SME0823 - Modelos Lineares Generalizados - Lista 4

Danilo Augusto Ganancin Faria – Nº USP: 9609172

20 de dezembro de 2019

Exercício 1

Os dados deste exercício correspondem a um experimento de dose-resposta conduzido para avaliar a influência do extrato vegetal "aquoso frio de folhas" na morte de um determinado tipo de mosquito.

Foram encontramos as seguintes variáves: `dose`, mosquitos expostos (`n`) e mosquitos mortos (`y`), que é nossa variável reposta.

Para realizar a leitura dos dados utilizou-se o seguinte código.

```
# Código para ler os dados
dados <- scan("http://www.ime.usp.br/~giapaula/dose1.dat", what = list(dose = 0,
    n = 0, y = 0))
```

Análise de dados preliminar

Tabela 1: Variáveis dose e n

Variável	
dose	0 15 20 25 30 35 40
n	50 50 50 50 50 50 50

Nota-se na Tabela 1 que a `dose` iniciou em 0, depois teve um salto para 15 e em seguida teve um acréscimo de 5 em 5 até chegar no limite de 40 (unidades de medida). A quantidade de mosquitos utilizada neste experimento para cada dose foi constante igual a 50 mosquitos.

```
summary(dados$y)
sd(dados$y)
var(dados$y)

# Cálculo da proporção de mosquitos mortos por dose
prop <- (dados$y)/(dados$n)
prop
```

Tabela 2: Medidas resumo da variável resposta y

Variável	Min.	Med.	Máx.	Média	DP	Var.
y	4	29	47	25,43	17,73	314,29

De acordo com a Tabela 2, é possível observar que a quantidade média de mosquitos mortos é de aproximadamente 26, e também que o número mínimo e máximo de mortes provocadas pelo extrato vegetal "aquoso frio de folhas" são de 4 e 47 mosquitos, respectivamente.

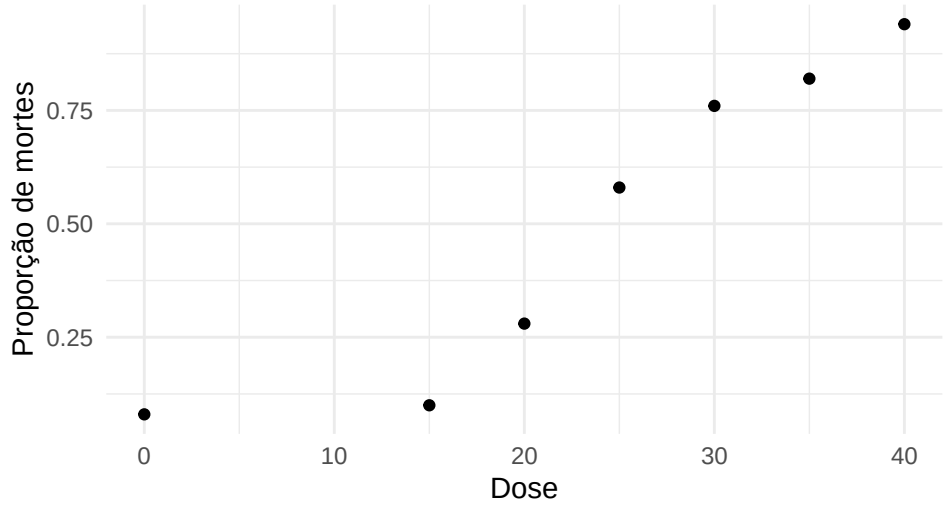

Figura 1: Gráficos de pontos da proporção observada de mosquitos mortos por dose

Pelo gráfico da Figura 1, temos uma indicação de que o modelo logístico pode ser apropriado para realizar o ajuste do modelo linear generalizado.

Ajustes dos modelos lineares generalizados

```
# Matriz X
matX <- cbind(dados$y, dados$n - dados$y)
matX
```

Acima foi criado uma matriz intitulada `matX` em que sua primeira coluna contém o número de mosquitos mortos e a segunda o número de mosquitos sobreviventes em relação a cada uma das doses.

Por se tratar de um modelo com resposta binária, ou seja, que admite apenas dois resultados, ajustou-se o modelo supondo uma distribuição Binomial para `y`.

```
# Ajuste do modelo MLG com função de ligação logito
fit.mod1 <- glm(matX ~ dados$dose, family = binomial(link = "logit"))

# Ajuste do modelo MLG com função de ligação probito
fit.mod2 <- glm(matX ~ dados$dose, family = binomial(link = "probit"))

# Ajuste do modelo MLG com função de ligação cloglog
fit.mod3 <- glm(matX ~ dados$dose, family = binomial(link = "cloglog"))
```

```r
# Ajuste do modelo MLG com função de ligação loglog
loglog <- function() structure(list(linkfun = function(mu) -log(-log(mu)), linkinv = function(eta) pmax
    1 - .Machine$double.eps), .Machine$double.eps), mu.eta = function(eta) {
    eta <- pmin(eta, 700)
    pmax(exp(-eta - exp(-eta)), .Machine$double.eps)
}, dmu.deta = function(eta) pmax(exp(-exp(-eta) - eta) * expm1(-eta), .Machine$double.eps),
    valideta = function(eta) TRUE, name = "loglog"), class = "link-glm")

fit.mod4 <- glm(matX ~ dados$dose, family = binomial(link = loglog()))

# Ajuste do modelo MLG com função de ligação cauchito
fit.mod5 <- glm(matX ~ dados$dose, family = binomial(link = "cauchit"))

# Resumo das medidas dos modelos MLG ajustados
summary(fit.mod1)
summary(fit.mod2)
summary(fit.mod3)
summary(fit.mod4)
summary(fit.mod5)

# Data frame com a proporção observada e com os valores ajustados de cada
# modelo MLG
df <- data.frame(prop, round(fit.mod1$fitted.values, 2), round(fit.mod2$fitted.values,
    2), round(fit.mod3$fitted.values, 2), round(fit.mod4$fitted.values, 2),
    round(fit.mod5$fitted.values, 2))
colnames(df) = c("prop", "fit.mod1", "fit.mod2", "fit.mod3", "fit.mod4", "fit.mod5")
df
```

