R para Ciência de Dados

Regressão Logística

Profa Carolina Paraíba e Prof Gilberto Sassi

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal da Bahia

Agosto de 2023

R para Ciência de Dados: Regressão Logística

R para Ciência de Dados: Regressão Logística

- Introdução
- Modelo de Regressão Logística
- Inferência
- Exemplos
- Regressão Logística no R

Modelos Lineares Generalizados

A classe de modelos lineares generalizados (MLG), proposta por Nelder e Wedderburn (1972), estende a classe dos modelos lineares normais.

A ideia básica por traz dos MLG consiste em permitir que a distribuição da variável resposta pertença à família exponencial de distribuições e flexibilizar a relação funcional entre a média da variável resposta $E(Y) = \mu$ e as covariáveis por meio de um preditor linear η e uma função de ligação g.

Modelos Lineares Generalizados

Os MLG envolvem três componentes:

1. **Componente aleatória:** representada por um conjunto de variáveis aleatórias (v.a.'s) independentes, $Y_1, \ldots, Y_n$, provenientes de uma mesma distribuição pertencente à família exponencial com

$$E(Y_i) = \mu_i; \quad i = 1, \ldots, n.$$

Modelos Lineares Generalizados

2. **Componente sistemática:** cada Y_i está associada a um conjunto de p variáveis explicativas, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, que definem um preditor linear dado por

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta}, \ i = 1, \ldots, n,$$

onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ é um vetor de parâmetros.

Modelos Lineares Generalizados

3. **Função de ligação:** uma função g monótona e diferenciável que relaciona a componente aleatória à componente sistemática, isto é, relaciona a média ao preditor linear,

$$g(\mu_i) = \eta_i.$$

A função de ligação que transforma a média μ_i no parâmetro natural θ_i é a *função de ligação canônica*. Para esta função de ligação, tem-se $g(\mu_i) = \eta_i = \theta_i$, isto é, o preditor linear modela diretamente o parâmetro canônico.

Modelo de Regressão Logística

Um caso particular de MLG é o modelo de regressão logística, que relaciona um conjunto de variáveis regressoras ao parâmetro binomial de variáveis binárias por meio da função de ligação logito.

Regressão Logística Binária: usada quando a resposta é binária (ou seja, tem dois resultados possíveis).

Regressão Logística Nominal: usada quando há três ou mais categorias sem ordenação natural dos níveis.

Regressão Logística Ordinal: usada quando existem três ou mais categorias com ordenação natural dos níveis, mas a ordenação dos níveis não significa necessariamente que os intervalos entre eles sejam iguais.

Variáveis binárias

Em muitas aplicações, a variável resposta de interesse representa um número fixo m de observações que podem assumir uma de duas possíveis categorias.

Por exemplo, a resposta pode ser "vivo" ou "morto", ou "presente" e "ausente".

Usualmente, usamos os termos sucesso e fracasso para as duas categorias.

Seja $\boldsymbol{z} = (z_1, \ldots, z_m)$ um vetor de observações de $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$, tal que os Z_j' são independentes com

$$P(Z_j = 1) = \pi_j \quad \text{e} \quad P(Z_j = 0) = 1 - \pi_j. \tag{1}$$

Variáveis binárias

Então, Z_j é uma variável binária.

Note que Z_j é uma variável categórica que assume uma de duas categorias.

A variável binária Z_j é dita ser um ensaio de Bernoulli, isto é $Z_j \sim Bernoulli(\pi_j)$.

Variáveis binárias

Se todos os π_j's são iguais, isto é $\pi_j = \pi$ para todo j, podemos definir a v.a.

$$Y = \sum_{j=1}^{m} Z_j,$$

que denota o número de sucessos em m ensaios independentes de Bernoulli e tem distribuição binomial com índice m e parâmetro π, $Y \sim Binomial(m, \pi)$, com

$$\begin{aligned} f_Y(y|\pi) &= \binom{m}{y} \pi^y (1-\pi)^{m-y} \\ &= \exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + m \log(1-\pi) + \log\binom{m}{x} \right\}, \end{aligned} \tag{2}$$

onde $\pi \in (0, 1)$, $y = \{0, \ldots, m\}$ e m é um inteiro positivo.

