

# Modelo linear generalizado

## Distribuição binomial

2023

Um modelo binomial com função de ligação canônica é ajustado aos dados do Exemplo 5, pag. 16 do livro Demétrio, C. G. B. (2002), *Modelos Lineares Generalizados em Experimentação Agronômica*, ESALQ (<https://docs.ufpr.br/~niveam/micro%20da%20sala/bom/Apostila%20de%20MLG.pdf>).

O modelo relaciona a probabilidade de que um inseto (pulgão) morra com a dose do inseticida (rotenone). Para cada dose foram utilizados  $m$  insetos no experimento dos quais  $y$  morreram após um certo tempo de exposição ao inseticida. Seis doses diferentes foram testadas. Esse tipo de estudo é conhecido como ensaio dose-resposta.

A linguagem R é utilizada no exemplo.

```
# Separador decimal nos resultados: ","  
options(OutDec = ",")
```

```
# Dados  
dose <- c(10.2, 7.7, 5.1, 3.8, 2.6, 0.0)  
m <- c(50, 49, 46, 48, 50, 49)  
y <- c(44, 42, 24, 16, 6, 0)  
resposta <- cbind(y, m - y)
```

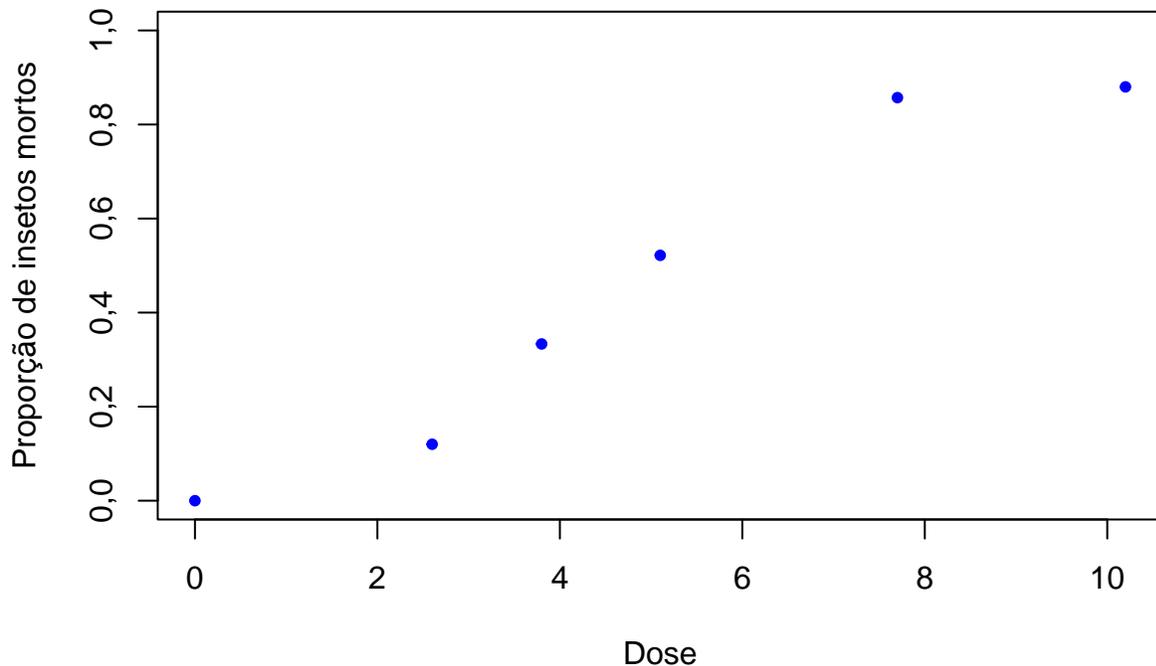
O objeto `resposta` representa a variável resposta e tem duas colunas: número de insetos mortos e número de insetos sobreviventes. O número de insetos no estudo foi

```
sum(m)
```

```
## [1] 292
```