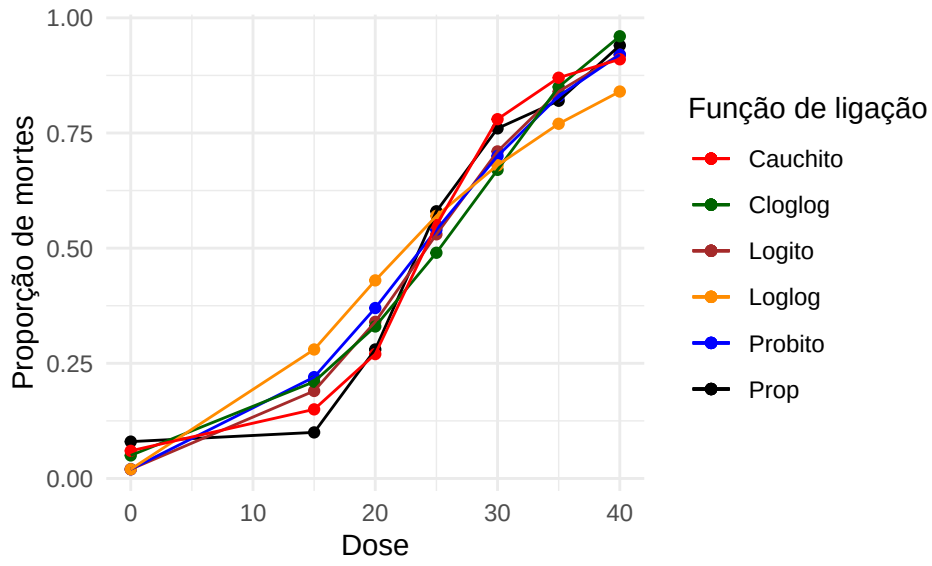

Figura 2: Gráficos de pontos e linhas da proporção observada de mosquitos mortos por dose de cada modelo ajustado

Na Figura 2, tem-se o gráfico de pontos e linhas da proporção de mortes observadas e dos valores ajustados de cada modelo com sua respectiva função de ligação.

O modelo com curva que melhor se aproxima do gráfico da proporção de mortes observadas, é o que utilizou como função de ligação a função `cauchito`. Temos um indício de que este modelo seja o que melhor se ajusta aos dados dentre todos os outros ajustados.

Diagnóstico dos modelos ajustados

- AIC e desvio residual

```
# Critério AIC para cada um dos MLG ajustados
AIC(fit.mod1)
AIC(fit.mod2)
AIC(fit.mod3)
AIC(fit.mod4)
AIC(fit.mod5)
```

Tabela 3: Critério AIC e desvio residual dos modelos MLG ajustados

Função de ligação	AIC	Desvio residual
Logito	40,10	10,18
Probito	43,79	13,87
Cloglog	39,84	9,93
Loglog	58,31	28,39
Cauchito	33,29	3,38

Uma das formas de se avaliar um modelo é calculando seu critério de informação de Akaike (AIC), sabe-se

quanto menor é seu valor, melhor é o ajuste do modelo. Uma medida alternativa que pode ser usada para complementar o uso do AIC é o Desvio residual, quanto menor seu valor, melhor é o ajuste do modelo. De acordo com as medidas apresentadas na Tabela 3, o melhor modelo ajustado é aquele em que utilizou-se como função de ligação a `cauchito`.

- Gráfico de envelope

Esta metodologia é utilizada com o intuito de verificar possíveis afastamentos das suposições feitas para o modelo, em especial para o componente aleatório e para a parte sistemática bem como a existência de observações discrepantes com alguma interferência desproporcional ou inferencial nos resultados dos ajustes.

```
# Envelope para o MLG com função de ligação logito
ntot <- dados$n

fit.model <- fit.mod1
source("http://www.ime.usp.br/~giapaula/envelr_bino")
```

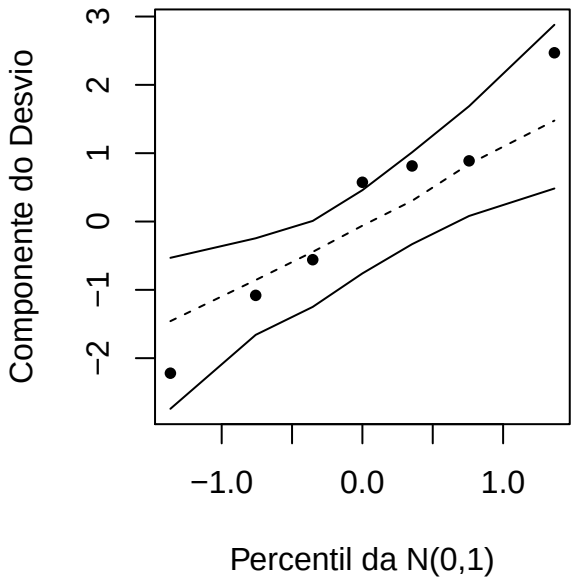

Figura 3: Gráfico de envelope para o modelo com função de ligação logito

Pode-se observar na Figura 3 que apenas uma observação está fora do envelope, neste caso, pode ser um indicativo de que o modelo não está totalmente bem ajustado.

```
# Envelope para o MLG com função de ligação probito
fit.model <- fit.mod2
source("http://www.ime.usp.br/~giapaula/envelr_bino")
```

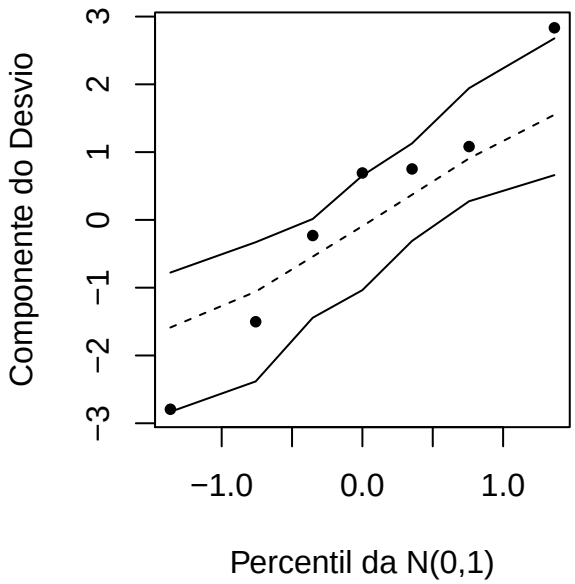

Figura 4: Gráfico de envelope para o modelo com função de ligação probito

Assim como na Figura 3, a Figura 4 mostra uma observação fora do envelope, é um indicativo de que o modelo não foi ajustado de forma correta.

```
# Envelope para o MLG com função de ligação cloglog
fit.model <- fit.mod3
source("http://www.ime.usp.br/~giapaula/envelr_bino")
```

