# Modelos de regressão para dados discretos (parte 1): dados binários

Prof. Caio Azevedo

# Motivação

- As metodologias, incluindo os modelos de regressão, vistas até agora, são apropriadas para análise de dados categorizados.
- Veremos como analisar situações nas quais a variável resposta é discreta mas os dados não estão categorizados (e as vezes não podem ser categorizados).

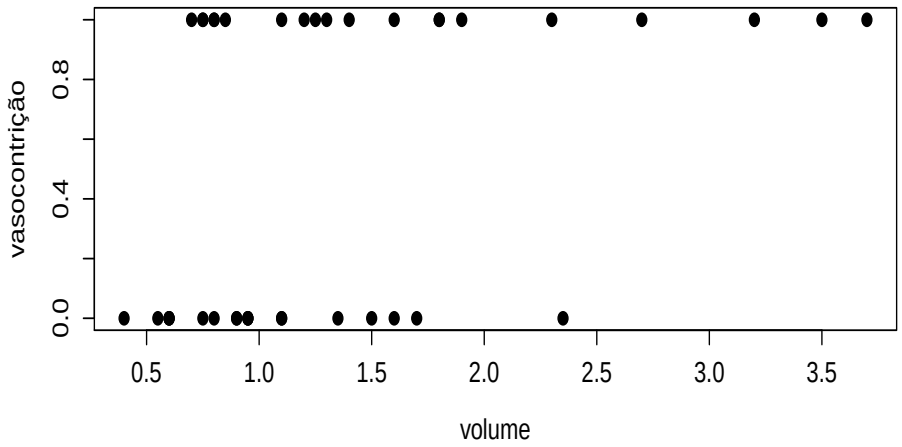
# Exemplo 11: Estudo sobre vasoconstrição

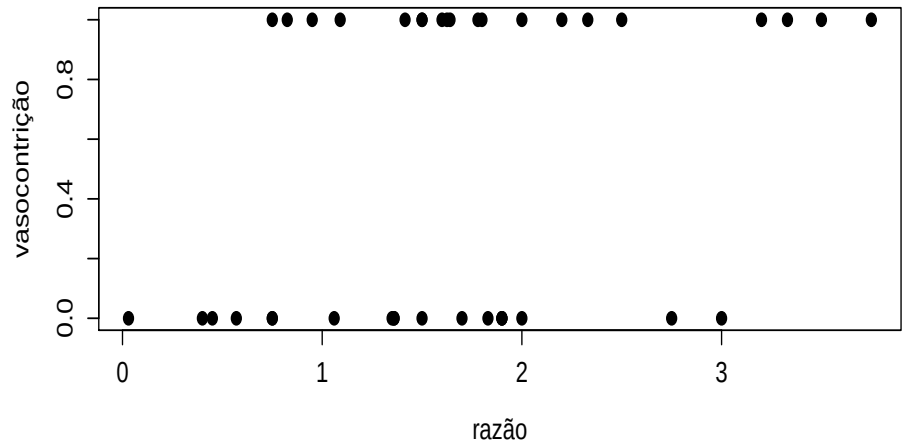
- Consideraremos os dados sobre um estudo de vasoconstrição (veja Paula (2013) e referências nela constantes).
- Nesse estudo, foram medidos de 3 pacientes o volume e a razão de ar inspirado, como também a ocorrência ou não de vasoconstrição (contração de vasos sanguíneos) na pele dos dedos da mão. O primeiro paciente contribuiu com 9 observações, o segundo com 8 e o terceiro com 22.

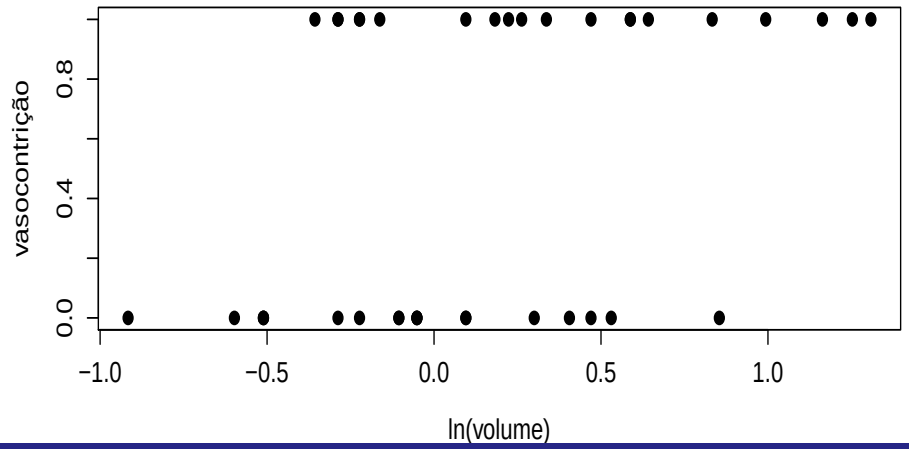
# Exemplo 11 (cont.)

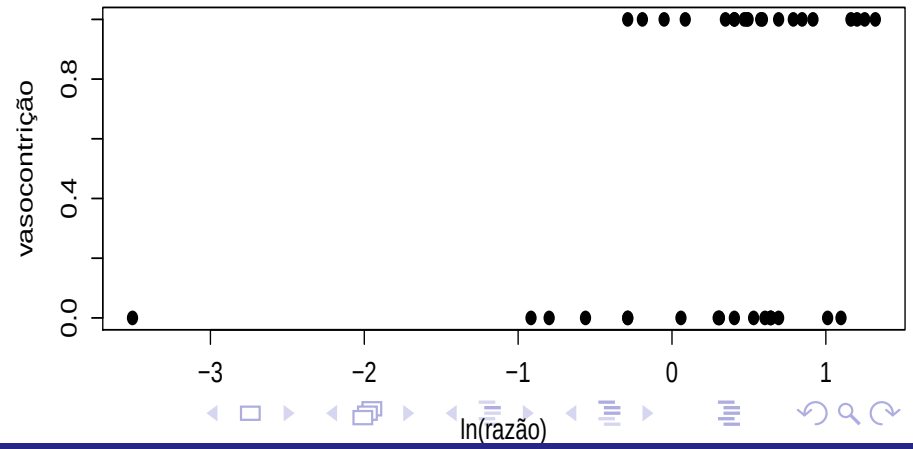
- Objetivo: verificar se a quantidade de ar (volume e razão, variáveis explicativas) influenciam a ocorrência de vasoconstrição (resposta).
- Seja $Y_i$ a variável aleatória que assume valor 1, se ocorreu vasocontrição no $i-$ésimo paciente e 0, caso contrário.
- As vezes é mais apropriado trabalhar como o ln (logaritmo natural) das variáveis explicativas (para, por exemplo, medir melhor o impacto de cada uma na variável resposta, principalmente se esta não for contínua).

# Gráficos de dispersão individuais

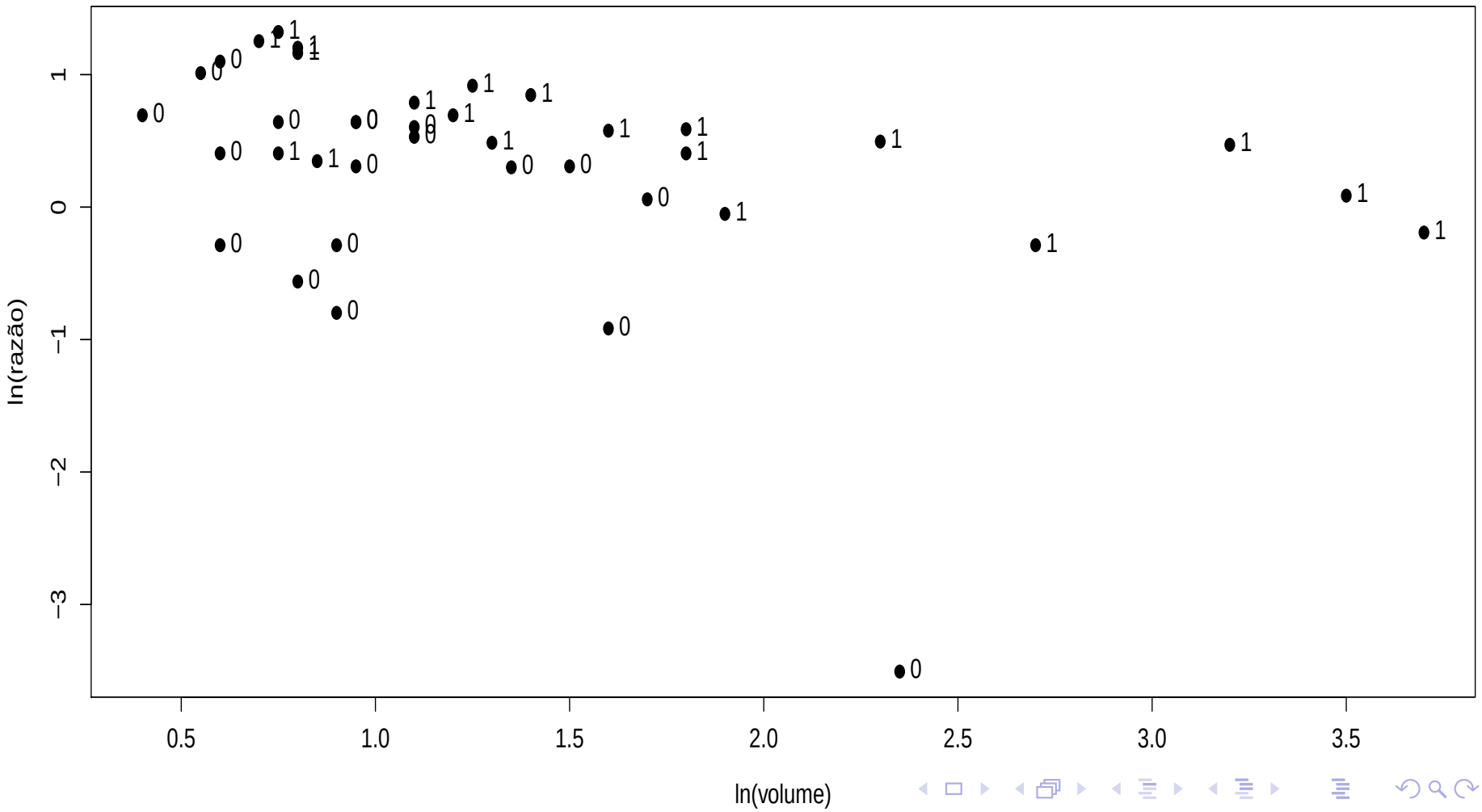


# Gráficos de dispersão: ln (razão) $\times$ ln(volume)



# Medidas resumo ln (razão) e ln(volume)

| | Involume | | Inrazão | |
|---|---|---|---|---|
| | Resposta | | | |
| Medida resumo | 0 | 1 | 0 | 1 |
| Média | -0,06 | 0,37 | 0,05 | 0,58 |
| Mediana | -0,05 | 0,30 | 0,31 | 0,54 |
| DP | 0,45 | 0,54 | 1,03 | 0,46 |
| Var. | 0,20 | 0,29 | 1,07 | 0,22 |
| \|CV(%)\| | 723,00 | 147,00 | 2223,00 | 81,00 |
| Min. | -0,92 | -0,36 | -3,51 | -0,29 |
| Max. | 0,85 | 1,31 | 1,10 | 1,30 |

# Modelo de regressão (geral) para dados binários

$$Y_i \overset{ind.}{\sim} \text{Bernoulli}(p_i)$$

$$F^{-1}(p_i) = \sum_{j=1}^{p} \beta_j x_{ji} \quad \rightarrow \quad p_i = F\left( \sum_{j=1}^{p} \beta_j x_{ji} \right), i = 1, 2, ..., n$$

- $Y_i$ : ocorrência (1) ou não (0) de algum evento.
- $x_{ji}$ : valor da variável explicativa $j$ associada ao indivíduo $i$; $\beta_j$ : parâmetro associado ao impacto de cada covariável na probabilidade de ocorrência do supracitado evento.
- $F(.)$ : função de distribuição acumulada de alguma variável aleatória (contínua) com suporte em $\mathcal{R}$. $F^{-1}(.)$ é conhecida como função de ligação.
- Modelo com intercepto: $x_{1i} = 1, \forall i$.

# Modelo de regressão para os dados de vasoconstrição

$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \text{Bernoulli}(p_i) \\
\text{logito}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \\
&\rightarrow \quad p_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}}, i = 1, 2, ..., n
\end{aligned}
$$

- $Y_i$ : ocorrência (1) ou não (0) de vaso constrição.
- $x_{1i}$ : logartimo natural do volume de ar inspirado da i-ésima observação; $x_{2i}$ : logartimo natural da razão de ar inspirado da i-ésima observação.
- $F(.)$ : corresponde à fda de uma distribuição logística padrão (portanto o nome regressão logística). Nesse caso, o logito(.) é a função de ligação.

