GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula 12

Introdução

- Exemplos de dados de contagem ou frequência incluem:
  - número de ciclones tropicais que atravessam uma determinada costa;
  - número de pacientes com infecção urinária;
  - número de internações por doenças respiratórias;
  - número de casos de malária nos municípios do Pará;
  - número de acidentes de trânsito registrados;
  - número de erros tipográficos por slide, etc.

- Usualmente, utiliza-se a distribuição de Poisson para modelar tais observações.

- Se Y é o número de ocorrências, sua função de probabilidade pode ser escrita como:

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \ldots$$

  onde μ é o número médio de ocorrências.

- Pode ser mostrado que:

$$E(Y) = \mu \quad \text{e} \quad Var(Y) = \mu$$

Introdução

- Então μ é a frequência média ou a taxa de ocorrência, por exemplo:
  - Número médio de clientes que compram um determinado produto entre 100 clientes que entram na loja;
  - Para batidas de carro, a taxa pode ser definida de diferentes maneiras - batidas por 1000 habitantes ou batidas por 100.000 km viajados por veículo.
- A escala de tempo também deve ser incluída na definição, por exemplo:
  - A taxa de acidentes de carro é usualmente especificada como uma taxa "por ano" (acidentes por 100.000 km/ano);
  - Na Austrália, a taxa de ciclones tropicais refere-se à temporada de ciclones, que vai de novembro a abril.
- A taxa é especificada em termos de "em exposição".
  - Para acidentes de trabalho, cada trabalhador está exposto durante o período do trabalho, então a taxa pode ser definida em termos de pessoas "em risco".

Introdução

- O efeito das variáveis explicativas na resposta Y é modelado por meio do parâmetro μ.
- Quando os eventos referem-se a quantidades variáveis de exposição que precisam ser levadas em consideração ao modelar a taxa de eventos, usamos uma regressão de Poisson.
- Em uma segunda situação, a exposição é constante e as variáveis explicativas são usualmente categóricas.
- Nesse caso, se existem somente algumas variáveis explicativas, os dados são resumidos em tabelas e a variável resposta é a contagem em cada célula da tabela (modelo log-linear).

Regressão de Poisson

- Sejam $Y_1, \ldots, Y_n$ variáveis resposta independentes cujos valores observados são contagens ou frequências (podendo representar o número de eventos observados da exposição E_i para o i-ésimo padrão da covariável),
- A primeira distribuição a ser considerada é:

$$Y_i \sim \text{Poisson}(\mu_i), \quad \text{onde } \mu_i > 0$$

- A função de ligação mais utilizada é:

$$\log \mu_i = x_i' \beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

que é a função de ligação canônica.

- Logo,

$$\mu_i = \exp(x_i' \beta),$$

- Neste caso, temos um MLG, pois:
  - a distribuição da variável resposta pertence à família exponencial canônica;
  - temos um preditor linear dado por $x_i' \beta$;
  - temos uma função de ligação dada por $\log \mu_i$.

Interpretação

- Pela equação da função de ligação, obtém-se que:

$$\mu_i = \exp(x_i'\beta)$$

- Então, a razão entre as taxas μ_i e μ_j pode ser descrita como:

$$\frac{\mu_i}{\mu_j} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots + \beta_p x_{ip})}{\exp(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk} + \cdots + \beta_p x_{jp})}$$

- Suponha que μ_i e μ_j diferem apenas no valor da variável explicativa x_k. Dessa forma, a razão das taxas (rate ratio ou razão de taxas, RR) se reduz a:

$$RR = \frac{\mu_i}{\mu_j} = \frac{\exp(\beta_k x_{ik})}{\exp(\beta_k x_{jk})} = \exp\left[\beta_k (x_{ik} - x_{jk})\right]$$

- Assim, β_k representa o logaritmo da razão de taxas associada a uma variação unitária em x_k.