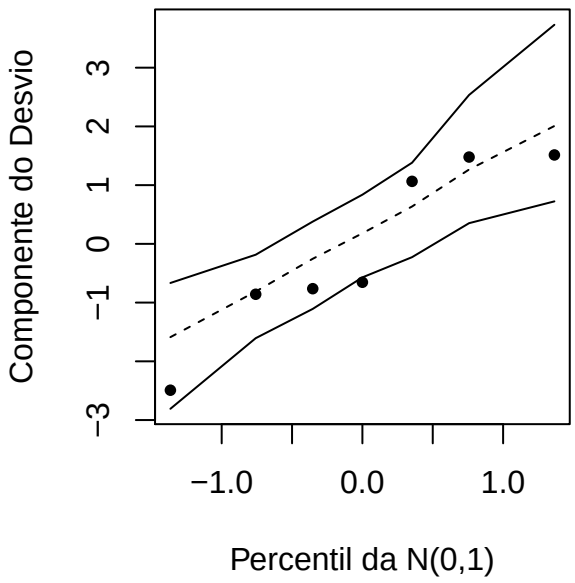

Figura 5: Gráfico de envelope para o modelo com função de ligação cloglog

Da mesma forma, na Figura 5, uma observação está fora do envelope, pode ser que o modelo não está ajustado corretamente.

```
# Envelope para o MLG com função de ligação loglog
fit.model <- fit.mod4
source("http://www.ime.usp.br/~giapaula/envelr_bino")
```

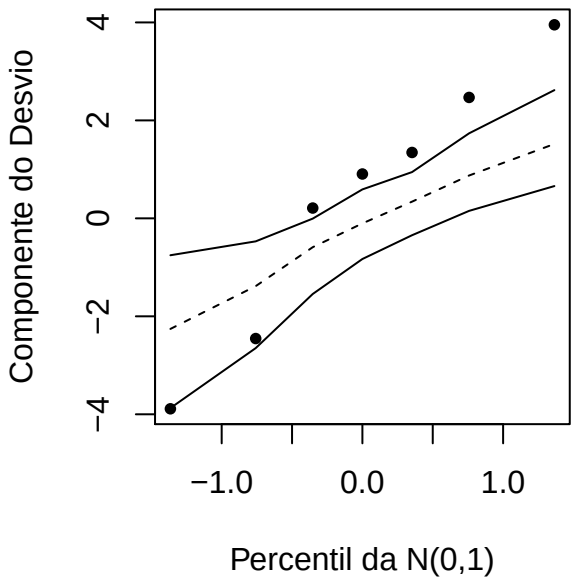

Figura 6: Gráfico de envelope para o modelo com função de ligação loglog

Na Figura 6, é possível constatar que seis observações estão fora do envelope, significa que este modelo não fez um bom ajuste, pode ser que alguma das suposições feitas a cerca do modelo foi violada, por exemplo, a suposição de que a variável resposta y tem distribuição Binomial.

```
# Envelope para o MLG com função de ligação cauchito
fit.model <- fit.mod5
source("http://www.ime.usp.br/~giapaula/envelr_bino")
```

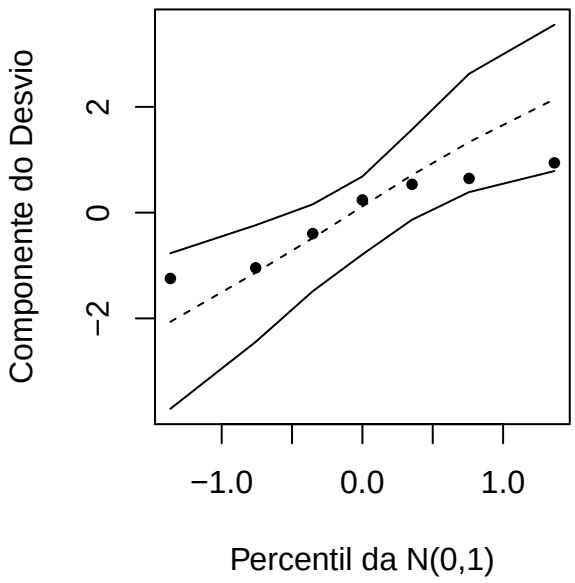

Figura 7: Gráfico de envelope para o modelo com função de ligação cauchito

Todas as observações encontram-se dentro do envelope, conforme consta na Figura 7, isto significa que o modelo foi ajustado de forma adequada.

- Gráficos de diagnóstico

Nas figuras a seguir são apresentados os gráficos de diagnóstico. O gráfico da Medida h *versus* Valor ajustado permite identificar os pontos de alavanca. Com o gráfico da Distância de Cook *versus* Índice é possível identificar os pontos influentes. Já o gráfico do Resíduo Componente do Desvio *versus* Índice, identifica os chamados pontos aberrantes ou *outliers*. E, por fim, o gráfico do Resíduo Componente do Desvio *versus* Valor ajustado avalia a função de ligação escolhida.

```
# Diagnóstico para o MLG com função de ligação logito
attach(dados)

fit.model <- fit.mod1
source("http://www.ime.usp.br/~giapaula/diag_bino")
```