# Modelo de regressão para os dados de vasoconstrição

- Interpretação dos parâmetros. Defina $l(p_i) = \text{logito}(p_i)$.
- Se $x_{1j} = x_{2j} = 0$, então $p_i = \dfrac{e^{\beta_0}}{1 + e^{\beta_0}}$.
- Defina $l_1(p_{i+1}) = \beta_0 + \beta_1(x_{1i} + 1) + \beta_2 x_{2i}$ e

  $l_1(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. Então

  $l_1(p_{i+1}) - l_1(p_i) = \beta_1 \rightarrow \dfrac{p_{i+1}/(1 - p_{i+1})}{p_i/(1 - p_i)} = e^{\beta_1}$ (**razão de chances em relação à primeira covariável**).
- Analogamente, defina $l_2(p_{i+1}) = \beta_0 + \beta_1 x_{1i} + \beta_2(x_{2i} + 1)$ e

  $l_2(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$. Então

  $l_2(p_{i+1}) - l_2(p_i) = \beta_2 \rightarrow \dfrac{p_{i+1}/(1 - p_{i+1})}{p_i/(1 - p_i)} = e^{\beta_2}$ (**razão de chances em relação à primeira covariável**).

# Inferência para o modelo

- Defina $\eta_i = \sum_{j=1}^{p} \beta_j x_{ji} = \mathbf{X}_i'\boldsymbol{\beta}$, em que $\mathbf{X}_i'$ é a i-ésima linha da matriz $\mathbf{X}$ e $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, em que $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)'$ e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$. Assim, temos que $Y_i \overset{ind.}{\sim}$ Bernoulli$(p_i)$, $p_i = F(\eta_i)$, $i = 1, 2...., n$.
- Verossimilhança

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

- Logverossimilhança.

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] \qquad (1)$$

# Inferência para o modelo

- Vetor escore

$$\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( \frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \frac{\partial p_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left( \frac{y_i - p_i}{p_i(1-p_i)} \right) \frac{\partial p_i}{\partial \boldsymbol{\beta}}$$

em que $\frac{\partial p_i}{\partial \boldsymbol{\beta}} = \frac{\partial F(\eta_i)}{\partial \boldsymbol{\beta}}$ é um vetor.

- Pela regra da cadeia e pelo fato de $F(.)$ ser uma fda, temos que

$$\frac{\partial F(\eta_i)}{\partial \boldsymbol{\beta}} = \frac{\partial F(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = f(\eta_i)\mathbf{X}_i,$$

em que $f(.)$ é a fdp associada à $F(.)$ e $\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{X}_i$.

# Inferência para o modelo

- Logo

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( (y_i - p_i) \frac{f(\eta_i)}{p_i(1-p_i)} \right) \mathbf{X}_i = \mathbf{X}'\mathbf{V}(\mathbf{y} - \boldsymbol{\mu})$$

em que $\mathbf{y} = (y_1, ..., y_n)'$, $\boldsymbol{\mu} = (p_1, ..., p_n)'$ e

$\mathbf{V} = \text{diag}\left(f(\eta_1)/(p_1(1-p_1)), ..., f(\eta_n)/(p_n(1-p_n))\right)$ (exercício).

- Podemos ainda escrever $\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} h_i(\boldsymbol{\beta}) g_i(\boldsymbol{\beta}) \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$, em que

$h_i(\boldsymbol{\beta}) = y_i - p_i$ e $g_i(\boldsymbol{\beta}) = \dfrac{f(\eta_i)}{p_i(1-p_i)}$.

# Inferência para o modelo

- A matriz Hessiana é dada por $H(\beta) = \dfrac{\mathbf{S}(\beta)}{\partial \beta'} = \dfrac{l(\beta)}{\partial \beta \partial \beta'}$.
- A matriz de informação de Fisher é dada por $\mathbf{I}(\beta) = -\mathcal{E}(\mathbf{H}(\beta))$ em que $\mathcal{E}(.)$ é calculada em termos da distribuição de $\mathbf{Y} = (Y_1, .., Y_n)'$. Note, ainda, que $\mathcal{E}(h_i(\beta)) = \mathcal{E}(Y_i - p_i) = 0$.
- Pela regrada da cadeia, temos que

$$\begin{aligned} \mathbf{H}(\beta) &= \sum_{i=1}^{n} \Big[ h_i(\beta) \frac{\partial \eta_i}{\partial \beta} \left( \frac{\partial g_i(\beta)}{\partial \beta} \right)' + h_i(\beta) g_i(\beta) \frac{\partial \eta_i}{\partial \beta \partial \beta'} \\ &+ g_i(\beta) \frac{\partial \eta_i}{\partial \beta} \left( \frac{\partial h_i(\beta)}{\partial \beta} \right)' \Big] \end{aligned}$$

# Inferência para o modelo

- Assim,

$$
\begin{aligned}
\mathbf{I}(\boldsymbol{\beta}) &= -\mathcal{E}\Big\{\sum_{i=1}^{n}\Big[h_i(\boldsymbol{\beta})\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}\left(\frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' + h_i(\boldsymbol{\beta})g_i(\boldsymbol{\beta})\frac{\partial \eta_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} \\
&+ g_i(\boldsymbol{\beta})\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}\left(\frac{\partial h_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'\Big]\Big\} \\
&= -\sum_{i=1}^{n} g_i(\boldsymbol{\beta})\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}\left(\frac{\partial h_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)' = \sum_{i=1}^{n} g_i(\boldsymbol{\beta})\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}\left(\frac{\partial p_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)',
\end{aligned}
$$

pois $\frac{\partial h_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial p_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, lembrando que $h_i(\beta) = (y_i - p_i)$.

# Inferência para o modelo

- Portanto,

$$\mathbf{I}(\beta) = \mathbf{X}'\mathbf{W}\mathbf{X},$$

em que $\mathbf{W} = \text{diag}\left(f(\eta_1)^2/(p_1(1-p_1)), ..., f(\eta_n)^2/(p_n(1-p_n))\right)$

- Se $p_i = \dfrac{e^{\eta_i}}{1+e^{\eta_i}}$ (regressão logística), então

$$\mathbf{S}(\beta) = \mathbf{X}'(\mathbf{y} - \mu)$$

e

$$\mathbf{I}(\beta) = \mathbf{X}'\mathbf{D}\mathbf{X}$$

(exercício), em que $\mathbf{D} = \text{diag}(p_1(1-p_1), ..., p_n(1-p_n))$ .

# Inferência para o modelo

- Independente da escolha de $F(.)$, o sistema de equações $\mathbf{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ não tem solução explícita e algum método de otimização numérica, como o algoritmo escore de Fisher, deve ser utilizado para obter-se as estimativas de MV.

# Algoritmo escore de Fisher

- Seja $\boldsymbol{\beta}^{(0)}$ uma estimativa inicial de $\boldsymbol{\beta}$ (chute inicial), então faça

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{S}(\boldsymbol{\beta}^{(t)}), t = 1, 2, .... \quad (2)$$

até que algum critério de convergência seja satisfeito, como

$$|l(\boldsymbol{\beta}^{(t+1)}) - l(\boldsymbol{\beta}^{(t)})| < \epsilon, \epsilon > 0,$$

em que $l(.)$ é a logverossimilhança (equação (1)).

# Algoritmo escore de Fisher

- A equação (2) pode ser reescrita como

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(t)}\mathbf{z}^{(t)},$$

em que $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1/2}\mathbf{D}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$.

# Mais sobre inferência

- Para $n$ suficientemente grande, $\widehat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$. Na prática, trabalhamos com $\mathbf{I}^{-1}(\widetilde{\boldsymbol{\beta}})$, em que $\widetilde{\boldsymbol{\beta}}$ é a emv de $\boldsymbol{\beta}$, obtida através do algoritmo Escore de Fisher.
- Defina, $\widehat{\sigma}_i^2$ : o i-ésimo elemento da diagonal principal de $\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})$ e $\widetilde{\sigma}_i^2$ : o i-ésimo elemento da diagonal principal de $\mathbf{I}^{-1}(\widetilde{\boldsymbol{\beta}})$ .
- Assim, hipóteses do tipo $H_0 : \beta_i = \beta_0$ vs $\beta_i \neq \beta_0$, podem ser testadas através da estatística $Z_t = \frac{\widehat{\beta}_i - \beta_0}{\sqrt{\widehat{\sigma}_i^2}}$, rejeitando-se $H_0$ quando p-valor $\leq \alpha$, p-valor $\approx 2P(Z \geq |z_t||H_0)$, $Z \sim N(0,1)$ e $z_t = \frac{\widetilde{\beta}_i - \beta_0}{\sqrt{\widetilde{\sigma}_i^2}}$, em que $\widehat{\beta}_i$ é o estimador de MV de $\beta_i$ e $\widetilde{\beta}_i$ a respectiva estimativa.

# Mais sobre inferência

- Hipóteses do tipo $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{M}$ vs $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{M}$ podem ser testadas através da estatística

$$Q_t = \left(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{M}\right)' \left(\mathbf{C}\mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}'\right)^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{M}\right).$$

  Sob $H_0$ e para $n$ suficientemente grande, $Q_t \approx \chi^2_c$, em que $c$ é o número de linhas de $\mathbf{C}$.

- Assim, rejeita-se $H_0$ se $p - valor \leq \alpha$, em que

  $p - valor \approx P(X \geq q_t | H_0)$, em que $X \sim \chi^2_c$

  $q_t = \left(\mathbf{C}\widetilde{\boldsymbol{\beta}} - \mathbf{M}\right)' \left(\mathbf{C}\mathbf{I}^{-1}(\widetilde{\boldsymbol{\beta}})\mathbf{C}'\right)^{-1} \left(\mathbf{C}\widetilde{\boldsymbol{\beta}} - \mathbf{M}\right).$

# Voltando ao conjunto de dados (os testes se referem à nulidade de cada parâmetro)

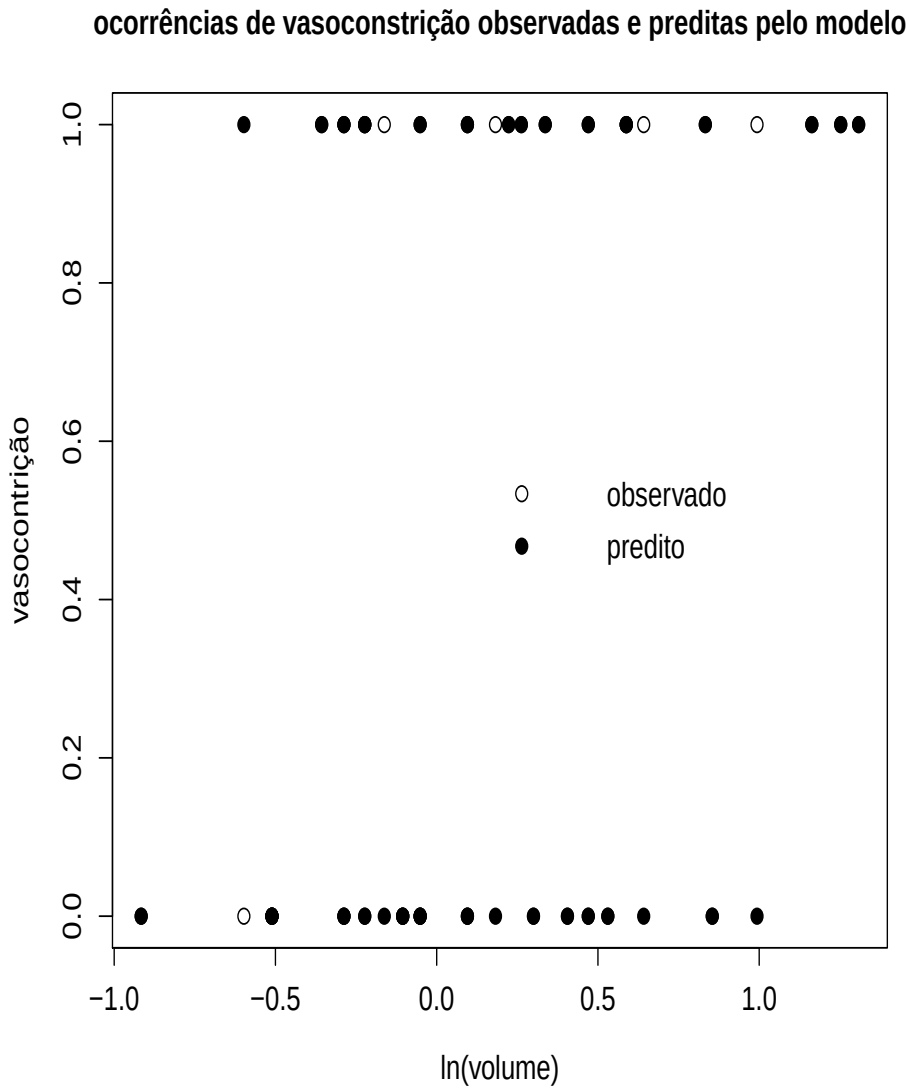
| Parâmetro | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\beta_0$ | -2,87 | 1,32 | -2,18 | 0,0295 |
| $\beta_1$ | 5,17 | 1,86 | 2,78 | 0,0055 |
| $\beta_2$ | 4,56 | 1,83 | 2,48 | 0,0131 |