Interpretação

- Então, se aumentarmos em uma unidade o valor de x_k, a razão de taxas aumenta (ou diminui) para $\exp(\beta_k)$, ou seja, resulta em um efeito multiplicativo $\exp(\beta_k)$ na taxa μ.
- No caso de uma variável explicativa binária, temos:

$$x_k = \begin{cases} 0, & \text{se o fator está ausente} \\ 1, & \text{se o fator está presente} \end{cases}$$

- A razão de taxas para presença versus ausência é:

$$RR = \frac{E(Y_i \mid \text{presente})}{E(Y_i \mid \text{ausente})} = \exp(\beta_k),$$

dado que todas as outras variáveis explicativas permaneçam constantes.

- Portanto, as estimativas dos parâmetros são interpretadas na escala exponencial, em termos de razões de taxas.

Estimação e Teste de Hipóteses

- Sejam $Y_1, \ldots, Y_n$ variáveis resposta independentes com distribuição de Poisson de média μ_i.
- A função de log-verossimilhança é dada por:

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \log \left[\prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right] = \sum_{i=1}^{N} [-\mu_i + y_i \log(\mu_i) - \log(y_i!)]$$

- Os valores ajustados, geralmente denotados por e_i, são dados por:

$$\hat{y}_i = \hat{\mu}_i = \exp(x_i' \hat{\boldsymbol{\beta}}), \quad \text{para } i = 1, \ldots, n.$$

Teste de Hipóteses

- Para medir a qualidade de ajuste de um modelo, utilizamos a estatística da razão de verossimilhança, dada por:

$$D = 2\left[\ell(\hat{\boldsymbol{\mu}}_{\text{max}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\right]$$

onde $\hat{\boldsymbol{\mu}}_{\text{max}}$ é o vetor de estimativas de máxima verossimilhança para o modelo maximal (ou modelo saturado).

- No caso da distribuição de Poisson, o estimador de máxima verossimilhança de μ_i no modelo maximal é:

$$\hat{\mu}_i = y_i$$

- Assim, para o modelo maximal, temos:

$$\ell(\hat{\boldsymbol{\mu}}_{\text{max}}; \mathbf{y}) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} \log(y_i!)$$

Teste de Hipóteses

- Para o modelo de interesse, a log-verossimilhança é dada por:

$$\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) = -\sum_{i=1}^{n} \hat{\mu}_i + \sum_{i=1}^{n} y_i \log(\hat{\mu}_i) - \sum_{i=1}^{n} \log(y_i!)$$

  onde $\hat{\mu}_i$ é calculada a partir da estimativa de máxima verossimilhança dos parâmetros $\hat{\boldsymbol{\beta}}$.

- Assim, a estatística da razão de verossimilhança (ou deviance) é:

$$D = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$$

- Em termos de observados (o_i) e esperados (e_i) e como para a maioria dos modelos $\sum o_i = \sum e_i$, podemos escrever:

$$D = 2 \sum_{i=1}^{n} o_i \log\left(\frac{o_i}{e_i}\right)$$

- A estatística D tem distribuição assintótica χ^2_{n-p}, onde p é o número de parâmetros estimados em $\boldsymbol{\beta}$ e, como nos demais modelos, serve para medir a qualidade de ajuste.

Teste de Hipóteses

- Outra estatística para medir a qualidade de ajuste é o qui-quadrado de Pearson, que no caso da distribuição de Poisson é dado por:

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

- Sob a hipótese de que o modelo se ajusta bem aos dados,

$$X^2 \sim \chi^2_{n-p},$$

onde p é o número de parâmetros estimados.

