beta_0 + \beta_1(x + c)\right) = \mu(x)\exp\left(c\beta_1\right).$$

Assim, a razão das médias em $x + c$ e em x é

$$\frac{\mu(x+c)}{\mu(x)} = \frac{\exp\left(\beta_0 + \beta_1(x+c)\right)}{\exp\left(\beta_0 + \beta_1 x\right)} = \exp\left(c\beta_1\right).$$

Isso leva a uma maneira conveniente de interpretar o efeito de x: A variação percentual na resposta média que resulta de uma alteração de c unidades em x é $PC = 100(e^{c\beta_1} - 1)\%$.

Se houvesse variáveis adicionais no modelo, a mesma interpretação se aplicaria, desde que mantivéssemos constantes as variáveis explicativas adicionais. Quando há termos de interação ou transformações envolvendo a variável explicativa de interesse, a razão de médias é mais complicada, mas pode ser derivada de maneira semelhante à mostrada para razões de chances na Seção 2.2.5.

4.2.2 Estimação e inferência de parâmetros

O modelo de regressão Poisson assume que as observações $y_1, y_2, \cdots, y_n$ são independentes. Por exemplo, eles não são serialmente correlacionados nem formam clusters e, portanto, a verossimilhança é formada pelo produto das funções de probabilidade individuais.

Seguindo as mesmas etapas apresentadas anteriormente, isso leva à log-verossimilhança

$$\begin{aligned}\log\left(L(\beta_0, \cdots, \beta_p \,|\, y_1, \cdots, y_n)\right) \\ \sum_{i=1}^{n} -\exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \\ + y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) - \log(y_i!).\end{aligned}$$

Como de costume, diferenciamos a log-verossimilhança em relação a cada parâmetro β_j, definimos cada uma das $p+1$ equações resultantes iguais a 0 e resolvemos o sistema de equações para encontrar os MLEs $\widehat{\beta}_0; \widehat{\beta}_1, \cdots, \widehat{\beta}_p$.

Como foi o caso da regressão logística na Seção 2.2.1, essas equações normalmente não fornecem soluções de forma fechada; portanto, as soluções devem ser encontradas usando procedimentos numéricos iterativos.

Uma vez que temos as estimativas MLEs para os coeficientes de regressão, podemos calcular MLEs para qualquer função dos coeficientes tomando a mesma função dos MLEs.

Por exemplo, o valor ajustado ou previsão é

$$\widehat{\mu}_i = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}\right),$$

e a alteração estimada na média para uma alteração de c unidades em x_j é

$$\widehat{PC} = e^{c\,\widehat{\beta}_j},$$

mantendo as outras variáveis explicativas constantes.

Inferência

As abordagens padrão para inferência na estimativa de máxima verossimilhança são usadas aqui.

Embora os métodos de Wald sejam relativamente fáceis de executar, eles nem sempre funcionam bem. Métodos de razão de verossimilhança são escolhas melhores quando as rotinas computacionais estão disponíveis.

Exemplo 4.5: Consumo de álcool.