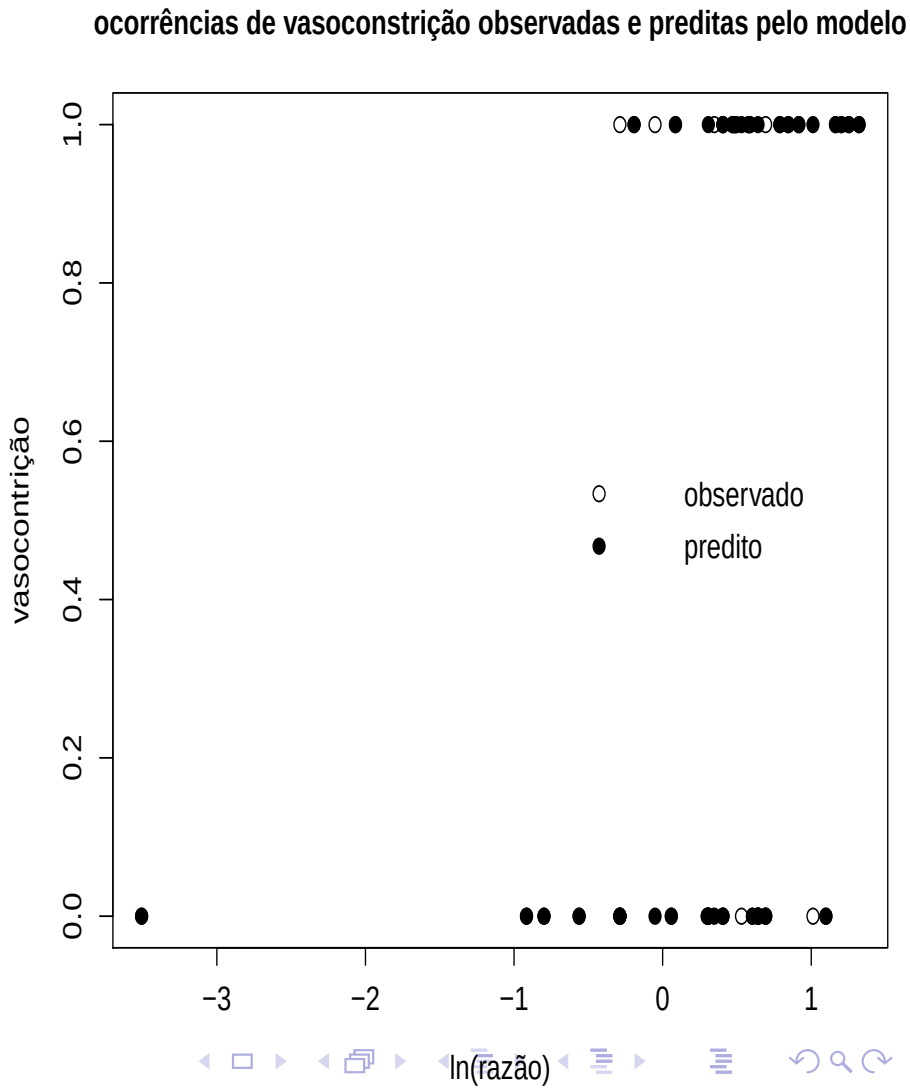
Todos os parâmetros são significativos.

# Probabilidades e valores preditos

- Probabilidades de ocorrência de vasoconstrição preditas:
$$\widetilde{\pi}_i = \frac{e^{\widetilde{\beta}_0+\widetilde{\beta}_1 x_{1i}+\widetilde{\beta}_2 x_{2i}}}{1+e^{\widetilde{\beta}_0+\widetilde{\beta}_1 x_{1i}+\widetilde{\beta}_2 x_{2i}}}$$
- Ocorrências de vasoconstrição preditas: simula-se $u, U \sim U(0,1)$, se $\widetilde{\pi}_i \geq u$, então $\widetilde{Y}_i = 1$, caso contrário, $\widetilde{Y}_i = 0$.

# Valores observados e preditos pelo modelo



# Perguntas

- Como gerar intervalos de confiança para $\frac{e^{\beta_0}}{1+e^{\beta_0}}$, $e^{\beta_1}$ e $e^{\beta_2}$?
  - Método delta.
  - Fazer um IC para o parâmetro original e depois calcular o IC para a transformação.
  - Reamostragem.
- Como verificar as suposições do modelo?
  - Estatísticas de qualidade de ajuste.
  - Resíduo componente do desvio.
- Vamos nos concentrar na regressão logística ($F^{-1}(p_i) = \text{logito}(p_i)$).

# Intervalos de confiança para funções de interesse

- Sejam $g_1(\boldsymbol{\beta}) \equiv \tau_1 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$, $g_2(\boldsymbol{\beta}) \equiv \tau_2 = e^{\beta_1}$ e $g_3(\boldsymbol{\beta}) \equiv \tau_3 = e^{\beta_2}$.
- Seja $\widehat{\boldsymbol{\beta}}$ o estimador de MV de $\boldsymbol{\beta}$. Já vimos que, para $n$ suficientemente grande, $\widehat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, em que $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbf{I}^{-1}(\boldsymbol{\beta})$.
- O método delta nos diz que, para $n$ suficientemente grande, $\widehat{\tau}_i \approx N(\tau_i, \boldsymbol{\Psi}_i \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\Psi}_i')$, em que

$$\boldsymbol{\Psi}_i = \left[ \begin{array}{ccc} \frac{\partial}{\partial \beta_0} g_i(\boldsymbol{\beta}) & \frac{\partial}{\partial \beta_1} g_i(\boldsymbol{\beta}) & \frac{\partial}{\partial \beta_2} g_i(\boldsymbol{\beta}) \end{array} \right]$$

# Intervalos de confiança para funções de interesse

- Nesse caso,

  $\boldsymbol{\Psi}_1 = \begin{bmatrix} \frac{e^{\beta_0}}{(1+e^{\beta_0})^2} & 0 & 0 \end{bmatrix}$, $\boldsymbol{\Psi}_2 = \begin{bmatrix} 0 & e^{\beta_1} & 0 \end{bmatrix}$,

  $\boldsymbol{\Psi}_3 = \begin{bmatrix} 0 & 0 & e^{\beta_2} \end{bmatrix}$

- Assim $IC(\tau_i, \gamma) = \left[\widehat{\tau}_i - z_{(1+\gamma)/2}\sqrt{\widehat{\psi}_i}; \widehat{\tau}_i + z_{(1+\gamma)/2}\sqrt{\widehat{\psi}_i}\right]$, em que $P(Z \geq z_{(1+\gamma)/2}) = \frac{1+\gamma}{2}$ e $\widehat{\psi}_i = \widehat{\boldsymbol{\Psi}}_i \widehat{\boldsymbol{\Sigma}}_\beta \widehat{\boldsymbol{\Psi}}_i'$, $Z \sim N(0,1)$ (lembrando que este é um IC assintótico).

# Intervalos de confiança para funções de interesse

| Parâmetro | IC (transformação) | IC (método delta) |
|---|---|---|
| $\tau_1$ | [0,00 ; 0,43] | [-0,08 ; 0,18 ] |
| $\tau_2$ | [4,59 ; 6862,99] | [-471,35 ; 826,48] |
| $\tau_3$ | [2,61 ; 3511,02] | [-249,12 ; 440,61] |

Neste caso, os IC's obtidos através do método delta, devem ser truncados à esquerda do zero. Exercício: obter os intervalos de confiança através de reamostragem.

# Verificação da qualidade de ajuste do modelo

- No modelo em questão, temos, essencialmente, as seguintes suposições a serem avaliadas.
  - Apesar do modelo ser heterocedástico ($\mathcal{V}(Y_i) = p_i(1 - p_i)$), a variância por ele imposta pode ser menor do que a observada (superdispersão) ou maior do que a observada (subdispersão).
  - As observações são independentes.
  - A função de ligação (nesse caso $F^{-1}$, $F$ uma fda) é apropriada.

# Verificação da qualidade de ajuste do modelo

- Função desvio: Seja $l(\boldsymbol{\mu}, \mathbf{y})$ a logverossimilhança do modelo ($\mathcal{E}(\mathbf{Y}) = \boldsymbol{\mu} = F(\boldsymbol{\eta})$). Lembremos que, para o modelo Bernoulli $\mu_i = p_i$.
- Para o modelo saturado $n = p$, ou seja, em que representamos a média de cada observação por ela mesma, o estimador de MV de $\mu_i$ é $\widehat{\mu}_i = Y_i$. Nesse caso o estimador de $l(\boldsymbol{\mu}, \mathbf{y})$ é dado por $l(\mathbf{Y}; \mathbf{y})$.
- Seja $l(\widehat{\boldsymbol{\mu}}, \mathbf{y})$ o estimador de MV da logverossimilhança sob o modelo em estudo, em que $\widehat{\boldsymbol{\mu}} = F(\widehat{\boldsymbol{\eta}})$ e $\widehat{\boldsymbol{\eta}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

# Verificação da qualidade de ajuste do modelo

- A função desvio (ou simplesmente desvio) é definida por

  $D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2\left\{l(\mathbf{Y}, \mathbf{y}) - l(\widehat{\boldsymbol{\mu}}, \mathbf{y})\right\}$

- No caso do modelo de regressão logística Bernoulli e com observações independentes, temos que

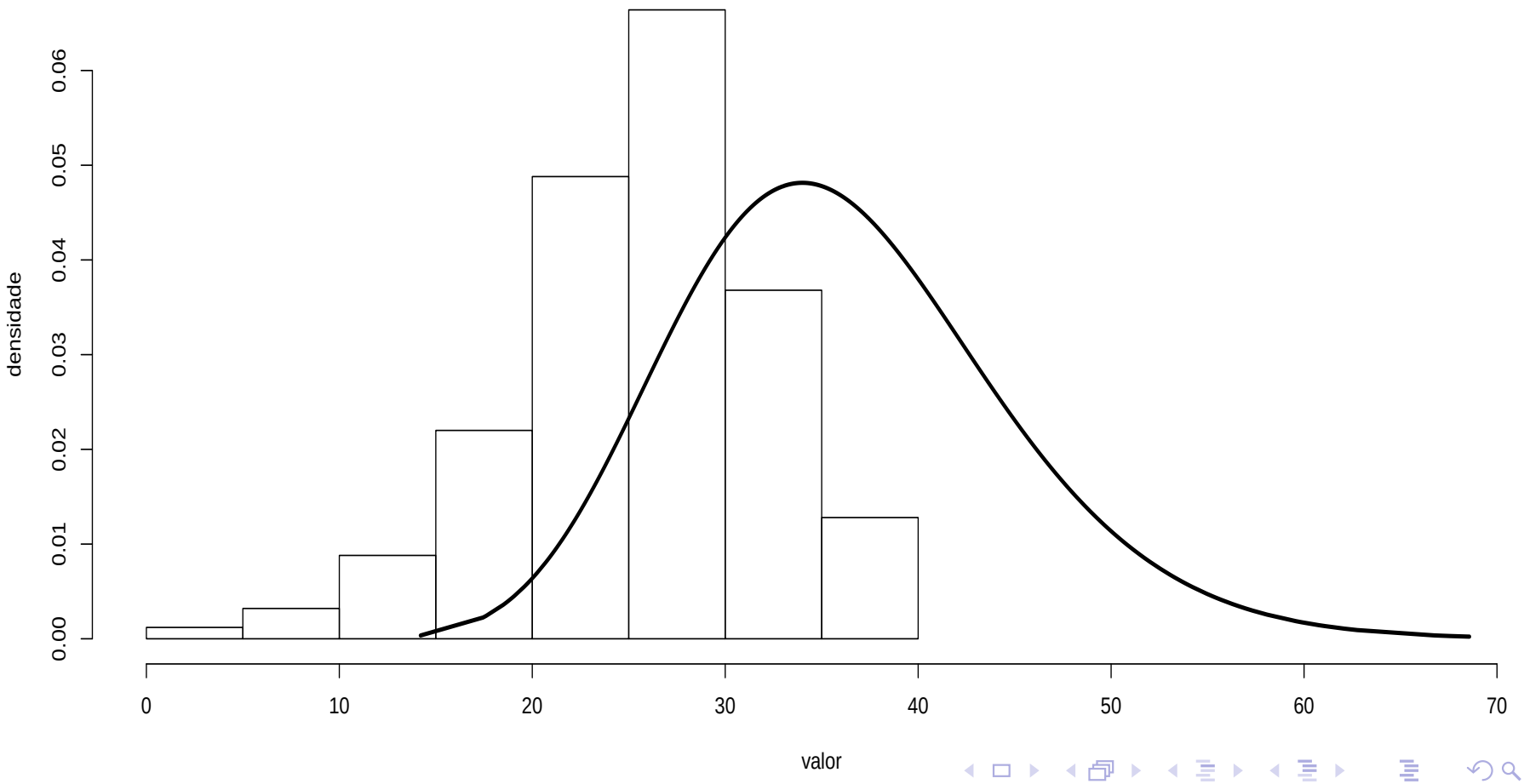
  $D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = -2\sum_{i=1}^{n}\left\{\ln(1-\widehat{\mu}_i)I_{\{0\}}(y_i) + \ln(\widehat{\mu}_i)I_{\{1\}}(y_i)\right\}$

- Contudo, em geral $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ não segue (mesmo assintoticamente) uma distribuição $\chi^2_{(n-p)}$, sob a hipótese de que o modelo em questão é adequado.