Os dados neste exemplo estão na forma agregada. Na forma bruta teríamos um vetor de respostas binárias  $\mathbf{y}$  ( $292 \times 1$ ) e um vetor  $\mathbf{x}$  ( $292 \times 1$ ) com os valores das doses, sendo que 10,2 estaria repetido 50 vezes, ..., 0,0 estaria repetido 49 vezes.

```
# Gráfico de dispersão  
pmortos <- y / m  
plot(dose, pmortos, pch = 20, ylab = "Proporção de insetos mortos",  
      xlab = "Dose", ylim = c(0, 1), col = "blue")
```



```
# Ligação canônica (default)
mcan <- glm(resposta ~ dose, family = binomial)
summary(mcan)
```

```
##
## Call:
## glm(formula = resposta ~ dose, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3,22566    0,36992  -8,720  <2e-16 ***
## dose         0,60513    0,06781   8,923  <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: 163,745  on 5  degrees of freedom
## Residual deviance:  10,258  on 4  degrees of freedom
## AIC: 33,479
##
## Number of Fisher Scoring iterations: 5
```

O objeto mcan contém diversos resultados do ajuste.

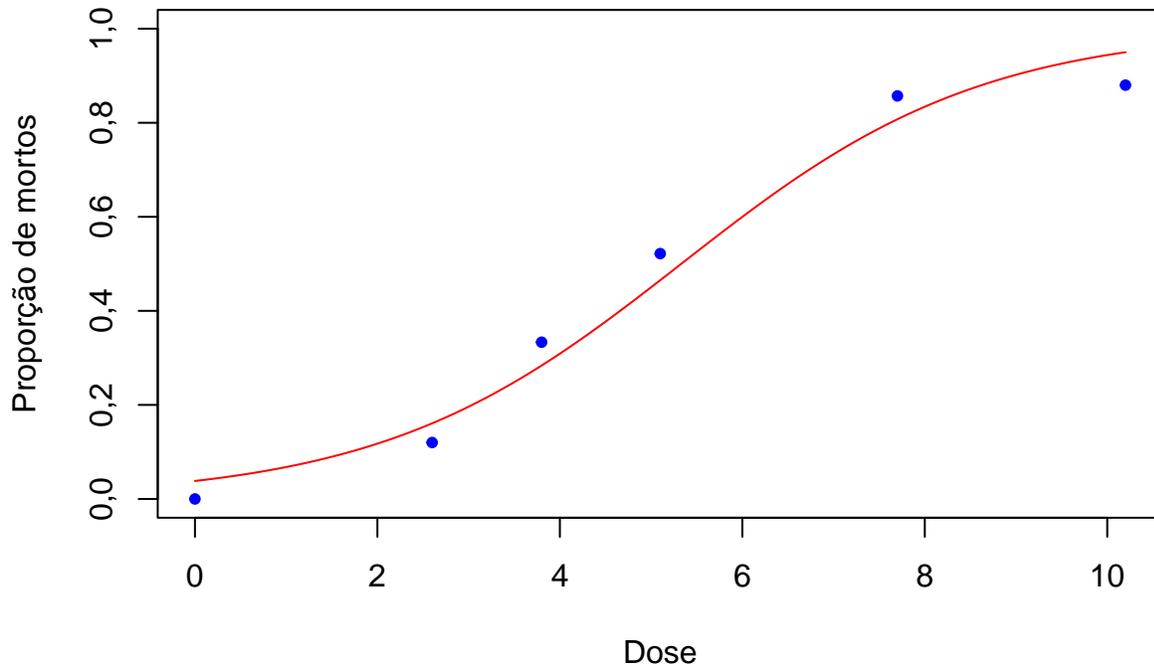
```
names(mcan)
```

```
## [1] "coefficients"      "residuals"        "fitted.values"
## [4] "effects"           "R"                 "rank"
## [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
```

```
## [22] "call"           "formula"         "terms"
## [25] "data"           "offset"          "control"
## [28] "method"        "contrasts"      "xlevels"
```

