CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

11 de julho, 2018

Aula 13 - Regressão para dados de contagens

Introdução

- Interesse em modelar a distribuição de uma variável referente a algum tipo de contagem em função de covariáveis. Normalmente, tem-se uma contagem de eventos em unidades de tempo ou espaço.

- O modelo de regressão mais comum nessas situações assume distribuição de Poisson com função de ligação logarítmica (canônica).

- Como para a distribuição de Poisson temos $\mu > 0$, a ligação logarítmica garantirá valores positivos para μ quaisquer que sejam os valores das covariáveis.

Exemplos de motivação

- Modelagem do número de sinistros de automóveis em função de características do motorista e do veículo;
- Número de casos de dengue em diferentes quadras de um município em função de variáveis geográficas e demográficas;
- Número de gols marcados por times de futebol nas partidas de um campeonato em função de variáveis referentes ao desempenho em campeonatos anteriores, ao investimento em jogadores, aos valores dos patrocínios, . . . ;
- Abundância de certa espécie animal em quadrantes de uma floresta em função de variáveis ambientais e climáticas.

Distribuição de Poisson

O modelo de Poisson configura uma distribuição de probabilidades discreta, obtida:

- Como aproximação à distribuição binomial quando $n \to \infty$ e $p \to 0$ de tal forma que np permaneça constante;
- Quando os eventos sob contagem ocorrem aleatoriamente ao longo do tempo (ou espaço), com probabilidade de ocorrência proporcional ao tamanho do intervalo e independente das contagens em outros intervalos (Processo de Poisson);
- Quando os tempos entre ocorrências de eventos são independentes e identicamente distribuídos segundo uma distribuição exponencial;
- A distribuição Poisson se caracteriza pela equidispersão ($E(y) = Var(y)$).

Processos de contagens

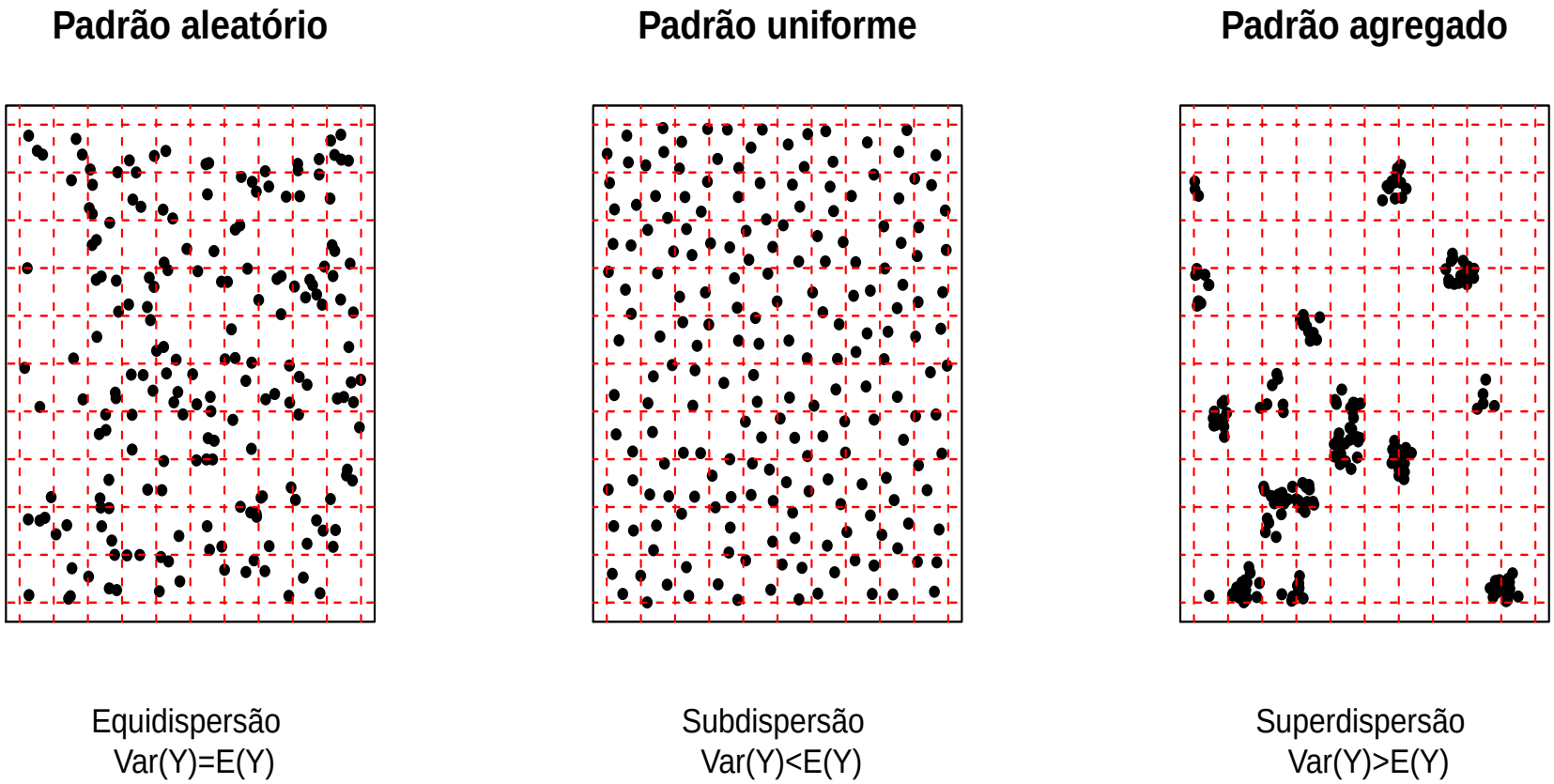

Figura 1: Ilustração de processos de contagens com diferentes padrões de dispersão.