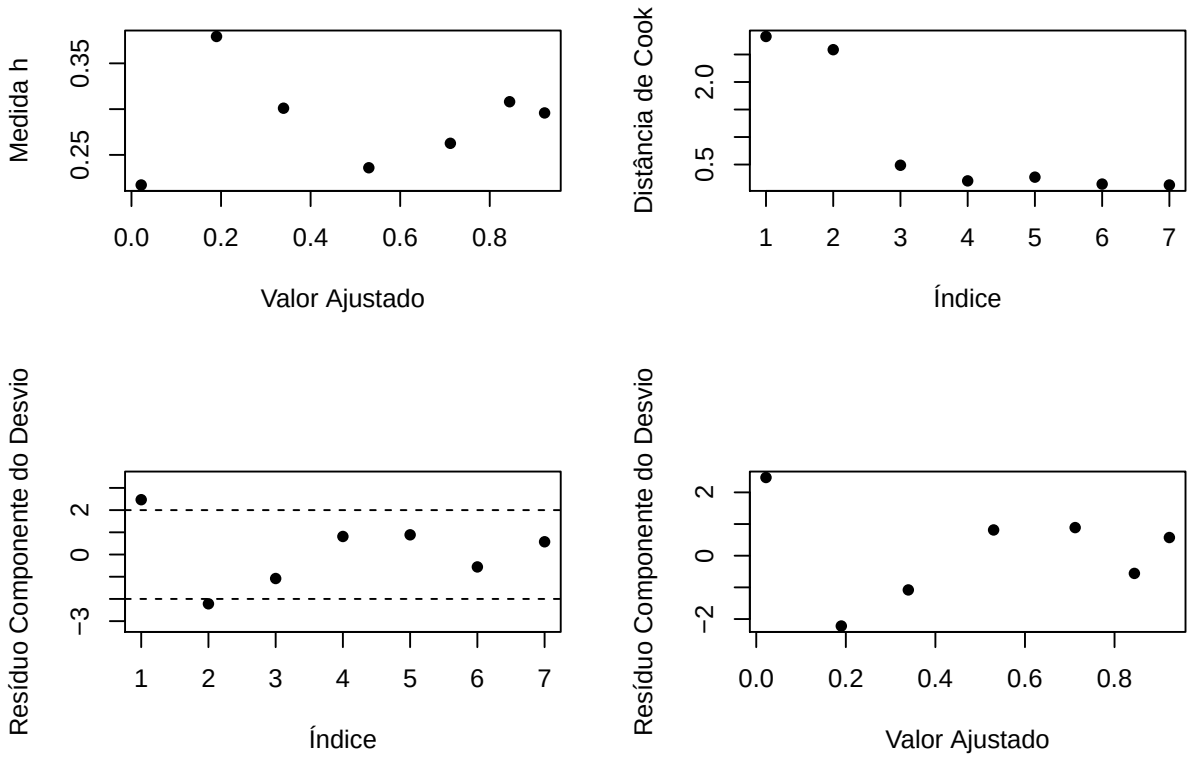

Figura 8: Gráfico de diagnóstico para o modelo com função de ligação logito

Na Figura 8, as observações de números 1 e 2 requerem maior atenção, uma vez que pelos gráficos de diagnósticos, tais observações podem ser consideradas como pontos de alavanca, pontos influentes e também pontos aberrantes.

```
# Diagnóstico para o MLG com função de ligação probito
fit.model <- fit.mod2
source("http://www.ime.usp.br/~giapaula/diag_bino")
```

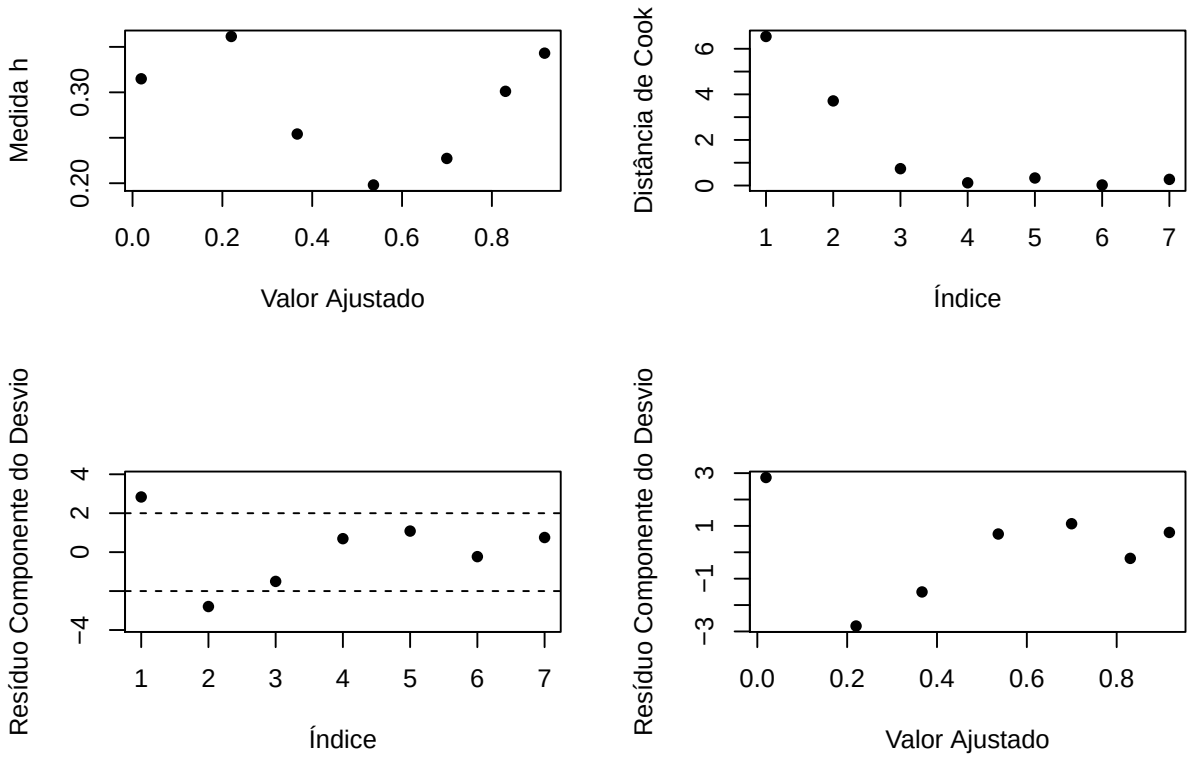

Figura 9: Gráfico de diagnóstico para o modelo com função de ligação probito

Os gráficos de diagnósticos da Figura 9 apontam que a observação de número 1 pode ser considerada como um ponto influente e aberrante. Já a observação de número 2, como sendo ponto de alavanca e aberrante.

```
# Diagnóstico para o MLG com função de ligação clolog
fit.model <- fit.mod3
source("http://www.ime.usp.br/~giapaula/diag_bino")
```