Variáveis binárias

O modelo binomial pertence à família exponencial com parâmetro natural

$$\theta = \log\left(\frac{\pi}{1-\pi}\right). \tag{3}$$

A esperança e variância da v.a. $Y \sim Binomial(m, \pi)$ são dadas por

$$E(Y) = m\pi \tag{4}$$

e

$$Var(Y) = m\pi(1-\pi). \tag{5}$$

Variáveis binárias

Para o caso geral de n variáveis independentes $Y_1, \ldots, Y_n$ correspondentes ao número de sucessos em n diferentes subgrupos, as frequências de sucessos e fracassos são mostradas na Tabela 1.

Note que para os tamanhos amostrais binomiais $\boldsymbol{m} = (m_1, \ldots, m_n)$, temos que o número total de observações binárias é $N = \sum_{i=1}^{n} m_i$.

Tabela 1: Frequências para n distribuições binomiais.

	Subgrupo		
	1	$\ldots$	n
Sucesso	Y_1	$\ldots$	Y_1
Fracasso	$m_1 - Y_1$	$\ldots$	$m_n - Y_n$
Total	m_1	$\ldots$	m_n

Modelo Regressão Logística

Suponha que as variáveis resposta Y_i's, $i = 1, \ldots, n$, associadas aos indivíduos (ou unidades experimentais) representem a soma de m_i sequências de respostas binárias independentes com probabilidade de sucesso comum π_i, ou seja, $Y_i \sim Binomial(m_i, \pi_i)$.

Temos que,

$$E(Y_i) = \mu_i = m_i \pi_i, \;\; i = 1, \ldots, n. \tag{6}$$

Como m_i é considerado conhecido, modelar a média da variável resposta μ_i é equivalente a modelar a probabilidade binomial π_i.

Modelo Regressão Logística

Em outras palavras:

- Se $Y_i \sim Binomial(m_i, \pi_i)$, onde Y_i é o número de sucessos em m_i ensaios de Bernoulli, temos que a média da v.a. Y_i, $\mu_i = E(Y_i) = m_i\pi_i$, depende de m_i.
- Então, vamos assumir que $y_1, \ldots, y_n$ são proporções binomiais tais que $m_i y_i \sim Binomial(m_i, \pi_i)$. Isto é, y_i é a proporção amostral de sucessos em m_i ensaios de Bernoulli e $E(Y_i) = \pi_i$ é independente de m_i.

Modelo Regressão Logística

Em muitos estudos, cada variável resposta Y_i pode estar associada a um vetor de covariáveis $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, que são informações que influenciam a probabilidade binomial π_i.

O interesse estatístico é verificar a relação entre π_i e as covariáveis $\boldsymbol{x}_i$.

Para investigar esta relação é conveniente estabelecer um modelo formal. Como a distribuição binomial pertence à família exponencial, esse problema pode ser visto como um caso particular de MLG.

Observação

Na prática, a construção do modelo necessita que algumas suposições sejam assumidas, por exemplo a independência entre as observações, linearidade da componente sistemática, entre outras. Essas suposições não podem ser garantidas, mas podem ser checadas.

Modelo Regressão Logística

Sob a estrutura de MLG, vamos supor que a dependência de π_i em $\boldsymbol{x}_i$ é dada pela combinação linear

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{7}$$

A menos que restrições sejam impostas aos β_j's, temos que $-\infty < \eta_i < +\infty$.

Então, para expressar os π_i's como uma combinação linear dos $\boldsymbol{x}_i$'s, devemos usar uma função de ligação g que leve os valores do intervalo unitário $(0, 1)$ a valores no reais.