Modelo de Poisson

- Uma variável aleatória Y segue o modelo de Poisson se sua função de probabilidades for dada por:

$$f(y;\mu) = P(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, ...; \quad \mu > 0. \tag{1}$$

- Como ressaltado anteriormente, como propriedade da distribuição Poisson temos:

$$E(y) = Var(y) = \mu. \tag{2}$$

- A distribuição de Poisson pertence à família exponencial de distribuições, tendo função de variância $V(\mu) = \mu$ e parâmetro de dispersão $\phi = 1$.

- Adicionalmente, a medida que μ aumenta, a distribuição Poisson torna-se mais simétrica, aproximando-se da distribuição Normal quando $\mu \to \infty$.

Modelo de Poisson

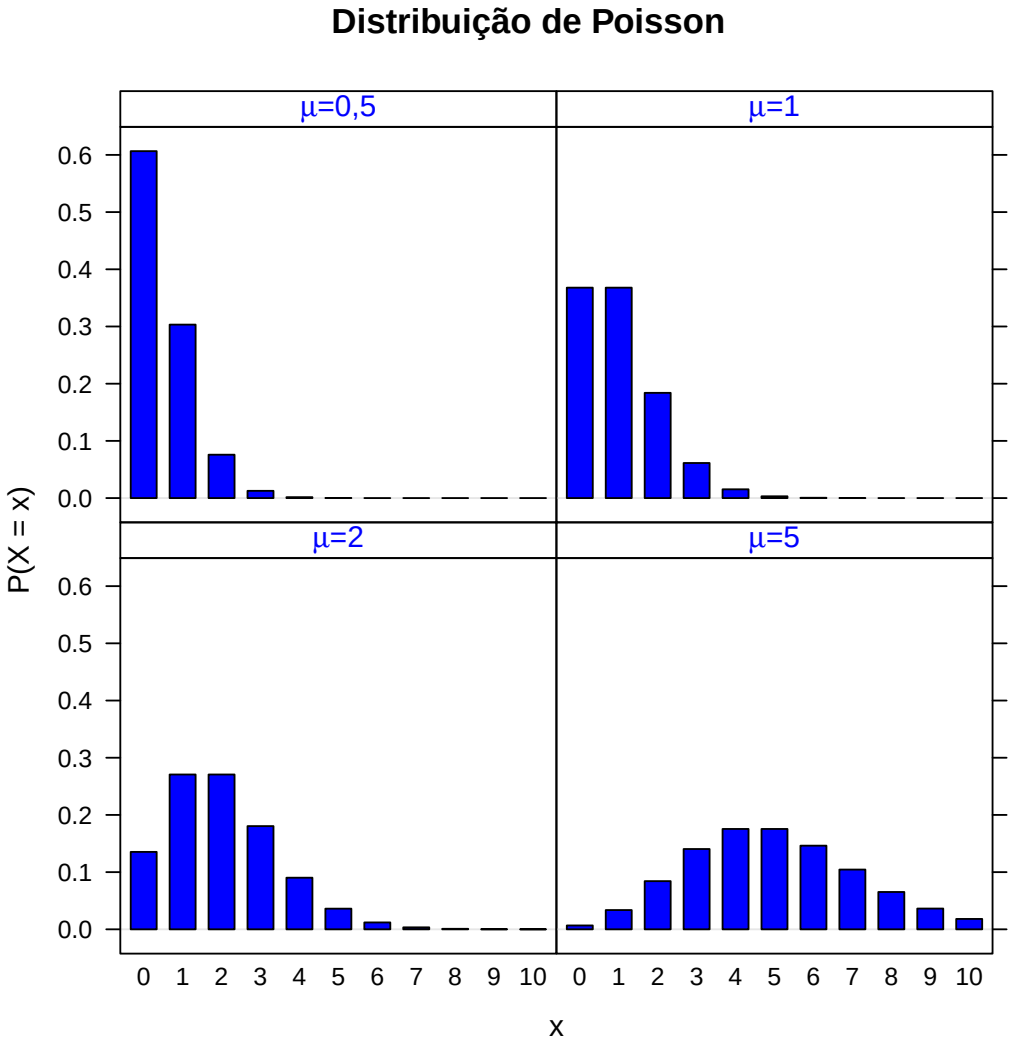

Figura 2: Distribuição de Poisson para diferentes valores de μ

Modelo de regressão log-linear

- No contexto de modelos lineares generalizados, vamos considerar $y_1, y_2, ..., y_n$ variáveis aleatórias independentes, com $y_i|\boldsymbol{x}_i \sim Poisson(\mu_i)$, $i = 1, 2, ..., n$.