- As estatísticas de bondade de ajuste X^2 (Pearson) e D (Deviance) são relacionadas. Utilizando uma expansão de Taylor, pode-se mostrar que:

$$D \approx X^2 \quad \text{quando } y_i \text{ está próximo de } \hat{\mu}_i.$$

Teste de Hipóteses

- Tanto a estatística de deviance (D) quanto o qui-quadrado de Pearson (X^2) podem ser comparadas com a distribuição qui-quadrado com $n - p$ graus de liberdade, onde p é o número de parâmetros estimados.
- Em geral, a aproximação à distribuição qui-quadrado é melhor para X^2 do que para D.
- Testes de hipóteses sobre os parâmetros β_j podem ser realizados utilizando as estatísticas Wald, Escore ou Razão de verossimilhanças.
- Intervalos de confiança também podem ser construídos de forma análoga. Por exemplo, para o parâmetro β_j

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \stackrel{\cdot}{\sim} N(0,1) \implies \frac{(\hat{\beta}_j - \beta_j)^2}{\text{Var}(\hat{\beta}_j)} \stackrel{\cdot}{\sim} \chi^2_1$$

Resíduos

- Os resíduos de Pearson são definidos por:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

- Esses resíduos podem ser padronizados para levar em conta a influência da observação i:

$$r_{pi} = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

onde h_{ii} é o i-ésimo elemento da diagonal da matriz hessiana, que mede a influência da observação sobre seu próprio valor ajustado.

- Os resíduos padronizados são assintoticamente normais e podem ser utilizados para:
  - verificar a linearidade do modelo;
  - identificar valores atípicos (outliers);
  - investigar possível associação com variáveis omitidas.

Resíduos

- Para a distribuição de Poisson, os resíduos de Pearson e a estatística qui-quadrado de bondade de ajuste estão relacionados por:

$$X^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

onde r_i representa o resíduo de Pearson da i-ésima observação.

- Os resíduos de deviance são definidos como:

$$d_i = \text{sinal}(o_i - e_i)\sqrt{2\left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right]}, \quad i = 1, \ldots, n$$

- Assim, a deviance total pode ser escrita como:

$$D = \sum_{i=1}^{n} d_i^2$$

- Os resíduos de deviance tendem a ser mais estáveis que os de Pearson em situações com contagens pequenas.

Exemplo: Mortes por AIDS

- As observações Y_i representam o número de mortes por AIDS na Austrália, por trimestre, de 1983 a 1986. Denotamos o tempo por t_i.
- Modelo proposto:

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 t_i$$

$$\mu_i = \exp(\beta_0 + \beta_1 t_i)$$

- A regressão de Poisson é um caso particular de modelo linear generalizado. Portanto, todos os métodos de estimação e inferência aplicáveis aos MLGs são válidos para a regressão de Poisson.

Exemplo: Mortes por AIDS

```
Call:
glm(formula = y ~ ti, family = poisson)
Deviance Residuals:
Min          1Q      Median    3Q      Max
-2.21346 -0.94490 -0.63783   0.01354  0.1377

Coefficients:
            Estimate  Std. Error z value
(Intercept) 0.37431   0.25179     1.487  2e-16
ti          0.24996   0.02219    11.265     **
---
```

Figura: Saída do ajuste do modelo de regressão de Poisson.

Exemplo 1: Mortes por AIDS

- O modelo estimado é:

$$\log(\hat{\mu}_i) = 0{,}374 + 0{,}250\, t_i$$

onde t_i representa o trimestre: $t_i = 1$ corresponde ao primeiro trimestre de 1983, $t_i = 2$ ao segundo trimestre, e assim por diante; $\hat{\mu}_i$ é a taxa bruta de óbitos por trimestre na Austrália entre 1983 e 1986.

- Se aumentarmos t_i em uma unidade, a taxa bruta de mortalidade aumenta em:

$$\exp(0{,}250) = 1{,}284$$

- Portanto, no período de 1983 a 1986, a taxa bruta de mortalidade por AIDS na Austrália aumenta, em média, 28% por trimestre.