# Verificação da qualidade de ajuste do modelo

- É aconselhável obter um p-valor para a estatística $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ por reamostragem.
- Algoritmo
  1. Ajuste o modelo (estime seus parâmetros) por MV e calcule o desvio (desvo).
  2. Para j = 1,...,R, faça
     - Simule $n$ variáveis Bernoulli de acordo com o modelo.
     - Ajuste o modelo considerando as variáveis simuladas anteriormente, e calcule o respectivo desvio (desvs).
  3. Assim, $p - valor = \frac{1}{R} \sum_{j=1}^{R} \mathbb{1}(desvs_j \geq desvo)$.

# Histograma da distribuição empírica do desvio e a curva teórica $\chi^2_{36}$



# Comentários

- Nesse caso, a aproximação assintótica se mostrou inapropriada.
- Além disso, $p - valor_{\text{reamostragem}} = 0,2880$ e $p - valor_{\text{assintótico}} = 0,7807$.
- Portanto, embora diferentes, os p-valores levam à mesma conclusão (o modelo está bem ajustado).

# Verificação da qualidade de ajuste do modelo

- Uma outra forma de verificar a qualidade do ajuste do modelo, é através da análise de resíduos.
- Utilizar o resíduo padronizado (semelhante aquele utilizado em modelos de regressão normais lineares), ou seja, $\frac{Y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}$ não é apropriado.
- Particularmente, esse resíduo não terá distribuição normal (mesmo sob a validade das hipóteses do modelo).

# Verificação da qualidade de ajuste do modelo

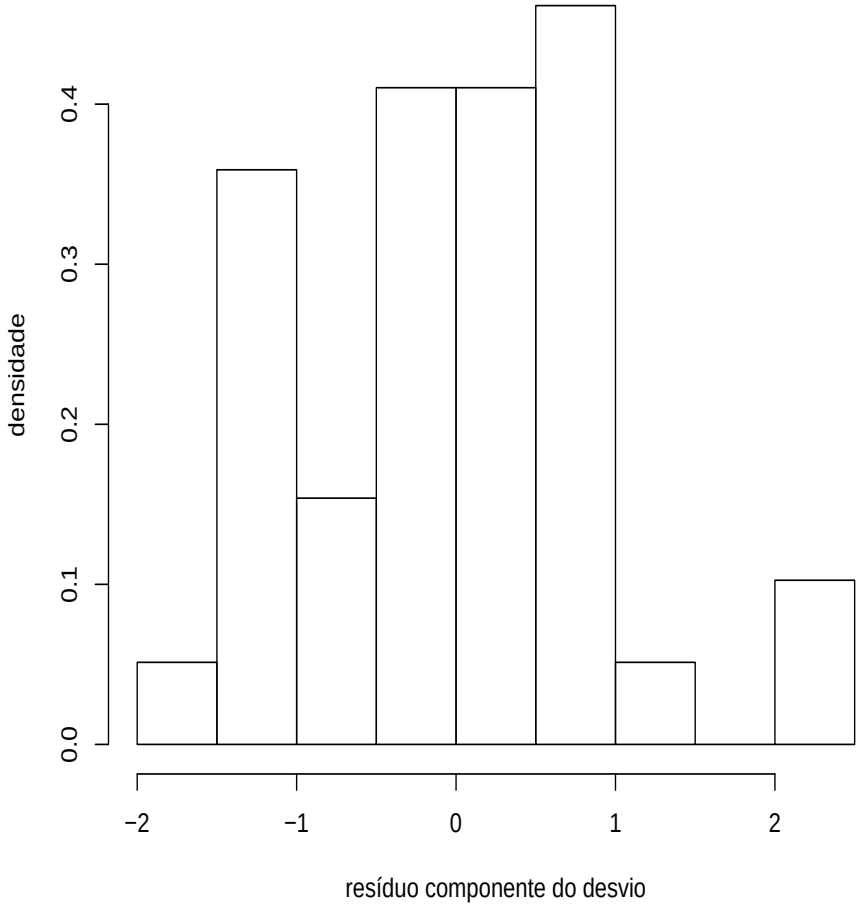
- Alternativa: resíduo componente do desvio (RCD). Nesse caso, é dado por

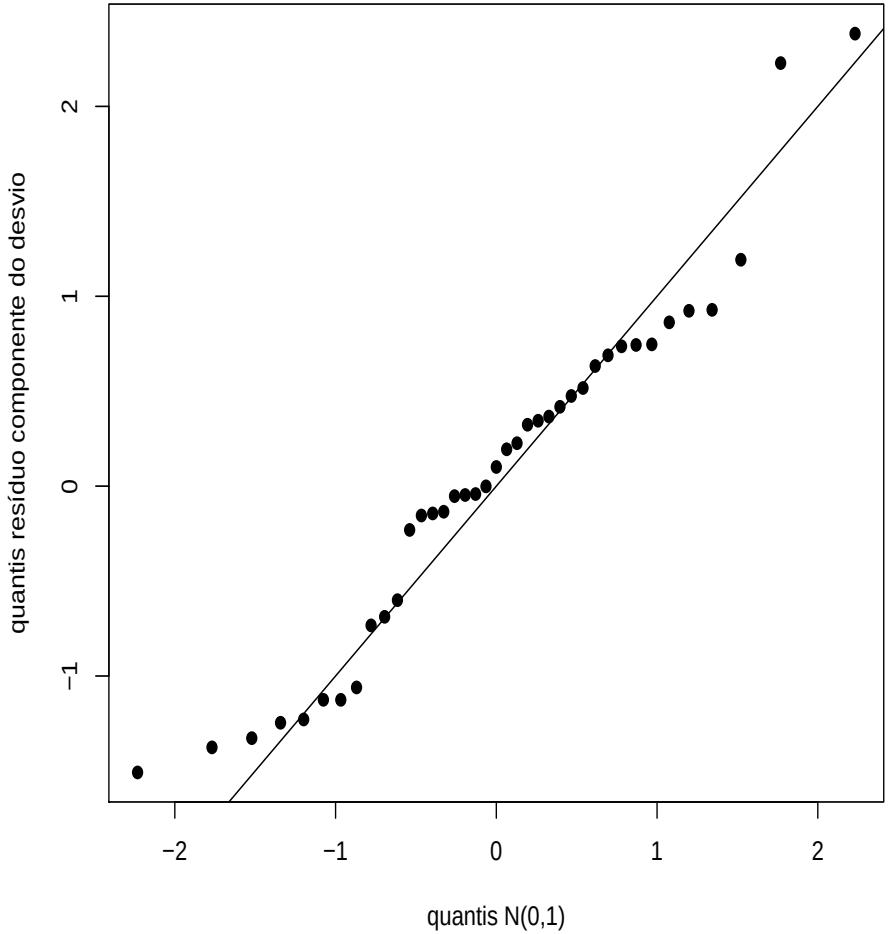
$$T_{D_i} = -\frac{(2|\ln(1-\widehat{p}_i)|)^{1/2}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{0\}}(y_i) + \frac{(2|\ln \widehat{p}_i|)^{1/2}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{1\}}(y_i)$$

em que $\widehat{h}_{ii} = \widehat{p}_i(1-\widehat{p}_i)\mathbf{x}_i' \left(\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X}\right)^{-1} \mathbf{x}_i$ e $\widehat{\mathbf{V}} = \text{diag}(\widehat{p}_1(1-\widehat{p}_1), ..., \widehat{p}_n(1-\widehat{p}_n))$.

- Para $n$ suficientemente grande e sob a validade das suposições do modelo, $T_{D_i} \approx N(0,1)$.
- Pergunta: construir um histograma e/ou qq-plots para os RCD's é apropriado (suficiente) para avaliar o comportamento dos resíduos?

# Histograma e qq-plot do rcd



# Procedimento para se gerar o gráfico de envelopes com o RCD

1) Ajuste o modelo de regressão (estima-se os parâmetros do modelo) obtendo-se as estimativas de MV ($\widetilde{\boldsymbol{\beta}}$) e calcule o RCD para cada observação, $(t_{D_i})$, $i = 1, 2, ..., n$.
2) De posse das estimativas de MV, repita os passos (a) e (b) $m$ vezes.
   a) Simule $n$ variáveis aleatórias ind. Bernoulli($\widetilde{p}_i$), com $\widetilde{p}_i = F(\widetilde{\boldsymbol{\eta}}_i)$, $\widetilde{\boldsymbol{\eta}}_i = \mathbf{X}_i'\widetilde{\boldsymbol{\beta}}$.
   b) Ajuste o modelo de regressão considerando as variáveis simuladas no item a) e obtenha o RCD para cada observação (i) em cada réplica (j).

# Procedimento para se gerar o gráfico de envelopes com o RCD

3) Ao final teremos uma matriz com os RCD's, ou seja $t^*_{D_{ij}}$, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\mathbf{T}_1 = \begin{bmatrix} t^*_{D_{11}} & t^*_{D_{12}} & \cdots & t^*_{D_{1m}} \\ t^*_{D_{21}} & t^*_{D_{22}} & \cdots & t^*_{D_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{n1}} & t^*_{D_{n2}} & \cdots & t^*_{D_{nm}} \end{bmatrix}$$

# Procedimento para se gerar o gráfico de envelopes com o RCD

4) Dentro de cada amostra, ordena-se, de modo crescente, os RCD's, obtendo-se $t^*_{D_{(i)j}}$ (estatísticas de ordem):

$$\mathbf{T}_2 = \begin{bmatrix} t^*_{D_{(1)1}} & t^*_{D_{(1)2}} & \cdots & t^*_{D_{(1)m}} \\ t^*_{D_{(2)1}} & t^*_{D_{(2)2}} & \cdots & t^*_{D_{(2)m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{(n)1}} & t^*_{D_{(n)2}} & \cdots & t^*_{D_{(n)m}} \end{bmatrix}$$

5) Obtem-se os limites $t^*_{(i)I} = \min\limits_{1 \leq j \leq m} t^*_{D_{(i)j}}$ e $t^*_{(i)S} = \max\limits_{1 \leq j \leq m} t^*_{D_{(i)j}}$, $j = 1, 2, ..., m$.

# Procedimento para se gerar o gráfico de envelopes com o RCD

5) Na prática considera-se $t^*_{(i)I} = \dfrac{t^*_{D_{(i)(2)}} + t^*_{D_{(i)(3)}}}{2}$ e $t^*_{(i)S} = \dfrac{t^*_{D_{(i)(m-2)}} + t^*_{D_{i(m-1)}}}{2}$ (refinamento das estimativas do mínimo e máximo), em que $t^*_{D_{(i)(r)}}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ...., n$.