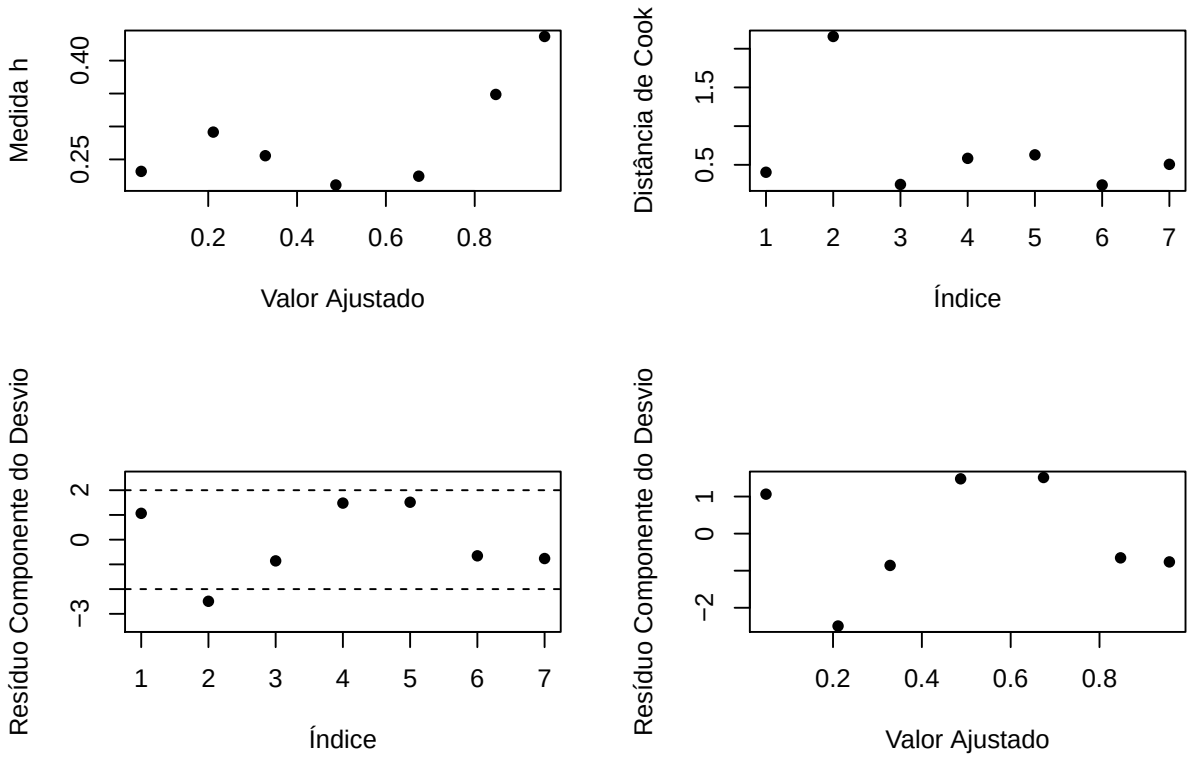

Figura 10: Gráfico de diagnóstico para o modelo com função de ligação cloglog

Na Figura 10, a observação de número 7 pode ser considerada um ponto de alavanca e a observação 2 como sendo um ponto influente e aberrante.

```
# Diagnóstico para o MLG com função de ligação loglog
fit.model <- fit.mod4
source("http://www.ime.usp.br/~giapaula/diag_bino")
```

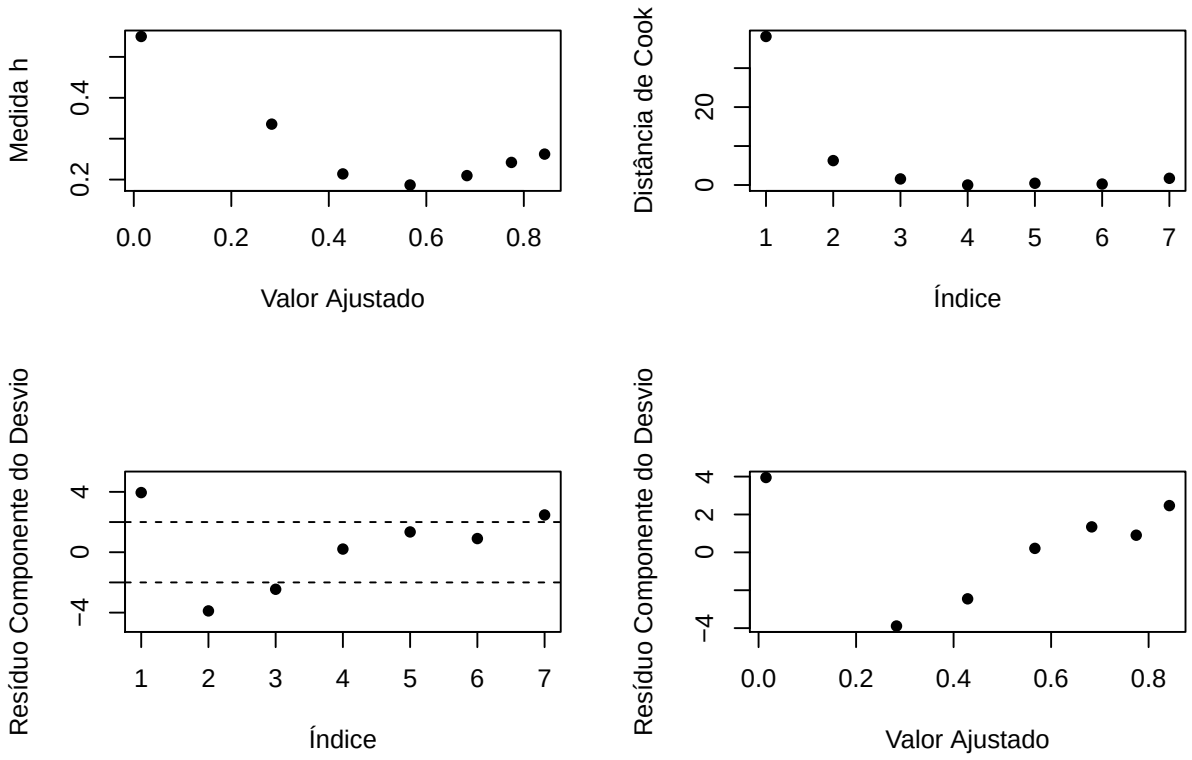

Figura 11: Gráfico de diagnóstico para o modelo com função de ligação loglog

De acordo com os gráficos de diagnósticos da Figura 11, a observação de número 1 pode ser considerada como pontos de alavanca, influente e também aberrante, além disso as observações de números 2, 3 e 7 como sendo pontos aberrantes.

```
# Diagnóstico para o MLG com função de ligação cauchito
fit.model <- fit.mod5
source("http://www.ime.usp.br/~giapaula/diag_bino")
```

