ME720 - Modelos Lineares Generalizados

Parte 13 - Modelos para Dados de Contagem

Profa. **Larissa Avila Matos**

Modelos para Dados de Contagem

Muitas variáveis de resposta tem contagens como possíveis resultados. Exemplos:

1. número de bebidas alcoólicas que você tomou na semana anterior;
2. número de dispositivos que você possui que podem acessar a Internet (laptops, telefones celulares inteligentes, tablets, etc.).

Essas contagens também ocorrem nas entradas das caselas nas tabelas de contingência que classificam as variáveis categóricas. Iremos ver modelos lineares generalizados (MLGs) para variáveis de resposta de contagens, como

1. modelos que assumem uma distribuição Poisson para a resposta. Esse modelo pode ser adaptado para modelar uma taxa quando a contagem é baseada em um índice, como espaço ou tempo;
2. modelos que assumem uma distribuição binomial negativa para a resposta; e
3. modelos que lidam com excesso de zeros na variável de resposta.

MLGs de Poisson para dados de contagem e taxas

A distribuição mais simples para dados de contagem, colocando sua massa no conjunto de valores inteiros não negativos, é a distribuição de Poisson. Suas probabilidades dependem de um único parâmetro, a média $\mu > 0$.

Vimos que, se Y é uma v.a. com distribuição de Poisson ($Y \sim P(\mu_i)$), a f.d.p. é dada por

$$f(\mathrm{y};\mu) = \frac{e^{-\mu}\mu^{\mathrm{y}}}{\mathrm{y}!} = \exp\left[\mathrm{y}\log(\mu) - \mu - \log(\mathrm{y}!)\right], \quad \mathrm{y} = 0, 1, 2, \ldots,$$

e $Y \in$ F.E., com $\mathbb{E}(Y) = \mathrm{Var}(Y) = \mu$.

Além disso, a distribuição de Poisson é uma distribuição unimodal e sua assimetria é descrita por

$$\frac{\mathbb{E}\left[(Y-\mu)^3\right]}{\sigma^3} = \frac{1}{\sqrt{\mu}}.$$

A medida que μ aumenta, a distribuição de Poisson é menos assimétrica e se aproxima de uma distribuição Normal, sendo uma aproximação razoavelmente boa quando $\mu > 10$.

A distribuição de Poisson é frequentemente usada para contagens de eventos que ocorrem aleatoriamente ao longo do tempo ou no espaço a uma taxa específica, quando os resultados em períodos ou regiões disjuntos são independentes.

Por exemplo, um fabricante de telefones celulares pode indicar que a distribuição de Poisson descreve razoavelmente bem o número de reclamações de garantia recebidas a cada semana.

A distribuição de Poisson também pode ser utilizada para a aproximação de uma distribuição Binomial quando o número de ensaios n é grande e π é muito pequeno, com $n\pi = \mu$.

Para a Binomial, se $n \to \infty$ e $\pi \to 0$ tal que $n\pi = \mu$ é fixo, a distribuição Binomial converge para a Poisson.

Demonstração

Por exemplo, se um fabricante tiver vendido 5.000 celulares de um tipo específico e cada um independentemente tiver probabilidade 0,001 de ter uma reivindicação de garantia em uma determinada semana, o número de reclamações por semana terá aproximadamente uma distribuição de Poisson com uma média de $5000(0,001) = 5$.

Modelo Poisson Log Linear

Modelo Poisson Log Linear

Assumindo que $Y_1, \cdots, Y_n$ são v.a.'s independentes, com $Y_i \sim \text{Poisson}(\mu_i)$.

Então, como já vimos, o modelo Poisson Log Linear, é dado por:

Modelo Poisson Log Linear:

$$\log(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n;$$

ou

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \text{ (na forma matricial).}$$

Ajustando um modelo Poisson Log Linear

Para $\eta_i = \log(\mu_i)$ e $\frac{\partial \mu_i}{\partial \eta_i} = \mu_i$, as equações de log-verossimilhança são dadas por

$$\sum_{i=1}^{n} (\mathrm{y}_i - \mu_i) x_{ij} = 0,$$

com visto anteriormente.

Para um modelo log-linear de Poisson vimos também que

$$\mu_i = \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = \left(e^{\beta_p}\right)^{x_{ip}} \ldots \left(e^{\beta_1}\right)^{x_{i1}}.$$

O aumento de uma unidade em x_{ij} tem o impacto multiplicativo de e^{β_j}, ou seja, a média em $x_{ij} + 1$ e igual a média em x_{ij} multiplicado por e^{β_j}, fixando as outras covariáveis.

A matriz da informação é dada por

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}; \mathrm{y})}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} x_{ij} x_{ik} \mu_i$$

Então, temos que a matriz Hessiana é definida-negativa, portanto a função de log-verossimilhança é côncava e possui um máximo único.

As matrizes de informação observada e esperada são idênticas. Portanto, o método Newton-Raphson é equivalente ao Escore de Fisher, uma conseqüência de usarmos a função de ligação canônica.

Além disso para o modelo log-linear de Poisson a matriz de covariância assintótica de $\widehat{\boldsymbol{\beta}}$ é dada por

$$\widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1},$$

onde $\boldsymbol{W}$ é uma matriz diagonal com elementos $w_{ii} = \frac{(\partial \mu_i / \partial \eta_i)^2}{\mathrm{Var}(Y_i)} = \mu_i$.

Desvio

O desvio para modelo Poisson Log Linear é

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left[\mathrm{y}_i \log\left(\frac{\mathrm{y}_i}{\widehat{\mu}_i}\right) - \mathrm{y}_i + \widehat{\mu}_i\right].$$

Se o modelo tem intercepto, temos que $\sum_{i=1}^{n}\widehat{\mu}_i = \sum_{i=1}^{n}\mathrm{y}_i$, então o desvio é dado por

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2\sum_{i}\mathrm{y}_i \log\left(\frac{\mathrm{y}_i}{\widehat{\mu}_i}\right).$$

Além disso temos que $\mathrm{Var}(\mathrm{y}_i) = V(\mu_i) = \mu_i$, com $\phi = 1$, a estatística de escore para comparar o modelo escolhido com o modelo saturado é

$$X^2 = \sum_{i=1}^{n}\frac{(\mathrm{y}_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} \quad \text{(Estatística de Pearson)}.$$

Referência

- Paula, G.A. (2013). Modelos de Regressão com Apoio Computacional.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.
- Faraway, J. J. (2006). *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapaman and Hall/CRC.
- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer