
# GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula 10

# Modelo logístico geral

- O modelo de regressão logística mais geral assume

$$\text{logito}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

  onde $\mathbf{x}_i$ é o vetor de variáveis contínuas ou variáveis dummy e $\boldsymbol{\beta}$ é o vetor paramétrico.

- Esse modelo é amplamente usado para analisar dados envolvendo respostas binárias ou binomiais.

- Os dados podem ser agrupados como frequências para cada padrão de covariável (isto é, observações com os mesmos valores de todas as variáveis explicativas).

- Ou ainda, cada observação pode ser codificada como 0 ou 1 e seu padrão de covariável listado separadamente.

# Modelo logístico geral

- Se os dados podem ser agrupados, a resposta $Y_i$ (número de sucessos para o padrão de covariável $i$) pode ser modelada pela distribuição Binomial.

- Se cada observação tem um padrão de covariável diferente, então $n_i = 1$ e a resposta $Y_i$ é binária e pode ser modelda pela Bernoulli.

**Exemplo:** Testes clínicos são realizados para comparar a eficiência de um novo procedimento cirúrgico frente a uma técnica já conhecida. Os testes foram realizados em 2 hospitais ($x_1 = 1, 2$). Em cada hospital, os pacientes foram distribuídos aleatoriamente para 2 procedimentos cirúrgicos ($x_2 = 1, 2$).

# Modelo logístico geral

- No primeiro mês de estudo, sete pacientes foram recrutados.
- Estes pacientes são listados na tabela abaixo pelo seu número de identificação e pelas classes de covariáveis.

| **Dados listados pelo nº do paciente** | | | **Dados listados pela classe da covariável** | | |
|---|---|---|---|---|---|
| Paciente nº | Covariável$(x_1, x_2)$ | Resposta$(Y)$ | Covariável $(x_1, x_2)$ | Tamanho$(n)$ | Resposta$(Y)$ |
| 1 | 1,1 | 0 | 1,1 | 2 | 1 |
| 2 | 1,2 | 1 | 1,2 | 3 | 2 |
| 3 | 1,2 | 0 | 2,1 | 1 | 0 |
| 4 | 2,1 | 0 | 2,2 | 1 | 1 |
| 5 | 2,2 | 1 | | | |
| 6 | 1,2 | 1 | | | |
| 7 | 1,1 | 1 | | | |

# Modelo logístico geral

- Os dados listados pelo padrão da covariável crescem em eficiência à medida que o número de pacientes aumenta.
- Nesse caso, as respostas têm a forma $\frac{y_i}{n_i}$, onde $0 < y_i < n_i$ é o número de sucessos em $n_i$ indivíduos no $i$-ésimo subgrupo (classe).
- Dados não agrupados podem ser considerados casos especiais em que $n_1 = n_2 = \cdots = n_N = 1$.
- O único problema em agrupar acontece se a ordem em que as observações aparecem for relevante.

# Modelo logístico geral

- O estimador de máxima verossimilhança do parâmetro $\boldsymbol{\beta}$, e consequentemente das probabilidades $\pi_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$, é obtido maximizando a função de log-verossimilhança:

$$\ell(\boldsymbol{\pi}; \mathbf{y}) = \sum_{i=1}^{N} \left[ y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right]$$

- A maximização é feita usando o método escore iterativo.

- A Deviance é dada por:

$$D = 2 \sum_{i=1}^{N} \left[ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]$$

# Modelo logístico geral: Deviance

- A Deviance também pode ser escrita como:

$$D = 2 \sum o \log\left(\frac{o}{e}\right)$$

onde:

  - $o$: frequências observadas, $y_i$ e $(n_i - y_i)$, das células da tabela de observações.
  - $e$: frequências esperadas estimadas, ou valores ajustados:

$$\hat{y}_i = n_i \hat{\pi}_i \quad \text{e} \quad n_i - \hat{y}_i = n_i - n_i \hat{\pi}_i$$

- O somatório é realizado em todas as $2 \times N$ células da tabela de observações.

# Modelo logístico geral: Deviance e Inferência

- A Deviance do modelo logístico não depende de um parâmetro de ruído $\sigma^2$ (como no modelo normal).

- A bondade de ajuste pode ser avaliada e hipóteses podem ser testadas usando a aproximação

$$D \overset{\cdot}{\sim} \chi^2_{N-p}$$

onde $p$ é o número de parâmetros estimados e $N$ o número de padrões de covariável.

- Os métodos de estimação e as distribuições amostrais para inferência dependem de resultados assintóticos.

- Para estudos pequenos ou com poucas observações por padrão de covariável, esses resultados assintóticos podem não ser adequados.

# Resíduos em regressão logística

- As questões estudadas na análise dos resíduos em modelos de regressão múltipla para resposta contínua também são relevantes no contexto de respostas binárias.
- Entre essas questões estão:
  - Inclusão ou exclusão de covariáveis;
  - Análise gráfica dos resíduos.
- Existem duas formas principais de resíduos, correspondendo às medidas de bondade de ajuste $D$ e $X^2$.
- Se existem $m$ diferentes níveis de covariáveis, então podemos calcular $m$ resíduos.

# Resíduos em regressão logística

- Seja $Y_k$ o número de sucessos, $n_k$ o número de ensaios e $\hat{\pi}_k$ a probabilidade de sucesso estimada para o $k$-ésimo nível de covariável.

- O resíduo de Pearson (ou qui-quadrado) é definido como:

$$X_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}, \quad k = 1, \ldots, m$$

- A soma dos quadrados dos resíduos resulta na estatística qui-quadrado de Pearson para bondade de ajuste:

$$\sum_{k=1}^{m} X_k^2 = X^2$$

# Resíduos em regressão logística

- Os resíduos de Pearson padronizados são definidos como:

$$r_{Pk} = \frac{X_k}{\sqrt{1 - h_{kk}}}$$

onde:

  - $X_k$ é o resíduo de Pearson do $k$-ésimo nível de covariável;
  - $h_{kk}$ mede o grau de influência da observação no ajuste do modelo (*leverage*) e corresponde ao $k$-ésimo elemento da diagonal da matriz de projeção $\mathbf{H}$.

# Resíduos em regressão logística

- Um segundo tipo de resíduo é o resíduo da Deviance.
- O valor total da Deviance pode ser escrito como:

$$D = \sum_{k=1}^{m} d_k^2$$

- Cada componente individual é dado por:

$$d_k = \text{sign}(y_k - n_k\hat{\pi}_k)\sqrt{2\left[y_k \log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k}\right)\right]}$$

- O termo $\text{sign}(y_k - n_k\hat{\pi}_k)$ garante que $d_k$ tenha o mesmo sinal do resíduo de Pearson $X_k$.

# Resíduos em regressão logística

- Os resíduos padronizados baseados na Deviance são definidos por

$$r_{Dk} = \frac{d_k}{\sqrt{1 - h_{kk}}}.$$

- As análises de resíduos em MLG devem ser conduzidas da mesma maneira que em modelos lineares normais.

- Se os dados são binários, ou se $n_i$ é pequeno para a maioria dos níveis das covariáveis, haverá poucos valores distintos dos resíduos e os gráficos serão pouco informativos.

- Nesse caso, deve-se confiar mais nas estatísticas agregadas de bondade de ajuste ($X^2$ e $D$), bem como em outros diagnósticos.

# Sobredispersão

- Sobredispersão ou variação extra-binomial é um fenômeno comum em modelagem de dados binários agrupados.
- Ocorre quando a variação observada excede aquela assumida pelo modelo.
- Em outras palavras, observações $y_i$ que são assumidas seguir uma distribuição Binomial podem apresentar variância maior que $n_i\pi_i(1-\pi_i)$.
- Uma abordagem é incluir um parâmetro extra $\phi$ tal que:

$$V(Y_i) = \phi\, n_i\pi_i(1-\pi_i)$$

- Interpretação de $\phi$:
  - $\phi = 1$: variabilidade binomial (modelo adequado);
  - $\phi > 1$: presença de sobredispersão (variação extra-binomial).
- Indícios de sobredispersão podem ocorrer quando as estatísticas de qualidade de ajuste (Deviance e $X^2$ de Pearson) são grandes em relação aos seus graus de liberdade ($N-p$).

# Sobredispersão: Considerações

- Alguns pontos aberrantes podem aumentar substancialmente o valor da Deviance e a *"simples eliminação desses pontos pode reduzir as evidências de sobredispersão"*.

- Para investigar o efeito de observações influentes, estão disponíveis na regressão logística as estatísticas: delta beta, delta qui-quadrado e delta deviance.

- A sobredispersão pode ter duas causas principais:

  - Modelo especificado incorretamente: necessidade de incluir termos adicionais, como interações ou termos quadráticos;

  - Falta de independência entre as observações.

# Sobredispersão: Estimação do parâmetro $\phi$

- O parâmetro $\phi$ pode ser estimado com base na estatística de Pearson ou na Deviance:

$$\hat{\phi} = \frac{X^2}{N-p} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i - \hat{\mu}_i)^2}{\widehat{\mathrm{V}}(y_i)}$$

$$\hat{\phi} = \frac{D}{N-p} = \frac{1}{N-p}\sum_{i=1}^{N} d_i^2$$

  onde $N - p$ é o total de graus de liberdade.

- Na prática, multiplica-se $\sqrt{\hat{\phi}}$ pelos erros-padrão estimados dos coeficientes $\beta$ para ajustar intervalos de confiança e testes de hipóteses.

# Conceito de Chance (Odds)

- Uma forma natural de quantificar as chances de um evento é utilizando probabilidades.
- Outra forma é a partir da razão de probabilidades.
- Se $A$ e $B$ são eventos tais que $A \cap B = \varnothing$ e $A \cup B = \Omega$, a razão de probabilidades é dada por:

$$\frac{P(A)}{P(B)} = \frac{P(A)}{1 - P(A)}$$

  sendo denominada de chances (odds) do evento $A$ relativo ao evento $B$.
- As chances do evento $A$ também podem ser calculadas como a razão entre o número de vezes que $A$ ocorre e o número de vezes que $A$ não ocorre.

# Exemplos de Chance

**Exemplo 1:** Uma chance de 4 significa que esperamos que as ocorrências sejam 4 vezes as não ocorrências do evento.

**Exemplo 2:** A probabilidade de nascimento de um indivíduo do sexo masculino é cerca de $0,515$. Então a chance desse evento é:

$$\frac{0{,}515}{0{,}485} \approx 1{,}062$$

A chance em favor do nascimento de um indivíduo do sexo masculino é de aproximadamente 106 para 100, ou seja, 106 nascimentos masculinos para cada 100 femininos.

# Razão de Chances

- O modelo logístico pressupõe que o logaritmo da chance é linearmente relacionado com as variáveis explicativas.

- Considere inicialmente o modelo logístico linear simples, em que $\pi(x)$ é a probabilidade de sucesso dado o valor $x$ de uma variável explicativa:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

  onde $\beta_0$ e $\beta_1$ são parâmetros desconhecidos.

- Esse modelo poderia, por exemplo, ser aplicado para analisar a associação entre uma determinada doença e a ocorrência (ou não) de um fator particular.

# Razão de Chances

- Suponha que sejam amostrados, independentemente, $n_1$ indivíduos com presença do fator ($x = 1$) e $n_2$ indivíduos com ausência do fator ($x = 0$). Seja $\pi(x)$ a probabilidade de desenvolvimento da doença após um certo período fixo.
- Assim,a chance de desenvolvimento da doença para indivíduos com o fator é

$$\frac{\pi(1)}{1-\pi(1)} = \exp(\beta_0 + \beta_1)$$

- Já para indivíduos sem o fator é

$$\frac{\pi(0)}{1-\pi(0)} = \exp(\beta_0).$$

- A razão de chances fica:

$$\psi = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{\pi(1)\{1-\pi(0)\}}{\pi(0)\{1-\pi(1)\}} = \exp(\beta_1),$$

dependendo apenas do parâmetro $\beta_1$.

# Razão de chances

**Exemplo**: Suponha que a variável resposta corresponda ao uso de anticoncepcional e estamos interessados na razão do número esperado de usuários para cada não usuário. Como variável explicativa temos um fator com dois níveis: urbano ($x = 1$) e rural ($x = 0$). Suponha que as chances em favor do uso sejam de 4 para 1 em áreas urbanas e de 2 para 1 em áreas rurais.

- Então, a razão de chances nas áreas urbanas para as chances em áreas rurais é 2.
- Neste caso, o número esperado de usuários para cada não usuário em áreas urbanas é duas vezes o de áreas rurais.
- Em termos de porcentagem, o número esperado de usuários para cada não usuário é 100% maior em áreas urbanas comparado às áreas rurais.

# Razão de chances

- O modelo nesse caso é

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x,$$

  onde $x = 1$ se urbana e $x = 0$ se rural.

- Para $x = 1 \Longrightarrow \frac{\pi_1}{1-\pi_1} = \exp(\hat{\beta}_0 + \hat{\beta}_1) = 4$

- Para $x = 0 \Longrightarrow \frac{\pi_0}{1-\pi_0} = \exp(\hat{\beta}_0) = 2$

- Então:

$$\psi = \frac{\dfrac{\pi_1}{1-\pi_1}}{\dfrac{\pi_0}{1-\pi_0}} = \exp(\hat{\beta}_1) = \frac{4}{2} = 2$$

- A chance em áreas urbanas é 2 vezes a chance em áreas rurais quando comparamos usuários e não usuários.

# Razão de chances

- No caso em que a razão de chances resulta em um número menor que 1, a interpretação pode ser feita da seguinte forma:

- Suponha que no exemplo a razão de chances resulte em 0,2. Então, a chance em áreas rurais é 5 vezes a chance em áreas urbanas quando comparamos usuários e não usuários de anticoncepcional.

- Em termos de porcentagem, o número esperado de usuários para cada não usuário é 80% menor em áreas urbanas comparado às áreas rurais. (Ou seja, a chance em áreas urbanas é 80% menor do que em áreas rurais).

# Razão de chances: Exemplo

**Exemplo**: Estudo para avaliar o efeito da taxa e do volume de ar inspirado por uma pessoa na probabilidade de ocorrência de um acidente vascular.

Resposta igual a 1: sucesso, ocorrência do evento.
Resposta igual a 0: fracasso, não ocorrência do evento.

- Neste caso, os dados aparecem não agrupados e as variáveis explicativas foram medidas em escala contínua.

- O modelo a ser ajustado é da forma:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \,\text{taxa}_i + \beta_2 \,\text{volume}_i$$

  onde $\pi_i$ é a probabilidade de que o $i$-ésimo indivíduo sofra um acidente vascular.

# Razão de chances: Exemplo

```
> summary(modelo1)

Call: glm(formula = resposta ~ taxa + volume, family
=binomial(link = "logit"),    data = resp)

Coefficients:
             Estimate     Std. Error     z value   Pr(>|z|)
(Intercept)    -9.5296         3.2332      -2.947   0.00320 **
taxa            2.6491         0.9142       2.898   0.00376 **
volume          3.8822         1.4286       2.717   0.00658 **
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.772  on 36  degrees of freedom AIC: 35.772
```

# Razão de chances: Exemplo

- A chance estimada do $i$-ésimo indivíduo sofrer um acidente vascular é dada por:

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 \,\text{taxa}_i + \hat{\beta}_2 \,\text{volume}_i\right)$$

- Para o terceiro indivíduo:

$$\frac{\hat{\pi}_3}{1-\hat{\pi}_3} = \exp\{-9{,}530 + 2{,}649(2{,}5) + 3{,}882(1{,}25)\} = 6{,}994$$

- Assim, para uma pessoa que tem taxa de ar inspirado igual a 2,5 e volume igual a 1,25, a ocorrência de um acidente vascular é de aproximadamente 7 para 1.