- Adicionalmente, considere $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ os vetores de covariáveis correspondentes a cada observação na amostra.

- Considerando função de ligação logarítmica, o modelo log-linear de Poisson fica definido por:

$$y_i|\boldsymbol{x}_i \sim Poisson(\mu_i)$$
$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}.$$

Modelo de regressão log-linear

- Para o modelo log-linear de Poisson, as equações de estimação, baseadas na maximização da (log) verossimilhança, ficam dadas por:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 0, 1, ..., p \quad (\textit{verificar}!) \tag{3}$$

- O modelo log-linear fica expresso na escala da resposta (média) por:

$$\mu_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}}. \tag{4}$$

- Verifica-se facilmente que os efeitos, para o modelo log-linear, são multiplicativos. Basta observar que:

$$\mu_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}} = e^{\beta_0}(e^{\beta_1})^{x_{i1}}(e^{\beta_2})^{x_{i2}}...(e^{\beta_p})^{x_{ip}}. \tag{5}$$

Modelo log-linear - interpretação dos parâmetros

- Considere o modelo log-linear com apenas uma covariável (x). Assumindo que x seja uma variável numérica, então:

$$\frac{\mu_{x+1}}{\mu_x} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}, \tag{6}$$

- Assim, o aumento de uma unidade em x tem efeito multiplicativo na média igual a e^{β_1}.

Modelo log-linear - interpretação dos parâmetros

- Para um aumento de k unidades em x, verifica-se facilmente que o efeito multiplicativo na média fica dado por $e^{k\beta_1}$.

- Para o caso de uma covariável dicotômica (com categorias A e B), inserida no modelo por meio de uma variável indicadora de B, temos:

$$\frac{\mu_B}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 1}}{e^{\beta_0+\beta_1\times 0}} = e^{\beta_1}. \tag{7}$$

- Assim, e^{β_1} corresponde à razão das médias para as categorias B e A (efeito multiplicativo de B).

Modelo log-linear - interpretação dos parâmetros

- Se houvesse uma terceira categoria (C), então seriam necessárias duas variáveis indicadoras (x_1, indicadora de B; x_2, indicadora de C). Assim, teríamos:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{8}$$

- As razões de médias ficariam dadas por:

$$\frac{\mu_B}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_1}; \tag{9}$$

$$\frac{\mu_C}{\mu_A} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 0+\beta_2\times 0}} = e^{\beta_2}; \tag{10}$$

$$\frac{\mu_C}{\mu_B} = \frac{e^{\beta_0+\beta_1\times 0+\beta_2\times 1}}{e^{\beta_0+\beta_1\times 1+\beta_2\times 0}} = e^{\beta_2-\beta_1}. \tag{11}$$

Modelo log-linear - interpretação dos parâmetros

- Caso o preditor linear contenha múltiplas variáveis, as interpretações são idênticas, devendo-se ressaltar, no entanto, que a interpretação do efeito para uma particular variável é válida fixando os valores das demais variáveis.

Modelagem de taxas

- Na análise de dados de contagens, é comum que os indivíduos na amostra apresentem diferentes *níveis de exposição* (pacientes acompanhados por períodos de tempo distintos; contagens de peixes em trechos de um rio com diferentes volumes de água, ...);

- Em situações desse tipo, é necessário incorporar o nível de exposição ao modelo, de maneira a modelar a taxa de ocorrência por *unidade de exposição*.

- Seja y_i a contagem correspondente ao indivíduo i, com exposição t_i (ex: y_i : número de animais de certa espécie em $t_i = 100m^2$ de uma reserva).

- Neste caso, temos uma taxa de y_i/t_i de animais por *unidade de área*, com valor esperado $\lambda_i = \mu_i/t_i$.

Modelagem de taxas

- O modelo log-linear, aplicado à modelagem de taxas, fica dado por:

$$log(\lambda_i) = log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \quad (12)$$

tal que:

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + log(t_i). \quad (13)$$

- Neste caso, a variável $log(t_i)$ deve ser incluída como covariável no preditor, forçando seu coeficiente a ser 1. (Chamamos $log(t_i)$ de termo *offset*).
- Escrevendo o modelo na escala da resposta (média), temos:

$$\mu_i = t_i e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}. \quad (14)$$

Qualidade do ajuste

- A deviance para o MLG Poisson fica dada por:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{n} \left[y_i \, log \left(\frac{y_i}{\hat{\mu}_i} \right) - y_i + \hat{\mu}_i \right]. \tag{15}$$

- Para a distribuição de Poisson, $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}$ quando $\mu_i \to \infty$ para todo i.
- A estatística X^2 de Pearson, definida por:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{16}$$

também tem distribuição χ^2_{n-p} quando $\mu_i \to \infty$ para todo i.

- Nessas condições, tanto a deviance quanto a estatística X^2 podem ser usadas para testar a hipótese nula de que o modelo está bem ajustado.