Para dados binários, é comum escolher uma transformação g que seja simétrica em torno de zero e que corresponda a uma função de distribuição acumulada (f.d.a.).

Modelo Regressão Logística

Quando g corresponde à função logística, temos

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{8}$$

Assim, temos a formulação do modelo de regressão logística, onde

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1+e^{\sum_{j=1}^{p} x_{ij}\beta_j}}; \quad i = 1, \ldots, n \tag{9}$$

ou

$$logito(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, \ldots, n. \tag{10}$$

Modelo Regressão Logística

- π_i é a probabilidade de que uma observação esteja numa categoria especificada da variável binária Y_i, usualmente chamada de *probabilidade de sucesso*.
- Observe que o modelo descreve a *probabilidade de um evento* ocorrer em função de um conjunto de covariáveis $\boldsymbol{x}_i$.
- Com o modelo logístico, estimativas de π_i sempre estarão entre 0 e 1:
  - o numerador é positivo porque é a potência de um valor positivo (e);
  - o denominador é $(1 +$ numerador$)$, então o resultado sempre será menor do que 1.

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Suponha que $p = 3$ e que $x_{i1} = 1$ para todo i, então, o modelo pode ser escrito em termos do logaritmo da chance (odds) de resposta positiva,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}. \tag{11}$$

Equivalentemente, o modelo pode ser escrito em termos da chance de resposta positiva,

$$\frac{\pi_i}{1-\pi_i} = \exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}. \tag{12}$$

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Por fim, a probabilidade de resposta positiva é

$$\pi_i = g^{-1}(\eta_i) = \frac{\exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}}{1 + \exp\{\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}\}}. \quad (13)$$

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Suponha que x_{i2} e x_{i3} não são funcionalmente relacionadas.

Para x_{i2} fixa, o modelo pode ser interpretado como segue: o efeito de uma unidade de mudança em x_{i3} é o aumento da chance por uma quantidade β_3; o efeito de uma unidade de mudança em x_{i3} é o aumento da chance de uma resposta positiva multiplicativamente pelo fator e^{β_3}.

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

As interpretações na escala da probabilidade são mais complicadas pois o efeito em π_i de uma unidade de mudança em x_{i3} depende dos valores de x_{i2} e x_{i3}.

A derivada de π_i em relação a x_{i3} é

$$\frac{\partial \pi_i}{\partial x_{i3}} = \beta_3 \pi_i (1 - \pi_i).$$

Então, uma pequena mudança em x_{i3} tem um efeito maior se π_i é próximo de 0.5 do que se π_i é próximo de 0 ou 1 (como medida na escala de probabilidade).

Modelo Regressão Logística

Interpretação de β: efeito na probabilidade e na razão de chances

Para o caso geral em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, temos que o parâmetro β_j refere-se ao efeito da j-ésima covariável no logaritmo da chance de resposta positiva (sendo que as outras covariáveis são mantidas fixas). Então e^{β_j} é o efeito multiplicativo do aumento de uma unidade na j-ésima covariável na chance de resposta positiva.

Estimação

Em MLG, o método de estimação mais comumente utilizado é o de máxima verossimilhança (MV).

Para $\boldsymbol{y} = (y_1, \ldots, y_n)$ um vetor de n observações de $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, em que cada $Y_i \sim Binomial(m_i, \pi_i)$, a função de log-verossimilhança de $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ é dada por

$$\ell(\boldsymbol{\pi}|\boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) + \log\binom{m_i}{y_i} \right]. \qquad (14)$$

Estimação

Temos que

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

Então, a função de log-verossimilhança do modelo de regressão logística é dada por

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=1}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} m_i \log\left[1 + \exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]. \quad (15)$$

Dstribuição assintótica de $\hat{\beta}$:

A distribuição assintótica de $\hat{\beta}$ é a base da construção de testes e intervalos de confiança, em amostras grandes, para os parâmetros dos MLG. Sob condições gerais de regularidade, para amostras grandes, tem-se

$$\hat{\beta} \overset{a}{\sim} N_p(\beta, \boldsymbol{I}^{-1}). \tag{16}$$

Testes de hipóteses (Teste de Wald)

Considere o teste de hipótese

$$H_0 : \beta_j = 0 \;\; \text{contra} \;\; H_1 : \beta_j \neq 0.$$

A estatística de Wald é definida por

$$Z = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \overset{a}{\sim} N(0, 1), \tag{17}$$

onde $\widehat{Var}(\hat{\beta}_j) = \widehat{Var}_{j,j}(\hat{\boldsymbol{\beta}}) = [(\boldsymbol{I}(\hat{\boldsymbol{\beta}}))^{-1}]_{j,j}$.

Assim, rejeita-se H_0 a um nível de $(1-\alpha)100\%$ de confiança se $|Z| > z_{1-\alpha/2}$.

Intervalos de confiança

De maneira geral, a estatística de Wald é a mais utilizada para construir intervalos de confiança $(1-\alpha)100\%$ assintóticos para cada um dos β_j's parâmetros.

Um intervalo de confiança $(1-\alpha)100\%$ para β_j é dado por

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\beta}_j)}, \tag{18}$$

onde $Var(\hat{\beta}_j) = \widehat{Var}(\hat{\boldsymbol{\beta}})_{j,j} = [(\boldsymbol{I}(\hat{\boldsymbol{\beta}}))^{-1}]_{j,j}$.

Teste da Razão de Verossimilhanças (TRV)

Suponha que deseja-se comparar dois modelos aninhados M_0 (modelo simplificado correspondente a H_0) e M_1 (modelo mais completo correspondente a H_1).

Seja $\hat{\boldsymbol{\pi}}_0$ os valores ajustados sob o modelo M_0 e $\hat{\boldsymbol{\pi}}_1$ os valores ajustados sob o modelo M_1.

Teste da Razão de Verossimilhanças (TRV)

Então, a estatística do TRV para comparar M_0 e M_1 é dada por

$$\begin{aligned} \Lambda &= 2[\ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y})] \\ &= 2\ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}) - 2\ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y}) \\ &= D_{M_0}(\hat{\boldsymbol{\pi}}) - D_{M_1}(\hat{\boldsymbol{\pi}}). \end{aligned} \tag{19}$$

Teste da Razão de Verossimilhanças (TRV)

A verossimilhança para um espaço menor M_0 não pode ser maior do que a verossimilhança sob um espaço maior M_1 isto é,

$$\ell(\hat{\boldsymbol{\pi}}_0|\boldsymbol{y}) \leq \ell(\hat{\boldsymbol{\pi}}_1|\boldsymbol{y}).$$

Então,

$$D_{M_1}(\hat{\boldsymbol{\pi}}) \leq D_{M_0}(\hat{\boldsymbol{\pi}}).$$

Assim, a estatística do TRV (19) é maior quando o modelo M_0 se ajusta mal aos dados quando comparado a M_1.

Exemplos

Exemplo 1.

Considere o conjunto de dados para o estado do Maine, nos EUA, com informações sobre renda per capita média e a localização de cada condado do estado.

O conjunto de dados está disponível na planilha *maine* do arquivo *dados_cdrl.xlsx*.

■

Exemplo 2.

A Tabela 2 mostra o número de besouros mortos após cinco horas de exposição a várias doses de dissulfeto de carbono gasoso.

O conjunto de dados está disponível na planilha *besouro* do arquivo *dados_cdrl.xlsx*.

Exemplos

Tabela 2: Dados de mortalidade de besouros.

Dose (x_i em $log_{10}CS2mgl^{-1}$)	Número de besouros (m_i)	Número de mortos (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Exemplos

Exemplo 3.