As proporções de insetos mortos e o modelo ajustado são mostrados na figura abaixo.

```
fprob <- function(x) {
  plogis(coef(mcan)[1] + coef(mcan)[2] * x)
}
plot(dose, prop, pch = 20, ylab = "Proporção de mortos",
      xlab = "Dose", ylim = c(0, 1), col = "blue")
curve(fprob, add = TRUE, col = "red")
```



```
## Intervalos de confiança assintóticos
# Aproximação pela dist. normal
conf <- 0.95 # coeficiente de confiança
z <- qnorm((1 + conf) / 2) # normal (z = 1,96)
smcan <- summary(mcan)
ep <- smcan$coefficients[, "Std. Error"]
icnorm <- cbind(mcan[["coefficients"]] - z * ep,
               mcan[["coefficients"]] + z * ep)
colnames(icnorm) <- c("Inf.", "Sup.")
print(icnorm, digits = 3)
```

```
##           Inf.   Sup.
## (Intercept) -3,951 -2,501
## dose         0,472  0,738
```

```
# Intervalos de confiança (veros. perfilada)
confint(mcan)
```

```
## Waiting for profiling to be done...
```

```
##           2,5 %   97,5 %
## (Intercept) -3,9952937 -2,540587
```

```
## dose      0,4807772  0,747802
```

## Funções de ligação

As funções de ligação logito (regressão logística), probito e complemento log-log são bastante utilizadas no modelo binomial. As expressões são dadas abaixo, sendo que  $\eta = \mathbf{x}^\top \boldsymbol{\beta} \in \mathbb{R}$  é o preditor linear,  $\pi = \mu/m$  e  $\Phi$  denota a função distribuição acumulada da distribuição normal(0, 1).

$$\log\left(\frac{\pi}{1-\pi}\right) = \eta, \quad \pi = \frac{e^\eta}{1+e^\eta} : \quad \text{logito}, \quad (1)$$

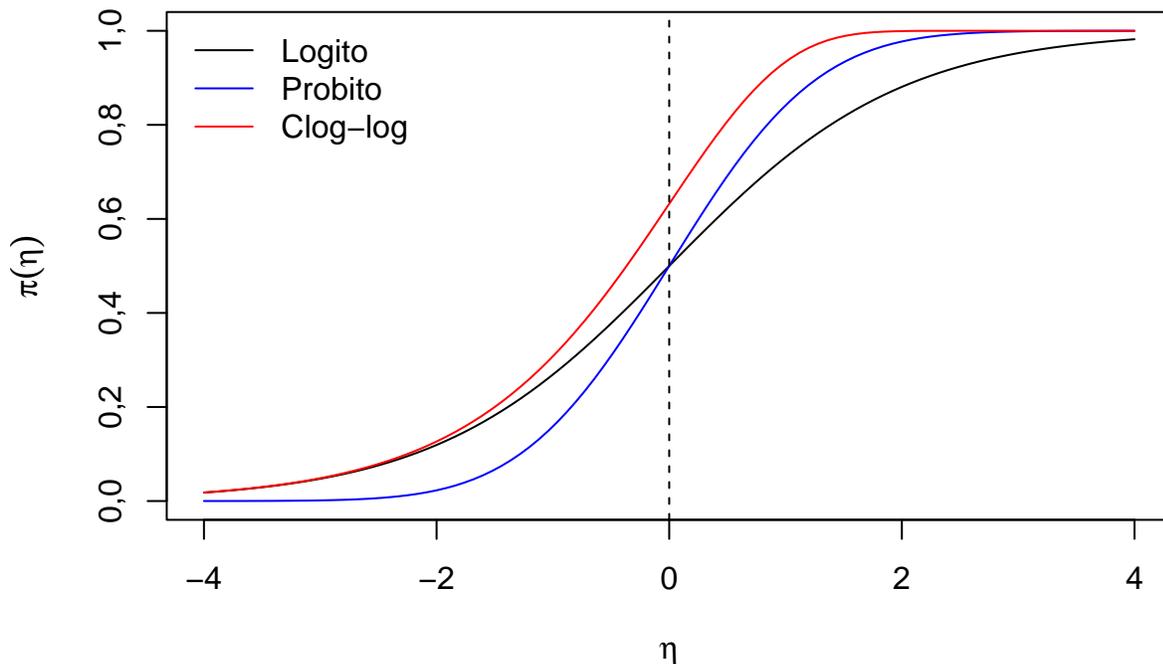
$$\Phi^{-1}(\pi) = \eta, \quad \pi = \Phi(\eta) : \quad \text{probito}, \quad (2)$$