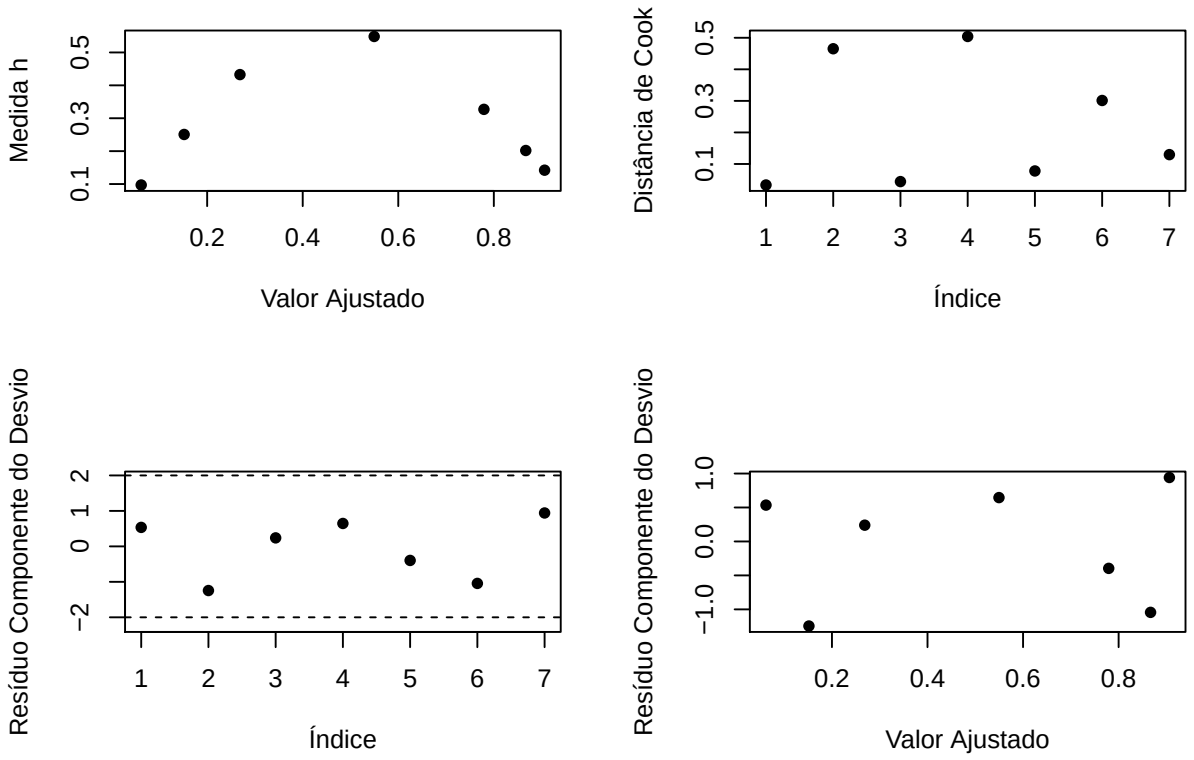

Figura 12: Gráfico de diagnóstico para o modelo com função de ligação cauchito

A Figura 12 apresenta os gráficos de diagnósticos em que é possível notar que a observação de número 4 pode ser apontada como pontos de alavanca e influente.

Conclusão

Após realizar as análises, o melhor modelo, de acordo com as medidas AIC, Desvio residual, análise de envelope e de diagnóstico, é o modelo MLG em que se utilizou como função de ligação a `cauchito`.

O modelo estimado para predizer a proporção de mosquitos mortos é dado por:

$$\hat{p_i} = \frac{1}{2} + \frac{\arctan(-5,074 + 0,209 * dose_i)}{\pi}.$$

Modelo final escolhido:

- Componente aleatório: $y_1, \ldots, y_7$ a.a. de $Y_i \sim Binomial(n, p_i)$, $i = 1, \ldots, 7$.
- Componente sistemático: $\eta_i = -5,074 + 0,209 * dose_i$, $i = 1, \ldots, 7$.
- Função de ligação: $g(\mu_i) = \eta_i = \tan\Big(\pi\Big(p_i - \frac{1}{2}\Big)\Big)$.

Exercício 2

Os dados deste exercício são dados na forma binária correspondentes a um estudo sobre doença coronária segundo diferentes variáveis. Foram avaliados 78 pacientes, identificou-se as seguintes variáveis: `dc`: doença coronária 1: possui, 0: não possui; `sexo`: 1: masculino e 0: feminino; `ecg`: nível de gravidade no exame do eletrocardiograma 0: nível baixo, 1: nível médio, 2: nível alto e, por fim, `idade`: idades entre 28 e 63 anos.

Análise da dados preliminar

```
# Código para fazer a leitura dos dados
dados2 <- read.table("http://www.uel.br/pessoal/silvano/Dados/chd4a.txt", header = T)
dados2 <- data.frame(dados2)
```

Tabela 4: Medidas resumo da variável idade

Variável	Min.	Med.	Máx.	Média	DP	Var.
idade	28	46,5	63	46,9	8,55	73,13

De acordo com a Tabela 4, é possível observar que a idade média dos pacientes é de aproximadamente 47 anos, e também que a idade mínima e máxima são de 28 e 63 anos, respectivamente.

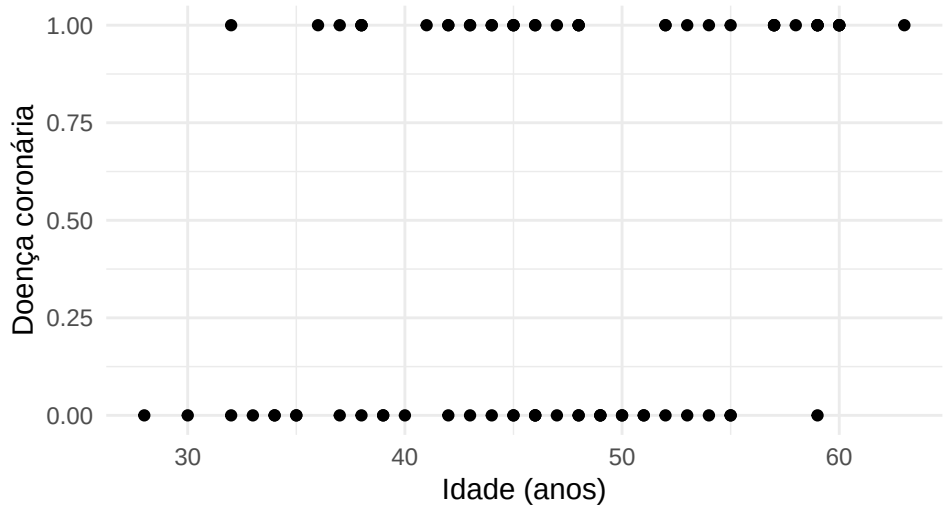

Figura 13: Gráfico de pontos da doença coronária pela idade

Com o intuito de ilustrar e comprovar que os dados deste exercício estão na forma binária, foi feita a Figura 13.