# Razão de chances: Exemplo

- Vamos observar o que ocorre quando variamos a taxa em uma unidade e mantemos o volume de ar constante.

$$\frac{\frac{\hat{\pi}_i}{1-\hat{\pi}_i}}{\frac{\hat{\pi}_j}{1-\hat{\pi}_j}} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1(\text{taxa}+1) + \hat{\beta}_2\,\text{volume}_i\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1(\text{taxa}) + \hat{\beta}_2\,\text{volume}_i\right)} = \exp(\hat{\beta}_1)$$

- Então, para cada unidade acrescida na taxa de ar inspirado, mantendo-se o volume constante, a razão de chances aumenta em:

$$\exp(\hat{\beta}_1) = \exp(2{,}649) \approx 14{,}14$$

# Razão de chances: Exemplo

- De forma análoga, para um aumento de uma unidade no volume de ar inspirado, mantendo-se a taxa constante, a razão de chances aumenta em:

$$\exp(\hat{\beta}_2) = \exp(3{,}882) \approx 48{,}52$$

- Como ambos os coeficientes estimados são positivos, aumentos nas variáveis implicam em aumentos na chance de ocorrência do evento

- Pela simetria da função logística, se tivéssemos modelado o evento complementar (não ocorrência do acidente vascular), obteríamos os coeficientes com sinais trocados, mas a estatística Deviance permaneceria igual.

# Intervalo de confiança para a razão de chances

- Considere o modelo:

$$Y_i \sim \text{Binomial}(n_i, \pi_i),$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x$$

onde $x$ é um fator:

  - $x = 0$: ausência do fator
  - $x = 1$: presença do fator

- Usualmente, é mais fácil interpretar os efeitos das variáveis explicativas em termos de razões de chances do que olhar diretamente para os parâmetros $\beta$.

# Intervalo de confiança para a razão de chances

- Baseado no modelo, podemos obter o quanto as chances aumentam na presença do fator:

$$\frac{\pi(0)}{1-\pi(0)} = \exp(\beta_0) \quad \text{(ausência do fator, } x = 0)$$

$$\frac{\pi(1)}{1-\pi(1)} = \exp(\beta_0 + \beta_1) \quad \text{(presença do fator, } x = 1)$$

- Portanto:

$$\psi = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \exp(\beta_1)$$

- Se $\beta_1 = 0$, então $\psi = 1$, o que corresponde a um "não efeito" da presença do fator.
- Denotando a razão de chances associada à variável $j$ por $\psi_j$, no modelo de regressão logística temos:

$$\psi_j = \exp(\beta_j), \quad \hat{\psi}_j = \exp(\hat{\beta}_j).$$

# Intervalo de confiança para a razão de chances

- O intervalo de confiança assintótico para $\psi_j$ com nível de confiança $100(1-\alpha)\%$ terá limites:

$$\left[\exp\left(\hat{\beta}_j - z_{1-\alpha/2}\,\widehat{SE}(\hat{\beta}_j)\right),\ \exp\left(\hat{\beta}_j + z_{1-\alpha/2}\,\widehat{SE}(\hat{\beta}_j)\right)\right]$$

- Por exemplo, intervalos de 95% de confiança para $\psi_j$ são calculados através de:

$$\exp\left(\hat{\beta}_j \pm 1,96 \cdot \widehat{SE}(\hat{\beta}_j)\right)$$

- Intervalos que não incluem o valor unitário correspondem a valores de $\beta_j$ significativamente diferentes de zero.

# Intervalo de confiança para a razão de chances

- Obter o intervalo de $\psi_j$ exponenciando os limites supõe simetria em escala log, o que é uma limitação, pois após a exponenciação o intervalo torna-se assimétrico e pode distorcer a interpretação.

- O método Delta permite obter a variância aproximada de uma função $g(\beta_j)$ de um estimador assintoticamente normal.

- Se $\hat{\beta}_j \sim N(\beta_j, \widehat{V}(\hat{\beta}_j))$ e $\psi_j = g(\beta_j) = \exp(\beta_j)$:

$$\widehat{V}(\hat{\psi}_j) \approx \left[g'(\hat{\beta}_j)\right]^2 \cdot \widehat{V}(\hat{\beta}_j)$$

- Como $g'(\beta_j) = \exp(\beta_j)$, temos:

$$\widehat{V}(\hat{\psi}_j) \approx \exp(2\hat{\beta}_j) \cdot \widehat{V}(\hat{\beta}_j)$$

- Assim, o intervalo de confiança é dado por:

$$IC(\psi_j) = \hat{\psi}_j \pm z_{1-\alpha/2} \sqrt{\widehat{V}(\hat{\psi}_j)}$$