$$\log(-\log(1-\pi)) = \eta, \quad \pi = 1 - \exp(-\exp(\eta)) : \quad \text{complemento log-log}. \quad (3)$$

As funções de ligação logito e probito são simétricas, significando que  $\pi(\eta) = 1 - \pi(-\eta)$  para todo  $\eta \in \mathbb{R}$ . Com a função de ligação complemento log-log, temos  $\pi(\eta) \geq 1 - \pi(-\eta)$  para todo  $\eta > 0$ .

```
## Inversa das funções de ligação
# Complemento log-log (Clog-log)
ploglog <- function(x) {
  return(1 - exp(-exp(x)))
}

curve(plogis(x), from = -4, to = 4, ylim = c(0, 1),
      xlab = expression(eta), ylab = expression(pi(eta)))
curve(pnorm(x), add = TRUE, col = "blue")
curve(ploglog(x), add = TRUE, col = "red")
abline(v = 0, lty = 2)
legend("topleft", c("Logito", "Probito", "Clog-log"), col = c("black", "blue", "red"),
      lty = 1, bty = "n")
```



**Nota 1.** Modelos com as funções de ligação probito e complemento log-log são ajustados com os comandos abaixo.

```
mprob <- glm(resposta ~ dose, family = binomial(link = "probit"))
mloglog <- glm(resposta ~ dose, family = binomial(link = "cloglog"))
```

## Razão de chances

Se  $\pi$  denota a probabilidade de um evento ocorrer,  $C = \pi/(1 - \pi)$  é a chance (*odds*) de que o evento ocorra, notando que  $C$  é uma função monótona crescente em  $\pi$ . Para as funções de ligação nas expressões (1)–(3), obtemos

$$C = \frac{e^\eta}{1 + e^\eta} = e^\eta \quad (4)$$

$$C = \frac{\Phi(\eta)}{1 - \Phi(\eta)} e$$

$$C = \frac{1 - \exp(-\exp(-\eta))}{\exp(-\exp(-\eta))}.$$

Considere  $\mathbf{x}_2 \neq \mathbf{x}_1$  (de modo que  $\eta_2 = \mathbf{x}_2^\top \boldsymbol{\beta} \neq \eta_1 = \mathbf{x}_1^\top \boldsymbol{\beta}$ ) com chances  $C_2$  e  $C_1$ . A razão de chances (*odds ratio*) é dada por  $RC = C_2/C_1$ . Tomando a função de ligação logito, da expressão (4) resulta

$$RC = \frac{e^{\eta_2}}{e^{\eta_1}} = e^{\eta_2 - \eta_1} = \exp((\mathbf{x}_2 - \mathbf{x}_1)^\top \boldsymbol{\beta}). \quad (5)$$

Retornando ao modelo binomial, temos  $Y \sim \text{binomial}(m, \pi)$  e  $E(Y) = \mu = m\pi$ . De acordo com o modelo e usando a função de ligação logito dada por (1), temos que

$$\log\left(\frac{\mu}{m - \mu}\right) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_1 + \beta_2 \text{dose}. \quad (6)$$

O lado esquerdo de (6) é chamado de logito da probabilidade  $\pi$ , que é representado por uma função linear dos coeficientes  $\boldsymbol{\beta}$ .