Uma pesquisadora está interessa em saber como as notas GRE (Graduate Record Exam scores), GPA (grade point average) e o prestígio da universidade onde candidatas/os cursaram a graduação afeta a admissão em um programa de pós-graduação.

O conjunto de dados está disponível na planilha *admissao* do arquivo *dados_cdrl.xlsx*.

Exemplo 4.

Considere o conjunto de dados de $n = 27$ pacientes com leucemia disponível na planilha *leucemia* do arquivo *dados_cdrl.xlsx*.

A variável resposta é binária e indica se ocorreu remissão da leucemia (REMISS), que é dada por um 1. As variáveis preditoras são celularidade da seção do coágulo da medula (CELL), porcentagem diferencial de esfregaço de blastos (SMEAR), porcentagem de infiltrado absoluto de células de leucemia da medula (INFIL), índice de marcação percentual das células de leucemia da medula óssea (LI), número absoluto de blastos no sangue periférico (BLAST) e a temperatura mais alta antes do início do tratamento (TEMP).

Regressão Logística no R

```r
library(readxl)
library(ggthemes)
library(ROCR)
library(caret)
library(tidymodels)
library(tidyverse)
```

Regressão Logística no R

```
dados <- read_xlsx("../dados/dados_cdrl.xlsx",
                   sheet = "diabetes")

dados <- dados |>
  mutate(diabetes = factor(diabetes),
         diab_bin = ifelse(diabetes == "neg", 0, 1))
```

Regressão Logística no R

```
fit <- glm(diab_bin ~ glucose,
           family = binomial(link = "logit"),
           data = dados)

tidy(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -6.10     0.630       -9.68 3.71e-22
## 2 glucose       0.0424   0.00476      8.91 5.07e-19
```

Regressão Logística no R

```r
summary(fit)
```

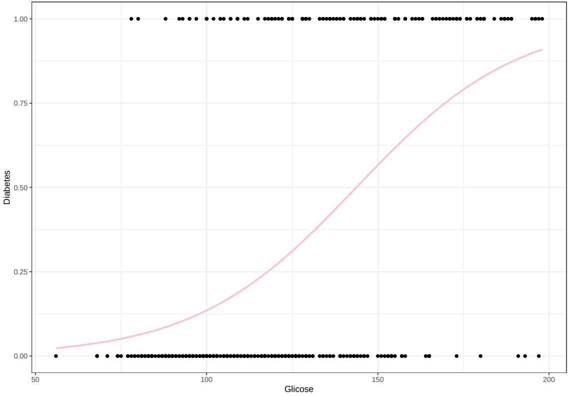
```
##
## Call:
## glm(formula = diab_bin ~ glucose, family = binomial(link = "logit"),
##     data = dados)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.095521   0.629787  -9.679   <2e-16 ***
## glucose      0.042421   0.004761   8.911   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 386.67  on 390  degrees of freedom
## AIC: 390.67
##
## Number of Fisher Scoring iterations: 4
```

Regressão Logística no R

```r
ggplot(dados, aes(x = glucose, y = diab_bin)) +
  geom_point() +
  stat_smooth(method = "glm", color = "pink",
              se = FALSE,
              method.args = list(family = binomial)) +
  labs(x = "Glicose", y = "Diabetes") +
  theme_bw()
```

Regressão Logística no R

```
## `geom_smooth()` using formula = 'y ~ x'
```



Regressão Logística no R

```r
# predições
probs <- predict(fit, type = "response")
pred_classe <- ifelse(probs < 0.5, 0, 1)

# acurácia
mean(pred_classe == dados$diab_bin)
```

```
## [1] 0.7678571
```

Regressão Logística no R

```r
cats <- factor(ifelse(probs < 0.5, "neg", "pos"))

conf_mat <- confusionMatrix(
  data = relevel(cats, ref = "pos"),
  reference = relevel(dados$diabetes, ref = "pos"))
```