Ajustes dos modelos lineares generalizados

```
# Ajuste do modelo MLG com função de ligação logito
mod1 <- glm(dc ~ sexo + ecg + idade, data = dados2, family = binomial(link = "logit"))

# Ajuste do modelo MLG com função de ligação probito
mod2 <- glm(dc ~ sexo + ecg + idade, data = dados2, family = binomial(link = "probit"))

# Ajuste do modelo MLG com função de ligação cauchito
mod3 <- glm(dc ~ sexo + ecg + idade, data = dados2, family = binomial(link = "cauchit"))

# Ajuste do modelo MLG com função de ligação cloglog
mod4 <- glm(dc ~ sexo + ecg + idade, data = dados2, family = binomial(link = "cloglog"))

# Ajuste do modelo MLG com função de ligação loglog
loglog <- function() structure(list(linkfun = function(mu) -log(-log(mu)), linkinv = function(eta) pmax
    1 - .Machine$double.eps), .Machine$double.eps), mu.eta = function(eta) {
    eta <- pmin(eta, 700)
    pmax(exp(-eta - exp(-eta)), .Machine$double.eps)
}, dmu.deta = function(eta) pmax(exp(-exp(-eta) - eta) * expm1(-eta), .Machine$double.eps),
    valideta = function(eta) TRUE, name = "loglog"), class = "link-glm")

mod5 <- glm(dc ~ sexo + ecg + idade, data = dados2, family = binomial(link = loglog()))

# Resumo das medidas dos modelos MLG ajustados
summary(mod1)
summary(mod2)
summary(mod3)
summary(mod4)
summary(mod5)
```

- Estatística de Hosmer-Lemeshow

A estatística de Hosmer-Lemeshow é utilizada com a finalidade de testar a bondade do ajuste, ou seja, comprova se o modelo proposto pode explicar de maneira correta o que se observa.

```
# Estatística Hosmer-Lemeshow
source("http://www.poleto.com/funcoes/gof.bino.txt")
attach(dados2)

ajuste <- mod1
hlmod1 <- gof.bino(ajuste, grupos = 10)$pvalue

ajuste <- mod2
gof.bino(ajuste, grupos = 10)
```

```r
hlmod2 <- gof.bino(ajuste, grupos = 10)$pvalue

ajuste <- mod3
gof.bino(ajuste, grupos = 10)
hlmod3 <- gof.bino(ajuste, grupos = 10)$pvalue

ajuste <- mod4
gof.bino(ajuste, grupos = 10)
hlmod4 <- gof.bino(ajuste, grupos = 10)$pvalue

ajuste <- mod5
gof.bino(ajuste, grupos = 10)
hlmod5 <- gof.bino(ajuste, grupos = 10)$pvalue
```

- Curva ROC

```r
# Curva ROC
r.bp <- Roc(list(mod1, mod2, mod3, mod4, mod5), data = dados2)

print(r.bp)

AUC1 <- unlist(unclass(r.bp$Auc[[1]]))
AUC2 <- unlist(unclass(r.bp$Auc[[2]]))
AUC3 <- unlist(unclass(r.bp$Auc[[3]]))
AUC4 <- unlist(unclass(r.bp$Auc[[4]]))
AUC5 <- unlist(unclass(r.bp$Auc[[5]]))
```

```r
# Curva ROC para os modelos MLG ajustados
plot(r.bp, auc = TRUE, text.col = c(1, 2, 3, 4, 5), ylab = "Sensitividade",
    xlab = "1 - Especificidade")
```

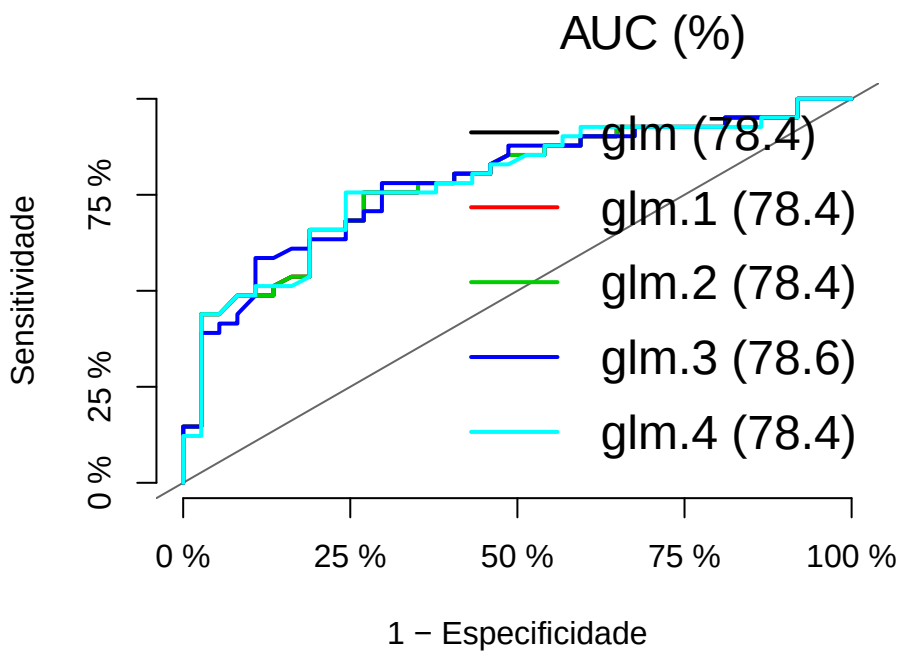

Figura 14: Curva ROC para os modelos MLG ajustados

Tabela 5: Critério AIC, desvio residual, estatística de Hosmer-Lemeshow e área abaixo da curva ROC dos modelos MLG ajustados

Função de ligação	AIC	Desvio residual	Hosmer-Lemeshow	AUC
Logito	94,81	86,81	0,67	0,78
Probito	94,79	86,79	0,65	0,78
Cauchito	94,99	86,99	0,75	0,78
Cloglog	94,00	86,00	0,62	0,79
Loglog	96,16	88,16	0,34	0,78