- Além disso, consideramos como a linha de referência $t^*_{(i)} = \frac{1}{m}\sum_{j=1}^{m} t^*_{D_{(i)j}}, i = 1, 2, ..., n.$

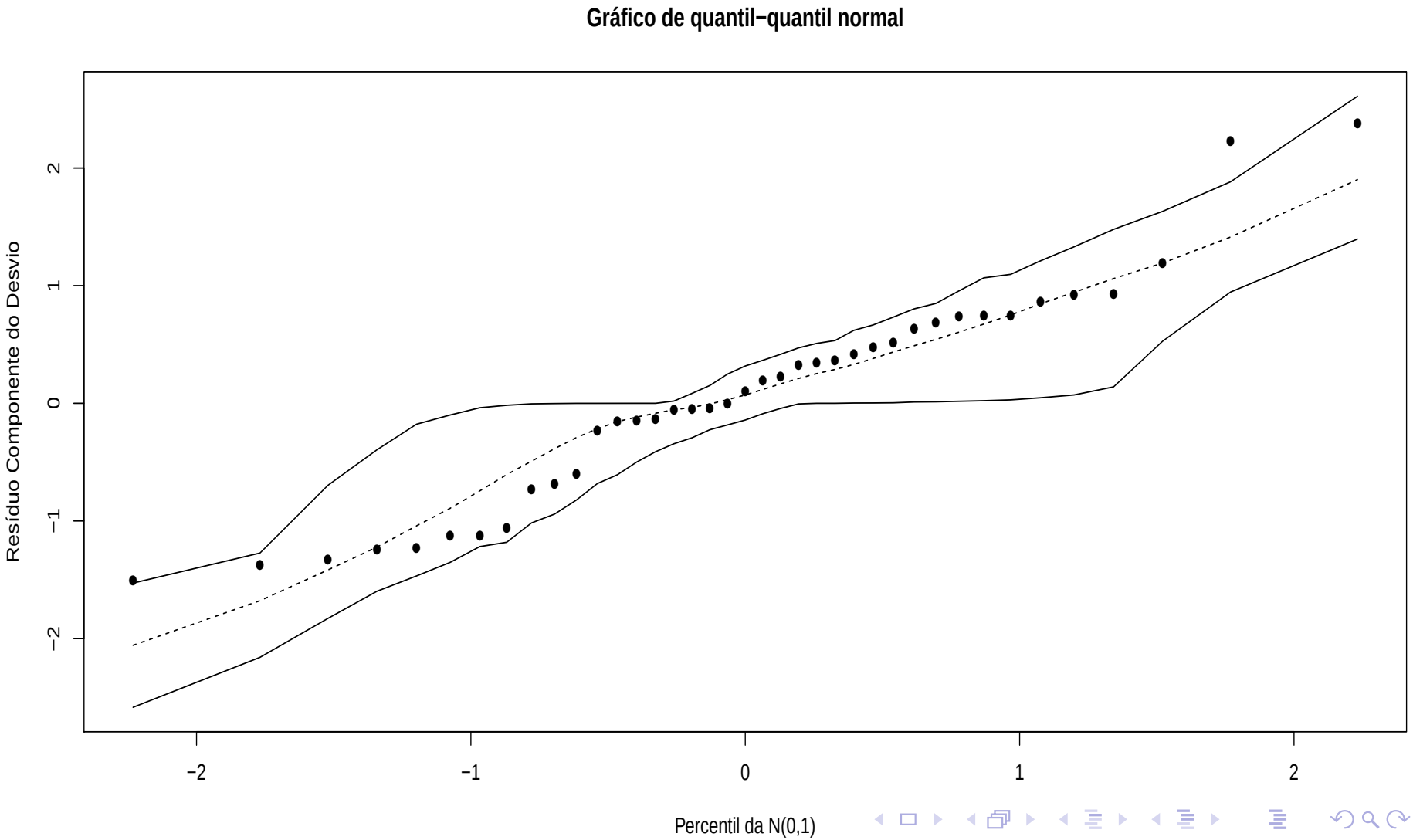
# Outros gráficos de interesse

- $t_{D_i} \times$ ordem da observação: pontos aberrantes, hetorogeneidade (heterocedasticidade) não capturada pelo modelo.
- $t_{D_i} \times F(\widetilde{\eta}_i)$(valor predito): pontos aberrantes.
- $\widetilde{z}_i \times \widetilde{\eta}_i$: adequabilidade da função de ligação e do preditor linear $(\eta_i)$, em que $\widetilde{z}_i = \widetilde{\eta}_i + \widetilde{W}_i^{-1/2}\widetilde{D}_i^{-1/2}(y_i - \widetilde{\mu}_i)$, em que $\eta_i = \mathbf{X}_i'\widetilde{\boldsymbol{\beta}}$ e

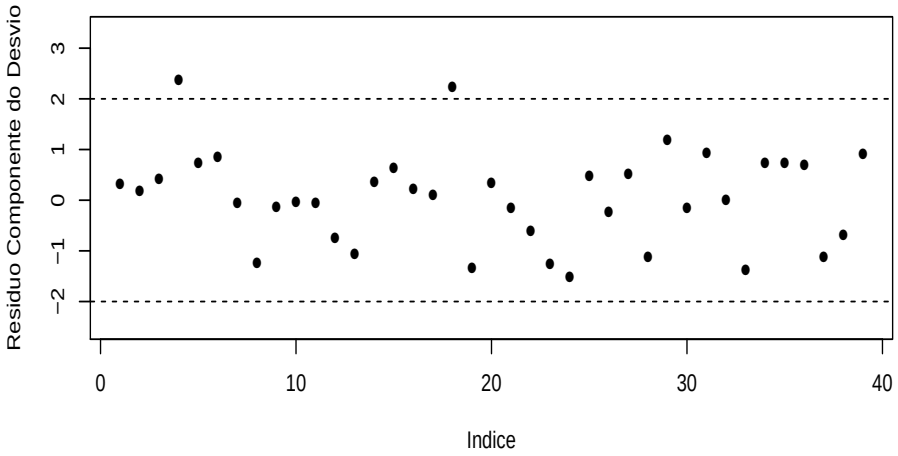
$$\mathbf{W} = \text{diag}\left(f(\widetilde{\eta}_1)^2/(\widetilde{p}_1(1-\widetilde{p}_1)), ..., f(\widetilde{\eta}_n)^2/(\widetilde{p}_n(1-\widetilde{p}_n))\right)$$

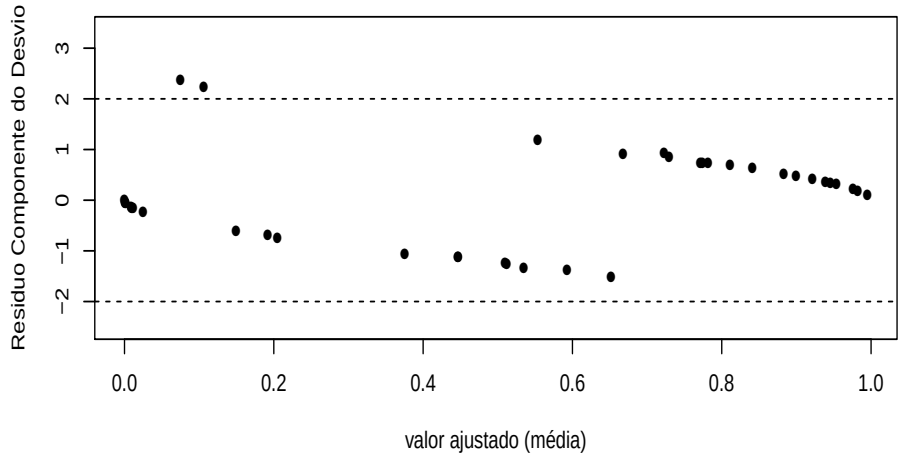
$$\mathbf{D} = \text{diag}(\widetilde{p}_1(1-\widetilde{p}_1), ..., \widetilde{p}_n(1-\widetilde{p}_n))$$

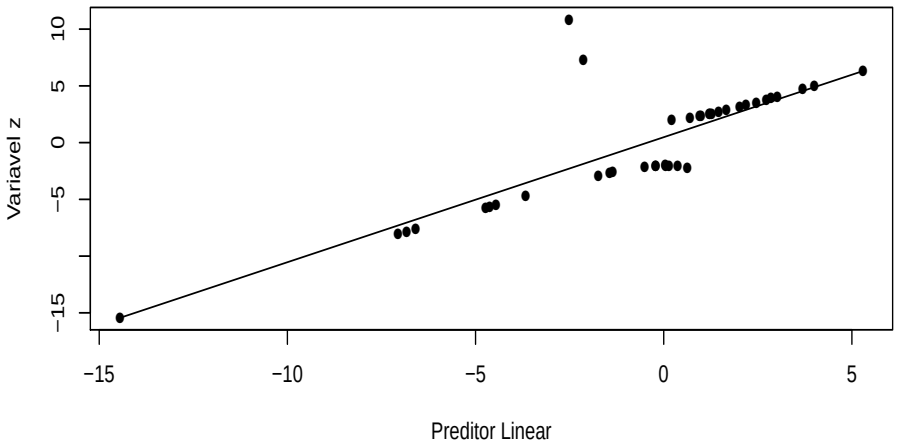
# Gráficos de envelopes para os RCD's

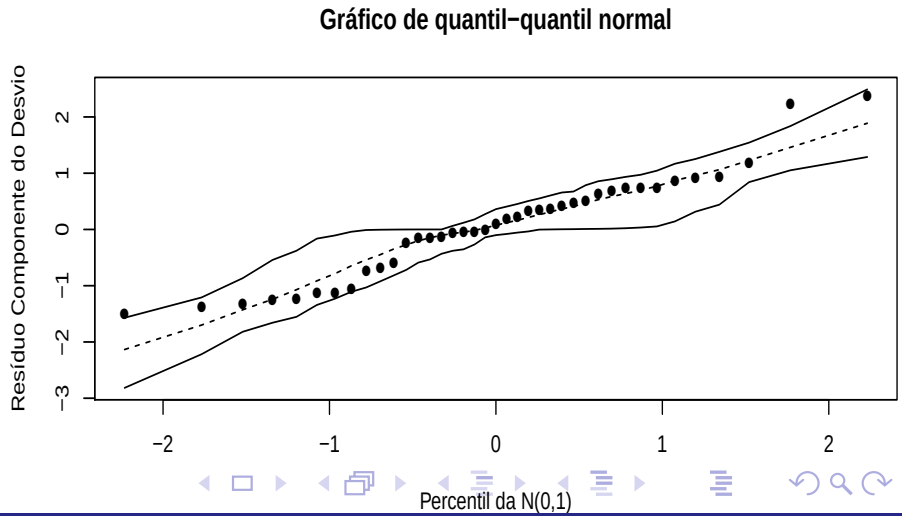


# Gráficos de envelopes para os RCD's



# Comentários

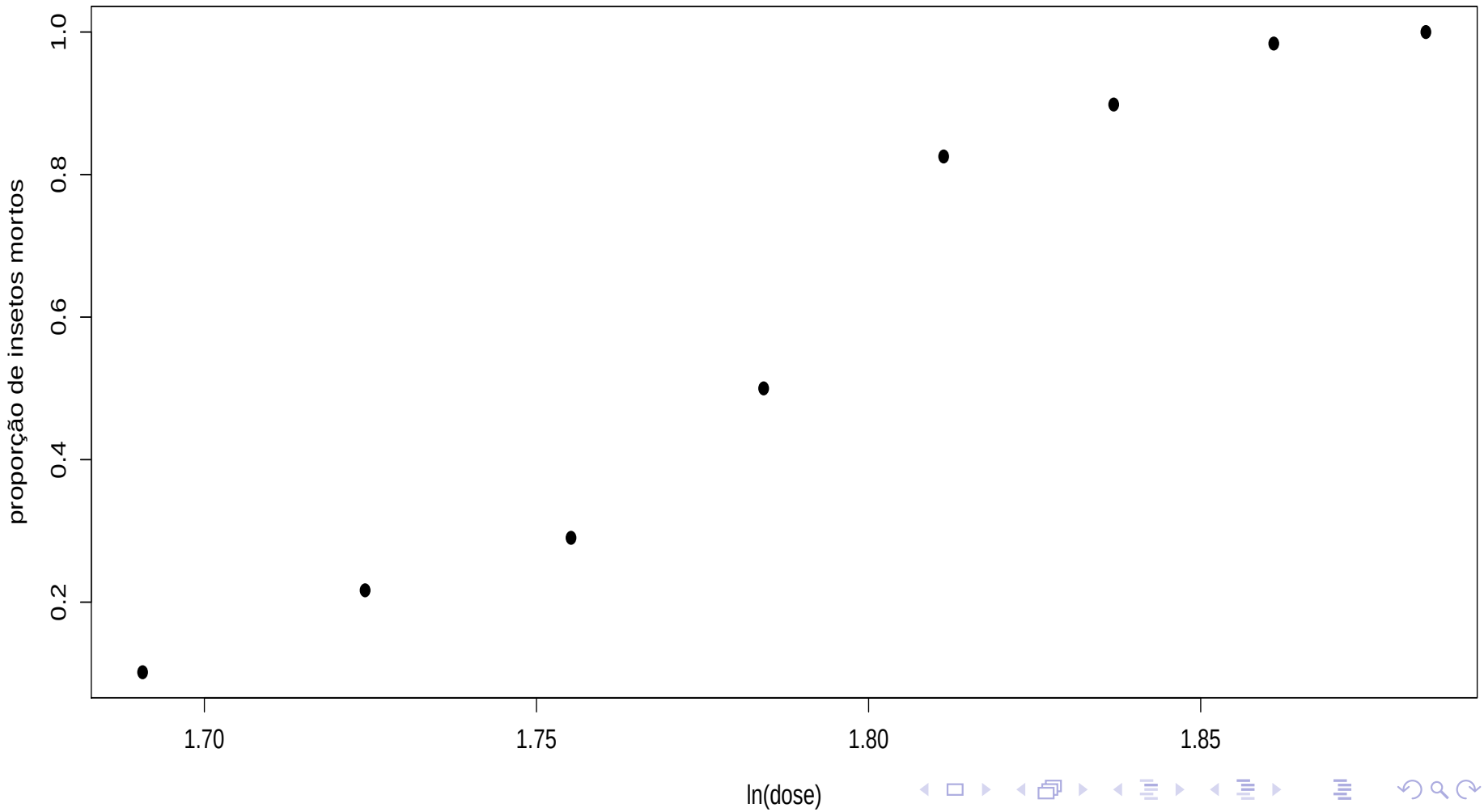
- A análise de diagnóstico indicou que o modelo se ajustou bem aos dados.
- Para finalizar: apresentar as estimativas pontuais e intervalares de vasoconstrição para diferentes valores do log(volume) e log(razão) de interesse do pesquisador (**exercício**).