Em seguida comparamos dois valores de doses,  $\text{dose}_1$  e  $\text{dose}_2$ , sendo que  $\text{dose}_2 = \text{dose}_1 + 1$ . Logo,  $\mathbf{x}_1 = (1, \text{dose}_1)^\top$ ,  $\mathbf{x}_2 = (1, \text{dose}_2)^\top = (1, \text{dose}_1 + 1)^\top$  e  $\mathbf{x}_2 - \mathbf{x}_1 = (0, 1)^\top$ . A partir de (5) obtemos

$$RC = \exp((0, 1)\boldsymbol{\beta}) = \exp(0 \times \beta_1 + 1 \times \beta_2) = e^{\beta_2}.$$

Portanto, a chance de um inseto morrer é multiplicada por  $e^{\beta_2}$  quando a dose é aumentada em uma unidade, qualquer que seja a dose. Desta forma, o coeficiente da variável **dose** ( $\beta_2$ ) tem uma interpretação.

Em um modelo com mais de uma covariável ( $p > 2$ ), a interpretação de cada coeficiente  $\beta_j$ ,  $j \neq 1$ , é semelhante tomando-se a covariável  $x_j$  acrescida de uma unidade ( $x_j + 1$ ) e mantendo-se constantes os valores de todas as demais covariáveis.

```
# Razão de chances
coef(mcan)
```

```
## (Intercept)      dose
## -3,2256633      0,6051256
```

```
cat("\n Razão de chances =", exp(coef(mcan)[2]))
```

```
##
## Razão de chances = 1,831482
```

**Nota 2.** Apresente um intervalo de confiança para a razão de chances.

## Dose letal

É de interesse estimar as doses de inseticida que eliminam 50% e 90% dos insetos, denotadas por  $DL_{50}$  e  $DL_{90}$ , respectivamente. Segue de (6) que

$$\text{dose} = \frac{\log\left(\frac{\pi}{1-\pi}\right) - \beta_1}{\beta_2}. \quad (7)$$

Da expressão (7) obtemos

$$DL_{50} = \frac{\log\left(\frac{0,5}{1-0,5}\right) - \beta_1}{\beta_2} = -\frac{\beta_1}{\beta_2} \quad (8)$$

e

$$DL_{90} = \frac{\log\left(\frac{0,9}{1-0,9}\right) - \beta_1}{\beta_2} = \frac{2,20 - \beta_1}{\beta_2}. \quad (9)$$

```
# Dose letal
coef(mcan)

## (Intercept)      dose
## -3,2256633      0,6051256

DL <- function(prob) {
  return((log(prob / (1 - prob))) - coef(mcan)[1]) / coef(mcan)[2])
}

cat("\n DL50 e DL90 = ", DL(c(0.5, 0.9)))

##
## DL50 e DL90 = 5,330569 8,961591
```

Aplicamos o método delta para obter um intervalo de confiança de 95% para a dose letal. Calculamos

$$\frac{\partial}{\partial \beta} DL = \left( -\frac{1}{\beta_2}, -\frac{\log\left(\frac{\pi}{1-\pi}\right) - \beta_1}{\beta_2^2} \right)^\top = -\frac{1}{\beta_2} (1, DL)^\top$$

e

$$\text{cov}(DL) = \frac{\partial}{\partial \beta^\top} DL \text{cov}(\hat{\beta}) \frac{\partial}{\partial \beta} DL,$$

sendo que um estimador de  $\text{cov}(\hat{\beta})$  é dado pela função `vcov`.

```
# IC de 95% para DL50 (método delta)
pDL <- 0.5
derivDL <- -c(1, DL(pDL)) / coef(mcan)[2]
epDL <- sqrt(t(derivDL) %*% vcov(mcan) %*% derivDL)
cat("\n IC DL", pDL * 100, ":", c(DL(pDL) - z * epDL, DL(pDL) + z * epDL))

##
## IC DL 50 : 4,809176 5,851961
```

**Nota 3.** Refaça o exemplo em linguagem Python.