Conforme é possível observar na Tabela 5, o modelo que melhor se ajustou aos dados foi o que utilizou como função de ligação a `cloglog`. Este modelo possui o menor AIC, menor Desvio residual, estatística de Hosmer-Lemeshow de 0,62 e AUC de maior valor igual a 0,79.

```
# Envelope para o MLG com função de ligação cloglog
fit.model <- mod4
source("http://www.ime.usp.br/~giapaula/envel_bino")
```

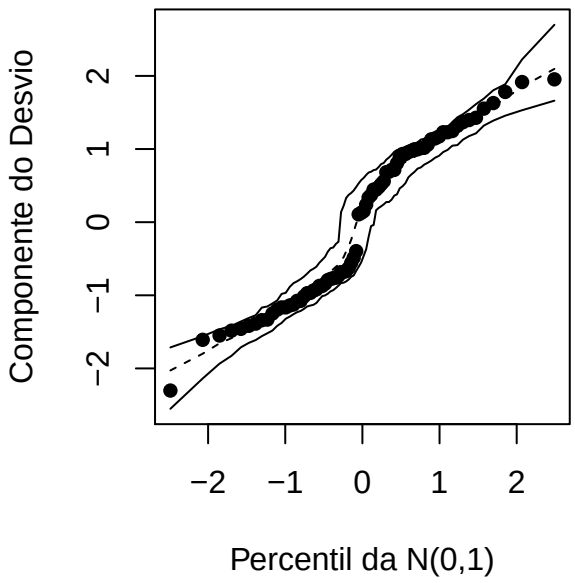

Figura 15: Gráfico de envelope para o modelo com função de ligação cloglog

O gráfico de envelope apresentado na Figura 15 mostra que todas as observações encontram-se entre as faixas de confiança do envelope simulado, isso significa que o modelo foi ajustado de forma correta.

```
# Diagnóstico para o MLG com função de ligação clolog
fit.model <- mod4
source("https://www.ime.usp.br/~giapaula/diag_bino")
```

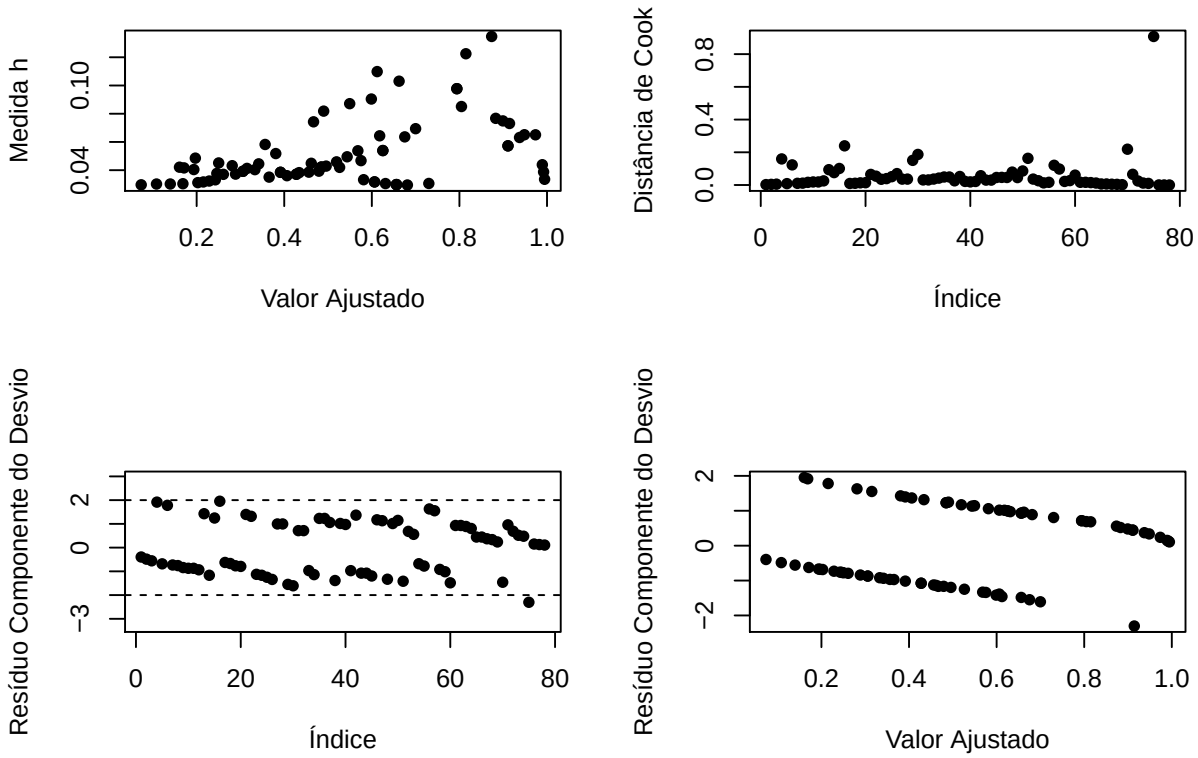

Figura 16: Gráfico de diagnóstico para o modelo com função de ligação cloglog

De acordo com os gráficos de diagnóstico da Figura 16, temos uma observação classificada como ponto de alavanca, outra como ponto influente e uma última classificada como ponto aberrante.

Conclusão

Após realizar as análises, o melhor modelo, de acordo com as medidas AIC, estatística de Hosmer-Lemeshow, curva ROC, análise de envelope e de diagnóstico, é o modelo MLG em que se utilizou como função de ligação a `cloglog`.

O modelo estimado para predizer se uma pessoa possa ter doença cardíaca em função da idade média é dado por:

$$\hat{p_i} = 1 - \exp\{-\exp\{-4,481 + 0,92 * sexo_i + 0,561 * ecg_i + 0,07 * idade_i\}\}.$$

Modelo final escolhido:

- Componente aleatório: $y_1, \ldots, y_{78}$ a.a. de $Y_i \sim Bernoulli(p_i)$, $i = 1, \ldots, 78$.
- Componente sistemático: $\eta_i = -4,481 + 0,92 * sexo_i + 0,561 * ecg_i + 0,07 * idade_i$, $i = 1, \ldots, 78$.
- Função de ligação: $g(\mu_i) = \eta_i = \log(-\log(1 - p_i))$, $i = 1, \ldots, 78$.