# Exemplo 12: mortalidade de besouros

- Dados relativos ao percentual de besouros mortos quando expostos à diferentes doses de disulfeto de carbono gasoso ($CS_2$).

| Dose: $log_{10} CS_2$ | nº Besouros expostos | nº Besouros mortos |
|---|---|---|
| 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 |

# Gráficos de dispersão



# Exemplo 12: mortalidade de besouros

- Modelo1

$$\begin{aligned} Y_i &\overset{ind.}{\sim} \text{binomial}(m_i, p_i) \\ \ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i, i = 1, 2, ..., 8 \end{aligned}$$

- $m_i$ : número de besouros expostos à dose i de $CS_2$.
- $Y_i$ : número de besouros expostos à dose i de $CS_2$ que morreram.
- $x_i$ : dose (log da concentração de $CS_2$) à que os besouros do grupo i foram expostos.

# Cont. do modelo 1

- Assim, $p_i = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$.
- $\beta_0$ é o logito $\left[\ln\left(\frac{p_i}{1-p_i}\right)\right]$ da proporção de besouros mortos submetidos à uma concentração de 1 unidade de $CS_2$. Ou seja, se $x_i = log_{10}(concent) = log_{10}(1) = 0$ então $p_i = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$.

# Cont. do modelo 1

- Sejam: $p_i = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$ e $p_{i+1} = \dfrac{e^{\beta_0+\beta_1(x_i+1)}}{1+e^{\beta_0+\beta_1(x_i+1)}}$.
- Assim: $\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ e $\ln\left(\frac{p_{i+1}}{1-p_{i+1}}\right) = \beta_0 + \beta_1(x_i+1)$.
- Logo: $\ln\left(\frac{p_{i+1}}{1-p_{i+1}}\right) - \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{p_{i+1}/(1-p_{i+1})}{p_i/(1-p_i)}\right) = \beta_1$.
- Portanto, $\dfrac{p_{i+1}/(1-p_{i+1})}{p_i/(1-p_i)} = e^{\beta_1}$ (razão de chances).

# Exemplo 12: mortalidade de besouros

- Modelo 2

$$Y_i|(\beta_0,\beta_1) \overset{ind.}{\sim} \text{binomial}(m_i, p_i)$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1(x_i - \overline{x}), \overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i, i = 1, 2, ..., 8$$

- Neste caso, $\beta_0$ é o logito $\left[\ln\left(\frac{p_i}{1-p_i}\right)\right]$ da proporção de besouros mortos submetidos à uma concentração igual à $\overline{x}$ unidades de $CS_2$. Ou seja, se $x_i = \frac{1}{8}\sum_{i=1}^{8} log_{10}(concent_i)$, então $p_i = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$.
- As outras quantidades, incluindo o parâmetro $\beta_1$, possuem as mesmas interpretações que no modelo 1, (substituindo-se $x_i$ por $x_i - \overline{x}$).

# Inferência para o modelo

- Defina $\eta_i = \sum_{j=1}^{p} \beta_j x_{ji} = \mathbf{X}_i'\boldsymbol{\beta}$, em que $\mathbf{X}_i'$ é a i-ésima linha da matriz $\mathbf{X}$ e $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, em que $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)'$ e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$. Assim, temos que $Y_i \overset{ind.}{\sim}$ binomial$(m_i, p_i), p_i = F(\eta_i), i = 1, ..., k$.
- Verossimilhança

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \begin{pmatrix} m_i \\ y_i \end{pmatrix} p_i^{y_i}(1-p_i)^{m_i - y_i} \propto \prod_{i=1}^{k} p_i^{y_i}(1-p_i)^{m_i - y_i}$$

- Logverossimilhança.

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln p_i + (m_i - y_i)\ln(1-p_i)] + \text{const.} \qquad (3)$$

# Inferência para o modelo

- Os desenvolvimentos relativos ao processo de inferência são muito semelhantes àqueles apresentados, considerando-se a distribuição de Bernoulli.
- Por simplicidade, vamos apresentar os resultados somente para o modelo de regressão logístico: $p_i = \dfrac{e^{\mathbf{X}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i'\boldsymbol{\beta}}}$.
- Vetor escore

$$\mathbf{S}(\boldsymbol{\beta}) = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}),$$

em que $\mathbf{X}$ é matriz de planejamento, $\mathbf{y} = (y_1, ..., y_k)'$ e $\boldsymbol{\mu} = (p_1, ..., p_k)'$.

# Inferência para o modelo

- Por outro lado, a informação de Fisher é dada por

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X},$$

em que $\mathbf{V} = \text{diag}\left(m_1 p_1(1-p_1), ..., m_k p_k(1-p_k)\right))$

- Novamente, o sistema de equações $\mathbf{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ não tem solução explícita e algum método de otimização numérica, como o algoritmo escore de Fisher, deve ser utilizado para obter-se as estimativas de MV.

# Algoritmo escore de Fisher

- Seja $\boldsymbol{\beta}^{(0)}$ uma estimativa inicial de $\boldsymbol{\beta}$ (chute inicial), então faça

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{S}(\boldsymbol{\beta}^{(t)}), t = 1, 2, .... \quad (4)$$

até que algum critério de convergência seja satisfeito, como

$$|l(\boldsymbol{\beta}^{(t+1)}) - l(\boldsymbol{\beta}^{(t)})| < \epsilon, \epsilon > 0,$$

em que $l(.)$ é a logverossimilhança (equação (3)).

- A equação (4) pode ser reescrita como

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}'\mathbf{V}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{(t)}\mathbf{z}^{(t)},$$

em que $\mathbf{z} = \boldsymbol{\eta} + \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

# Comentários

- Os resultados anteriores continuam válidos, com pequenas modificações.
- $\mathcal{V}(Y_i) = m_i p_i (1 - p_i)$.
- $\mathbf{I}(\beta) = \mathbf{X}'\mathbf{V}\mathbf{X}, \mathbf{V} = \text{diag}\left(m_1 p_1 (1 - p_1), ..., m_k p_k (1 - p_k)\right)$.

# Comentários

- Desvio:

$$
\begin{aligned}
& D(\mathbf{y}, \widehat{\boldsymbol{\mu}}) = \\
& 2 \sum_{i=1}^{k} \Big\{ y_i \ln[y_i/(m_i \widehat{p}_i)] + (m_i - y_i) \ln\left[(1 - y_i/m_i)/(1 - \widehat{p}_i)\right] \\
\times \quad & \mathbb{1}_{\{1,\ldots,(m_i-1)\}}(y_i) \\
- \quad & 2[m_i \ln(1 - \widehat{p}_i)] I_{\{0\}}(y_i) - 2[m_i \ln \widehat{p}_i] I_{\{m_i\}}(y_i) \Big\}. \qquad (5)
\end{aligned}
$$

Nesse caso, para $k$ fixado e $m_i \to \infty, i = 1, 2, ..., k$, sob a hipótese de que o modelo é adequado, $D(\mathbf{y}, \widehat{\boldsymbol{\mu}}) \approx \chi^2_{(k-p)}$.

# Comentários

- Resíduo componente do desvio (RCD). Nesse caso, é dado por

$$
\begin{aligned}
T_{D_i} &= -\frac{(2m_i|\ln(1-\widehat{p}_i)|)^{1/2}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{0\}}(y_i) + \frac{(2m_i|\ln\widehat{p}_i|)^{1/2}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{m_i\}}(y_i) \\
&+ \pm\sqrt{\frac{2}{1-\widehat{h}_{ii}}}\left\{ y_i \ln\left(\frac{y_i}{m_i\widehat{p}_i}\right)\right. \\
&+ \left.(m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - m_i\widehat{p}_i}\right)\right\}^{1/2} \mathbb{1}_{\{1,\ldots,m_i-1\}}(y_i).
\end{aligned}
$$

em que $\pm$ assume o mesmo sinal de $y_i - m_i\widehat{p}_i$,

$\widehat{h}_{ii} = m_i\widehat{p}_i(1-\widehat{p}_i)\mathbf{X}_i'\left(\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}_i$, e

$\widehat{\mathbf{V}} = \text{diag}(m_1\widehat{p}_1(1-\widehat{p}_1), ..., m_k\widehat{p}_k(1-\widehat{p}_k))$.

- $\mathbf{z} = \boldsymbol{\eta} + \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

# Estimativas dos parâmetros (os testes se referem à nulidade de cada parâmetro)

| Parâmetros | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\beta_0$ | -60,72 | 5,18 | -11,72 | $<$0,0001 |
| $\beta_1$ | 34,27 | 2,91 | 11,77 | $<$ 0,0001 |

Todos os parâmetros são significativos. Além disso, $D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}) = 11,23$, para $k - p = 8 - 2 = 6$ graus de liberdade, o que leva à um p-valor $= 0,08145$, o que sugere um ajuste apenas razoável.

# Estimativas das proporções de insetos mortos

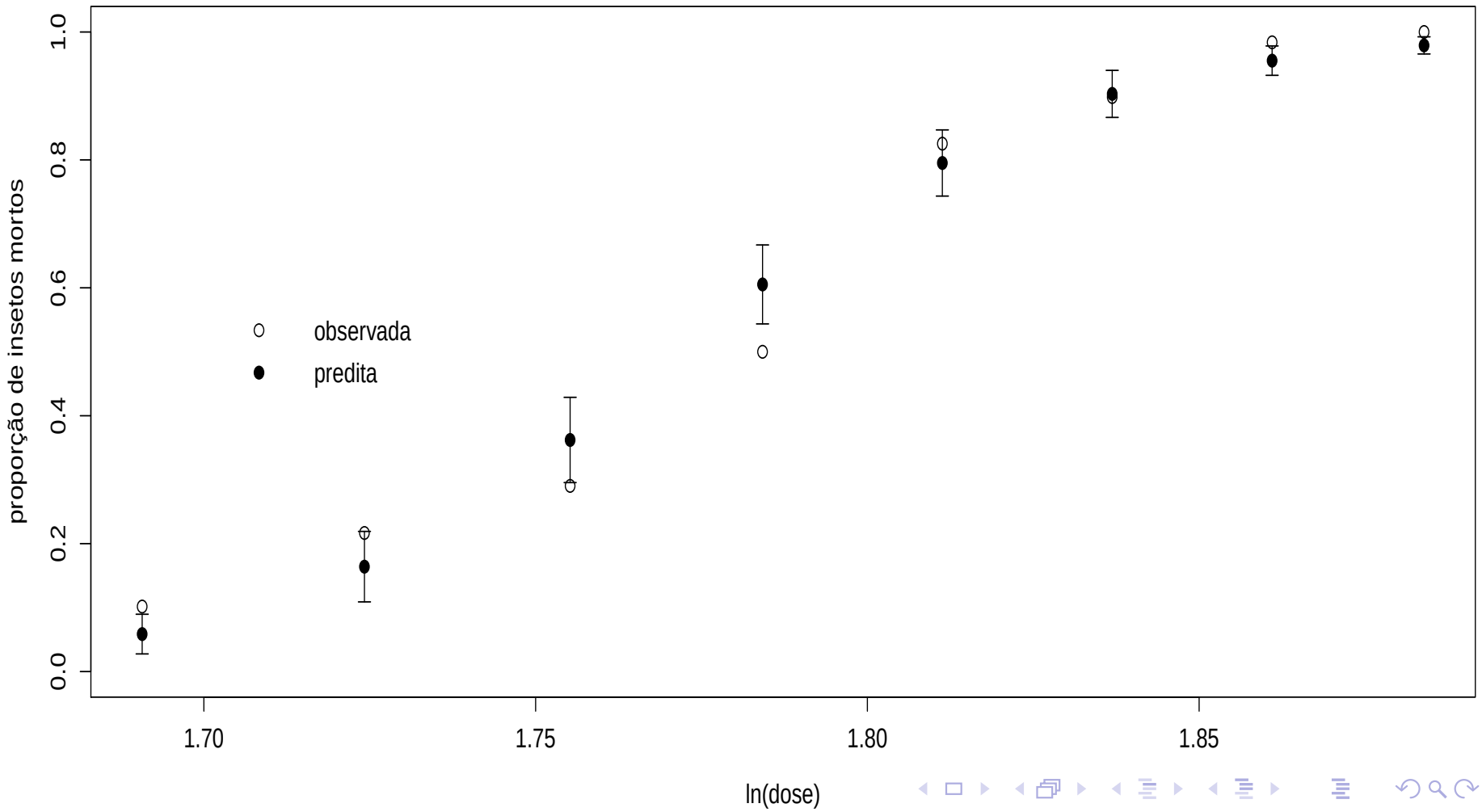
- A proporção de insetos mortos submetidos à dose $x_i$ predita pelo modelo é dada por $\widehat{p}_i = \dfrac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_i}}$.
- Pelo método delta, para $m_i, i = 1, 2, ....., 8$, suficientemente grandes, temos que $\widehat{p}_i \approx N(p_i, \boldsymbol{\Psi}_i \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\Psi}_i')$, em que

$$\boldsymbol{\Psi}_i = \begin{bmatrix} \frac{\partial}{\partial \beta_0} p_i & \frac{\partial}{\partial \beta_1} p_i \end{bmatrix}$$