- O intervalo de confiança de 95% para a razão das taxas é:

$$IC(RR) = \exp\{IC(\beta_1)\} = (1{,}231;\ 1{,}343)$$

Poisson como Aproximação da Binomial

- Existem casos em que é possível verificar claramente que a distribuição mais apropriada é a distribuição binomial.
- Isso ocorre porque a variável resposta é dada pelo número de sucessos limitado pelo tamanho da população (E_i).

$$Y_i \sim \text{Binomial}(E_i, \pi)$$

- No entanto, existe um resultado teórico importante que estabelece que, quando E é grande e π é pequeno, a distribuição binomial pode ser aproximada por uma distribuição de Poisson com média:

$$\mu = E\pi$$

- Assim:

$$Y_i \sim \text{Poisson}(E_i\pi)$$

é uma boa aproximação sob essas condições.

Poisson como Aproximação da Binomial

- Seja $Y_1, \ldots, Y_n$ o número de eventos observados em E_i indivíduos expostos sob a i-ésima condição de covariáveis.
- O valor esperado de Y_i pode ser escrito como:

$$E(Y_i) = \mu_i = E_i \theta_i$$

onde:

  - E_i é o tamanho da população exposta;
  - θ_i é a probabilidade de ocorrência (ou taxa) do evento.

- Por exemplo, suponha que Y_i seja o número de óbitos por doenças respiratórias em Niterói no ano de 2010.
- Nesse caso, Y_i depende da população residente (E_i); e de variáveis que afetam θ_i, como faixa etária, sexo, entre outras.

Poisson como Aproximação da Binomial

- A dependência da taxa de ocorrência θ_i em covariáveis é usualmente expressa como:

$$\theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

- Portanto, o modelo linear generalizado é:

$$Y_i \sim \text{Poisson}(\mu_i) \qquad \text{com} \qquad E(Y_i) = \mu_i = E_i \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

- Utilizando a função de ligação logarítmica, temos:

$$\log(\mu_i) = \log(E_i) + \mathbf{x}_i'\boldsymbol{\beta}$$

- O termo $\log(E_i)$ é conhecido como *offset*.

- O *offset* é uma constante conhecida incorporada ao procedimento de estimação, permitindo ajustar a taxa de ocorrência proporcionalmente ao tamanho da exposição E_i.

Exemplo: Óbitos por doenças respiratórias em Niterói

Faixa Etária	**Óbitos**		**População**	
	Fem	Masc	Fem	Masc
30 a 39 anos	4	6	39.219	35.577
40 a 49 anos	5	14	37.761	32.342
50 a 59 anos	16	22	35.573	28.286
60 a 69 anos	31	39	24.621	17.984
70 a 79 anos	68	87	16.366	10.188
80 anos ou mais	206	122	16.366	10.188

Fonte: Datasus, 2010.

Exemplo: Óbitos por doenças respiratórias em Niterói

```
Call:
glm(formula = y ~ sexo + Fx_etaria, family = poisson(), offset = log(n))

Deviance Residuals:
      1       2       3       4       5       6       7       8       9
-0.2430  -1.4053  -0.5652  -0.7429  -1.8472   2.0024   0.2127   1.1362   0.5243
     10      11      12
 0.7195   1.8816  -2.3176

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.07151    0.31931 -28.410  < 2e-16 ***
sexoM         0.29571    0.08059   3.670 0.000243 ***
Fx_etaria40-49  0.71089    0.39068   1.820 0.068818 .
Fx_etaria50-59  1.50281    0.35542   4.228 2.35e-05 ***
Fx_etaria60-69  2.52466    0.33809   7.467 8.18e-14 ***
Fx_etaria70-79  3.80400    0.32636  11.656  < 2e-16 ***
Fx_etaria80+    4.55359    0.32110  14.181  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1289.081  on 11  degrees of freedom
Residual deviance:   21.367  on  5  degrees of freedom
AIC: 96.775
```

Figura: Saída do ajuste do modelo de regressão de Poisson.

Exemplo: Óbitos por doenças respiratórias em Niterói

- A taxa de mortalidade por doenças respiratórias entre os homens residentes em Niterói no ano de 2010 é:

$$\exp(0{,}29571) = 1{,}344$$

  vezes a taxa correspondente para as mulheres.