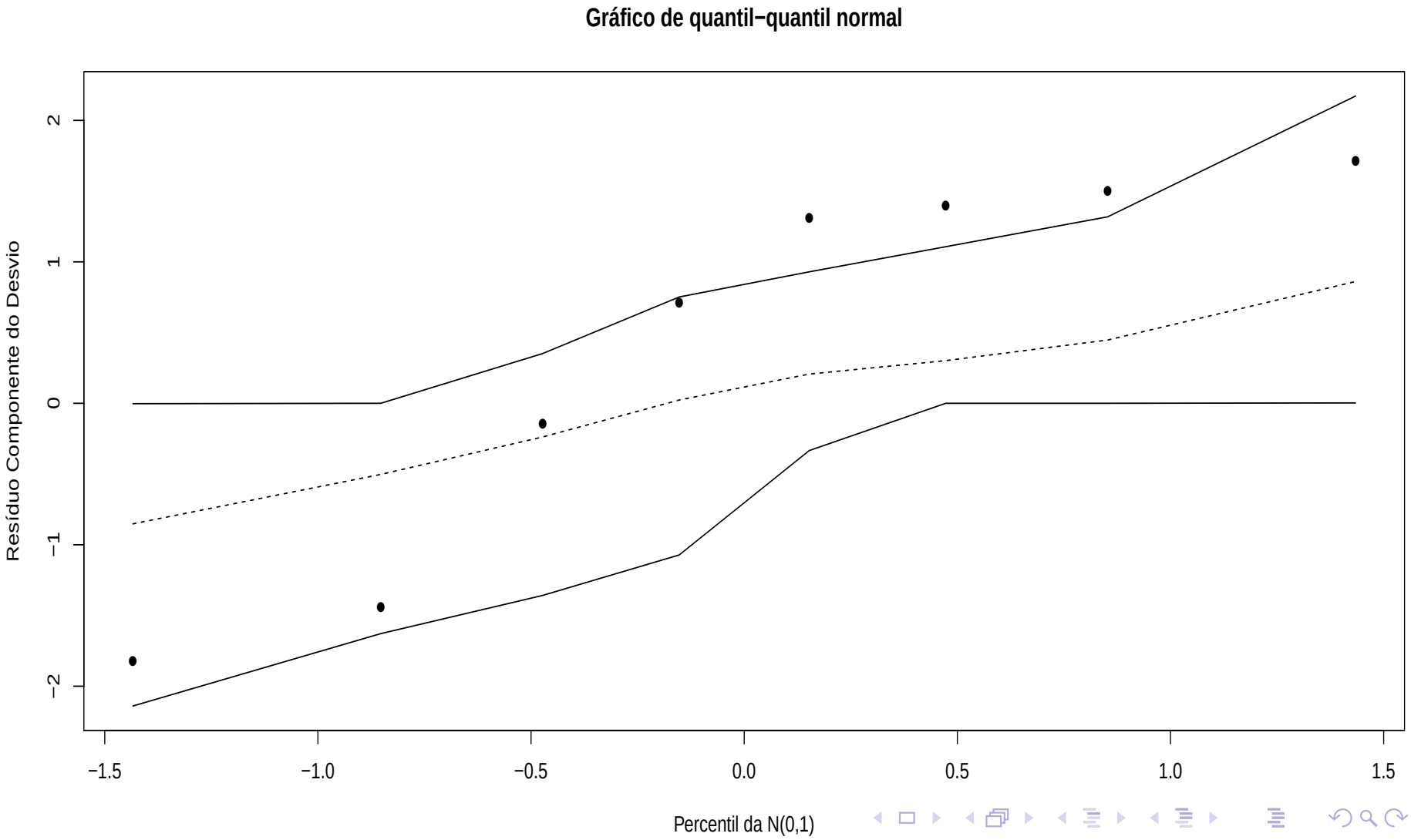
e $p_i = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. Pode-se provar que $\frac{\partial}{\partial \beta_0} p_i = p_i(1-p_i)$ e $\frac{\partial}{\partial \beta_1} p_i = p_i(1-p_i)x_i$ (**exercício**).

- Assim $IC(p_i, \gamma) = \left[\widehat{p}_i - z_{(1+\gamma)/2}\sqrt{\widehat{\psi}_i}; \widehat{p}_i + z_{(1+\gamma)/2}\sqrt{\widehat{\psi}_i}\right]$, em que $P(Z \geq z_{(1+\gamma)/2}) = \frac{1+\gamma}{2}$, $\widehat{\psi}_i = \widehat{\boldsymbol{\Psi}}_i \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{\Psi}}_i'$ e $Z \sim N(0,1)$ (lembrando que esse IC é assintótico).

# Proporções observadas $\times$ proporções preditas pelo modelo



# Gráficos de envelopes para os RCD's



# Comentários

- A análise de diagnóstico indicou que o modelo não se ajustou bem aos dados, portanto ele não pode ser utilizado para analisar os dados.
- Isso ocorreu, possivelmente, devido à função de ligação.
- Alternativas de análise: utilizar o mesmo modelo com outra função de ligação, p.e., baseada na distribuição normal assimétrica ou t assimétrica.

# Comentários

- Para finalizar (utilizando um modelo que se ajuste bem aos dados). Além de apresentar a figura anterior (com as proporções observadas e preditas) devemos estimar (pontual e intervalarmente) certas doses de letalidade de interesse do pesquisador (veja o livro do Prof. Gilberto, pags. 235 à 238).

# Voltando ao Exemplo 7: efeitos de certos fatores na sobrevivência de recém-nascidos

| idade | N. de cigarros | Sobrevivência | | |
|---|---|---|---|---|
| | | Não | Sim | Total |
| <30 | $< 5$ | $74(\theta_{(1)11})$ | $4327(\theta_{(1)12})$ | 4401 |
| | 5+ | $15(\theta_{(1)21})$ | $499(\theta_{(1)22})$ | 514 |
| 30+ | $< 5$ | $55(\theta_{(2)11})$ | $1741(\theta_{(2)12})$ | 1796 |
| | 5+ | $5(\theta_{(2)21})$ | $135\ (\theta_{(2)22})$ | 140 |

Cada linha corresponde à uma distribuição binomial.

# Modelo

- Modelo

$$
\begin{aligned}
Y_{(i)j1} &\overset{ind.}{\sim} \text{binomial}(m_{(i)j}, \theta_{(i)j1}) \\
\ln\left(\frac{\theta_{(i)j1}}{1-\theta_{(i)j1}}\right) &= \mu_i + \alpha_{(i)j}, i = 1, 2, j = 1, 2, \alpha_{(i)1} = 0, i = 1, 2.
\end{aligned}
$$

- $m_{(i)j}$ : número total de recém nascidos de mães que fumam uma quantidade $j$ de cigarros por dia e que pertencem ao grupo $i$ da idade.
- $Y_{(i)j1}$ : número de recém nascidos que vieram à óbito, de mães que fumam uma quantidade $j$ de cigarros por dia e que pertencem ao grupo $i$ da idade.
- $\boldsymbol{\beta} = (\mu_1, \mu_2, \alpha_{(1)2}, \alpha_{(2)2})'$. Note que este modelo é saturado (n = p).

# Logitos, parâmetros e quantidades observadas

$$
\begin{aligned}
\ln\left(\frac{\theta_{(1)11}}{1-\theta_{(1)11}}\right) &= \mu_1 \Rightarrow \theta_{(1)11} = \frac{e^{\mu_1}}{1+e^{\mu_1}} \\
\ln\left(\frac{\theta_{(1)21}}{1-\theta_{(1)21}}\right) &= \mu_1 + \alpha_{(1)2} \Rightarrow \theta_{(1)11} = \frac{e^{\mu_1+\alpha_{(1)2}}}{1+e^{\mu_1+\alpha_{(1)2}}} \\
\ln\left(\frac{\theta_{(2)11}}{1-\theta_{(2)11}}\right) &= \mu_2 \Rightarrow \theta_{(1)11} = \frac{e^{\mu_2}}{1+e^{\mu_2}} \\
\ln\left(\frac{\theta_{(2)21}}{1-\theta_{(2)21}}\right) &= \mu_2 + \alpha_{(2)2} \Rightarrow \theta_{(1)11} = \frac{e^{\mu_2+\alpha_{(2)2}}}{1+e^{\mu_2+\alpha_{(2)2}}}
\end{aligned}
$$

Além disso, $m_{(1)1} = 4401$, $m_{(1)2} = 514$, $m_{(2)1} = 1796$, $m_{(2)2} = 140$ e $y_{(1)11} = 74$, $y_{(1)21} = 15$, $y_{(2)11} = 55$, $y_{(2)21} = 5$. **Exercício: escrever as razões de chance de interesse em função dos parâmetros $\beta$, interpretando os parâmetros $(\alpha_{(1)2}, \alpha_{(2)2})'$.**

# Hipóteses de interesse

- Ausência de independência, entre sobreviênvia e n. de cigarros, para cada uma das subpopulações (idade).

$$H_0 : \begin{cases} \theta_{(1)11} = \theta_{(1)21} \\ \theta_{(2)11} = \theta_{(2)21} \end{cases} \leftrightarrow \begin{cases} \alpha_{(1)2} = 0 \\ \alpha_{(2)2} = 0 \end{cases}$$

vs $H_1$ : há pelo menos uma diferença

- Como testar as hipóteses acima? Através de testes individuais de nulidade, testes do tipo $\mathbf{C}\boldsymbol{\beta} = \mathbf{M}$, teste da razão de verossimilhanças, análise do desvio (quando os modelos envolvidos são não saturados).

# Testes da Razão de verossimilhanças

- Vamos supor para o vetor de parâmetros $\boldsymbol{\beta}$ a partição $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, em que $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são vetores de dimensão $q \times 1$ e $(p-q) \times 1$, respectivamente.
- Desejamos testar $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$.
- Sejam $D(\mathbf{y}; \widehat{\boldsymbol{\mu}}^{(0)})$ e $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ os desvios dos modelos: ajustados sob $H_0$ e irrestrito, respectivamente. Lembremos que $D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2\{l(\mathbf{Y}, \mathbf{y}) - l(\widehat{\boldsymbol{\mu}}, \mathbf{y})\}$.

# Testes da Razão de verossimilhanças

- Defina a seguinte estatística:

  $Q_{RV} = D(\mathbf{y}; \widehat{\boldsymbol{\mu}}^{(0)}) - D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = -2(l(\widehat{\boldsymbol{\mu}}, \mathbf{y}) - l(\widehat{\boldsymbol{\mu}}^{(0)}, \mathbf{y}))$ (**exercício**).

  Sob $H_0$, para $n$ suficientemente grande, $Q_{RV} \approx \chi^2_q$

- Assim, rejeita-se $H_0$ se $p - valor \leq \alpha$, em que

  $p - valor \approx P(X \geq q_{RV}|H_0)$, em que $X \sim \chi^2_q$

  $q_{RV} = D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}^{(0)}) - D(\mathbf{y}; \widetilde{\boldsymbol{\mu}})$.

# Análise do desvio

- Baseados no teste da RV, podemos ainda definir um outro procedimento para testar as hipóteses $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}$.
- A estatística $Q_{AD} = \dfrac{\left(D(\mathbf{y}; \widehat{\boldsymbol{\mu}}^{(0)}) - D(\mathbf{y}; \widehat{\boldsymbol{\mu}})\right)/q}{D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/(n-p)}$ sob $H_0$ e para $n$ suficientemente grande, é tal que $Q_{AD} \approx F_{(q,n-p)}$
- Note que só podemos utilizar esta abordagem para modelos não saturados (n $>$ p).
- Assim, rejeita-se $H_0$ se $p - valor \leq \alpha$, em que $p - valor \approx P(X \geq q_{AD}|H_0)$, em que $X \sim F_{(q,n-p)}$

$$q_{AD} = \frac{\left(D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}^{(0)}) - D(\mathbf{y}; \widetilde{\boldsymbol{\mu}})\right)/q}{D(\mathbf{y}; \widetilde{\boldsymbol{\mu}})/(n-p)}.$$

# Ajuste do modelo

| Parâmetro | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\mu_1$ | -4,069 | 0,117 | -34,70 | $<$ 0,0001 |
| $\alpha_{(1)2}$ | 0,564 | 0,2871 | 1,96 | 0,0495 |
| $\mu_2$ | -3,455 | 0,137 | -25,23 | $<$ 0,0001 |
| $\alpha_{(2)2}$ | 0,159 | 0,4756 | 0,33 | 0,7381 |

Há uma significância marginal da hipótese alternativa.