- Como as categorias da faixa etária estão alinhadas temporalmente, podemos afirmar que, à medida que aumenta a idade, a taxa de mortalidade também aumenta.

- Em particular, a taxa de mortalidade entre pessoas de 40 a 49 anos é:

$$\exp(0{,}71089) = 2{,}04$$

  vezes a taxa observada para pessoas de 30 a 39 anos.

Exemplo: Óbitos por doenças respiratórias em Niterói

Como comparar a taxa de mortalidade entre pessoas de 0 a 79 anos e aquelas de 80 anos ou mais? Essa diferença é significativa?

- Para comparar dois fatores de uma mesma variável (diferentes da categoria base), basta calcular a diferença entre os efeitos estimados:

$$\beta_{80+} - \beta_{70-79}$$

- No exemplo, temos:

$$\hat{\beta}_{80+} - \hat{\beta}_{70-79} = 4{,}55359 - 3{,}80400 = 0{,}74959$$

- Assim, a razão entre as taxas de mortalidade é:

$$\exp(0{,}74959) = 2{,}12$$

ou seja, a taxa entre pessoas de 80 anos ou mais é aproximadamente 2 vezes maior do que a taxa entre pessoas de 70 a 79 anos.

Exemplo: Óbitos por doenças respiratórias em Niterói

- Para verificar se essa diferença é significativa, podemos testar a hipótese H_0 : $\beta_{80+} - \beta_{70-79} = 0$.
- Em ambos os casos, precisamos calcular a variância de:

$$\widehat{\text{Var}}(\hat{\beta}_{80+} - \hat{\beta}_{70-79}) = \widehat{\text{Var}}(\hat{\beta}_{80+}) + \widehat{\text{Var}}(\hat{\beta}_{70-79}) - 2\,\widehat{\text{Cov}}(\hat{\beta}_{80+}, \hat{\beta}_{70-79})$$

- Usando a função `vcov(ajuste)` no R, obtemos as estimativas necessárias. Assim:

$$\widehat{\text{Var}}(\hat{\beta}_{80+} - \hat{\beta}_{70-79}) = 0{,}1065 + 0{,}1031 - 2 \times 0{,}1001 = 0{,}0094$$

- Logo, o intervalo de confiança de 95% é dado por:

$$\hat{\beta}_{80+} - \hat{\beta}_{70-79} \pm 1{,}96\sqrt{0{,}0094} = (0{,}5595; 0{,}9397)$$

- Como o intervalo não contém zero, a diferença entre as duas faixas etárias é significativa.

Sobredispersão

- Sabe-se que a distribuição de Poisson possui a seguinte propriedade:

$$Y \sim \text{Poisson}(\lambda) \quad \Rightarrow \quad E(Y) = \lambda = \text{Var}(Y)$$

- Essa propriedade dificilmente é observada na prática. Usualmente:

$$E(Y) < \text{Var}(Y)$$

  Quando isso ocorre, dizemos que há sobredispersão (ou superdispersão).

- É raro, mas é possível observar o oposto:

$$E(Y) > \text{Var}(Y)$$

  Nesse caso, temos o fenômeno de subdispersão.

Sobredispersão

- Consequências: Quando tratamos incorretamente a incerteza associada ao modelo, as variâncias são mal estimadas, resultando em:
  - erros-padrão incorretos;
  - testes de hipóteses e intervalos de confiança imprecisos;
  - conclusões incorretas sobre a significância dos parâmetros.
- Causas:
  - variáveis explicativas omitidas;
  - forma estrutural incorreta do modelo;
  - falta de independência entre as observações.
- Soluções: Considerar uma forma funcional diferente para o modelo:
  - Distribuição Binomial Negativa;
  - Modelos de Quasi-verossimilhança;
  - Inclusão de um efeito aleatório (modelos hierárquicos);
  - Modelos de Poisson com excesso de zeros.