# Outro testes

- Teste $\mathbf{C}\boldsymbol{\beta} = \mathbf{M}$, $q_t = 3,97$, $p - valor = 0,1373$. **Exercício: encontrar as matrizes C e M.**
- Teste da RV, $q_{RV} = 3,53$, $p - valor = 0,1714$.
- Assim, não rejeitamos a hipótese de independência simultânea.
- **Exercício:** Ajustar o modelo reduzido ($\alpha_{(2)1} = \alpha_{(2)2} = 0$) e estimar, pontual e intervalarmente, as probabilidades de interesse ($(\theta_{(1)11}, \theta_{(1)21}, \theta_{(2)11}, \theta_{(2)21})'$) sob esse modelo.

# Exemplo 13: preferência de consumidores com relação à marcas de carros

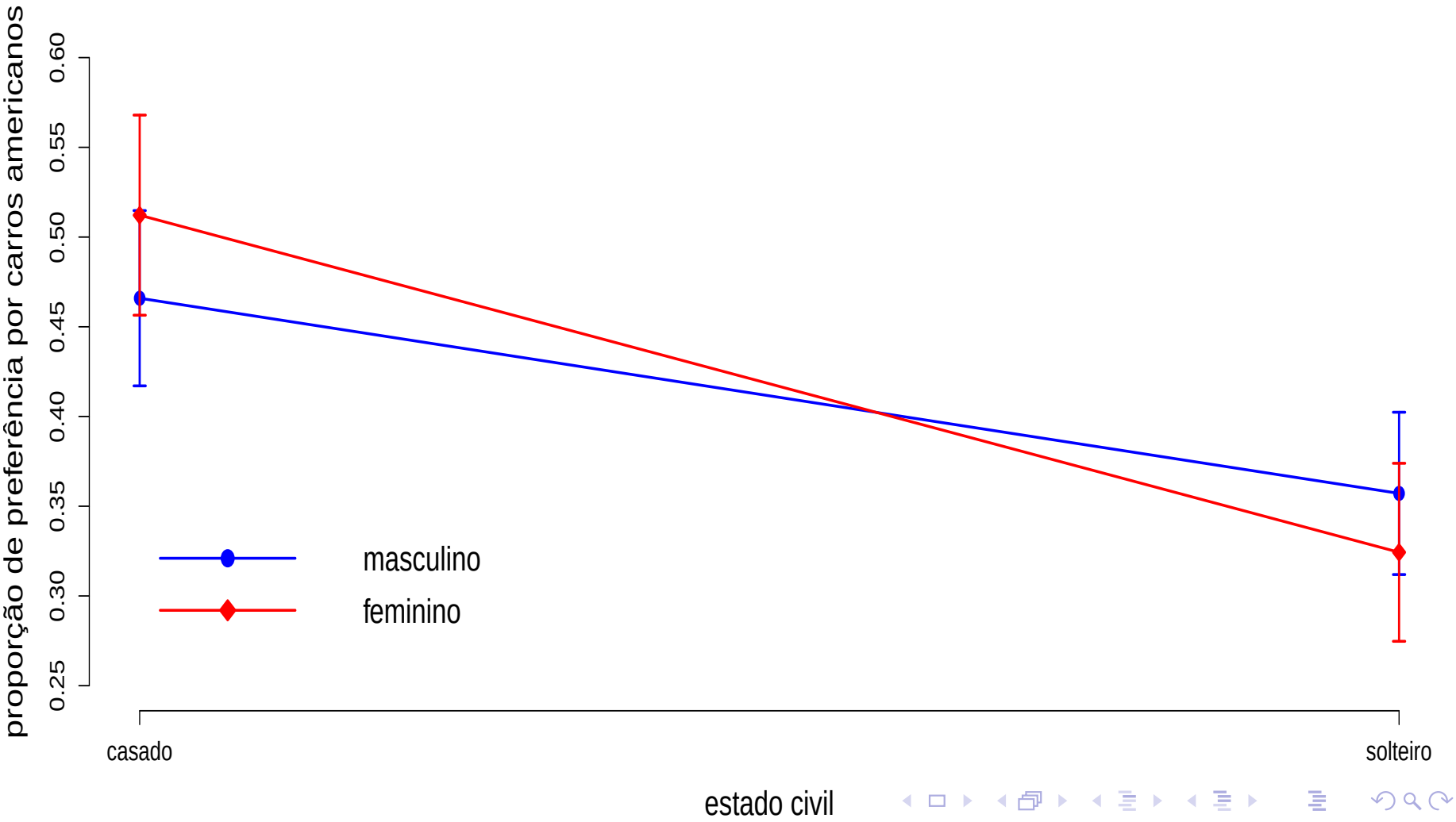
- Uma amostra aleatória de 263 consumidores foi considerada.
- As seguintes variáveis foram observadas para cada comprador: preferência do tipo de automóvel (1: americano, 0: japonês), idade (em anos), sexo (0: masculino; 1: feminino) e estado civil (0: casado, 1: solteiro).
- Variável resposta: preferência do tipo de automóvel.
- Para maiores detalhes ver Foster, Stine e Waterman (1998, pgs. 338-339).

# Análise descritiva

Os percentuais foram calculados dentro de cada categoria de gênero e estado civil (os percentuais dentro de cada linha somam 100%).

| | preferência | |
|---|---|---|
| gênero | japonês | americano |
| masculino | 57,64 | 42,36 |
| feminino | 54,62 | 45,38 |
| estado civil | | |
| casado | 51,18 | 48,82 |
| solteiro | 65,59 | 34,41 |

# Gráficos de perfis



# Modelo

- Modelo

$$
\begin{aligned}
Y_{ijk} &\overset{ind.}{\sim} \text{Bernoulli}(\theta_{ij}) \\
\ln\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, i = 1,2, j = 1,2, k = 1,2,...,n_{ij} \\
\alpha_1 = \beta_1 &= (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0, \forall i, j.
\end{aligned}
$$

- $n_{ij}$ : número total de consumidores pertecentes ao $i$-ésimo gênero (1: masculino, 2: feminino) e ao $j$-ésimo estado civil (1: casado,2:solteiro), $n_{11} = 88, n_{12} = 56, n_{21} = 82, n_{22} = 37$.
- $Y_{ijk}$ : 1 se o $k$-ésimo consumidor pertencente ao $i$-ésimo gênero e ao $j$-ésimo estado civil prefere carros americanos e 0, caso ele prefira carros japoneses.

# Modelo

- $\boldsymbol{\beta} = (\mu, \alpha_2, \beta_2, (\alpha\beta)_{22})'$.
- Logitos

$$
\begin{aligned}
\ln\left(\frac{\theta_{11}}{1-\theta_{11}}\right) &= \mu \Rightarrow \theta_{11} = \frac{e^{\mu}}{1+e^{\mu}} \\
\ln\left(\frac{\theta_{21}}{1-\theta_{21}}\right) &= \mu + \alpha_2 \Rightarrow \theta_{21} = \frac{e^{\mu+\alpha_2}}{1+e^{\mu+\alpha_2}} \\
\ln\left(\frac{\theta_{12}}{1-\theta_{12}}\right) &= \mu + \beta_2 \Rightarrow \theta_{12} = \frac{e^{\mu+\beta_2}}{1+e^{\mu+\beta_2}} \\
\ln\left(\frac{\theta_{22}}{1-\theta_{22}}\right) &= \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \Rightarrow \theta_{22} = \frac{e^{\mu+\alpha_2+\beta_2+(\alpha\beta)_{22}}}{1+e^{\mu+\alpha_2+\beta_2+(\alpha\beta)_{22}}}
\end{aligned}
$$

# Modelo

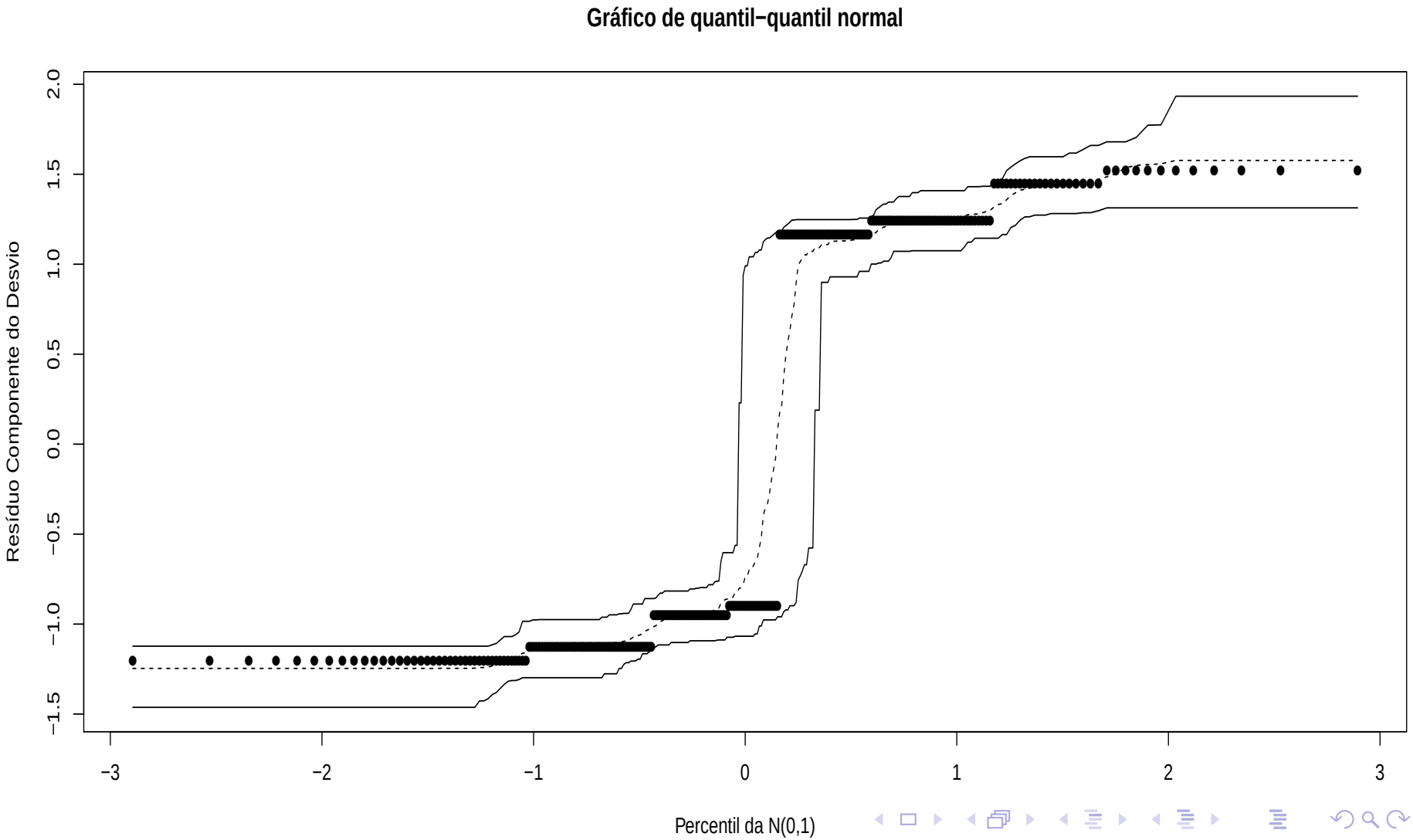
- Os parâmetros seguem as interpretações usuais, mas agora em termos das probabilidades e das razões de chances.
- **Exercício: provar que o parâmetro $(\alpha\beta)_{22}$ está relacionado com a presença de interação entre os fatores.**
- **Exercício: interprete os parâmetros $(\alpha_2, \beta_2)'$ em termos de razões de chances, dado a presença de interação.**
- **Exercício: provar que os parâmetros $(\alpha_2, \beta_2)'$ estão relacionados com a presença dos efeitos dos seus respectivos fatores, dado a ausência de interação.**

# Ajuste do modelo completo

| Parâmetro | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\mu$ | -0,137 | 0,214 | -0,639 | 0,5228 |
| $\alpha_2$ | 0,185 | 0,307 | 0,603 | 0,5465 |
| $\beta_2$ | -0,451 | 0,351 | -1,284 | 0,1991 |
| $(\alpha\beta)_{22}$ | -0,332 | 0,5437 | -0,610 | 0,5420 |

Aparentemente, nenhum coeficiente é significativo. Entretanto, vamos explorar o modelo um pouco melhor.

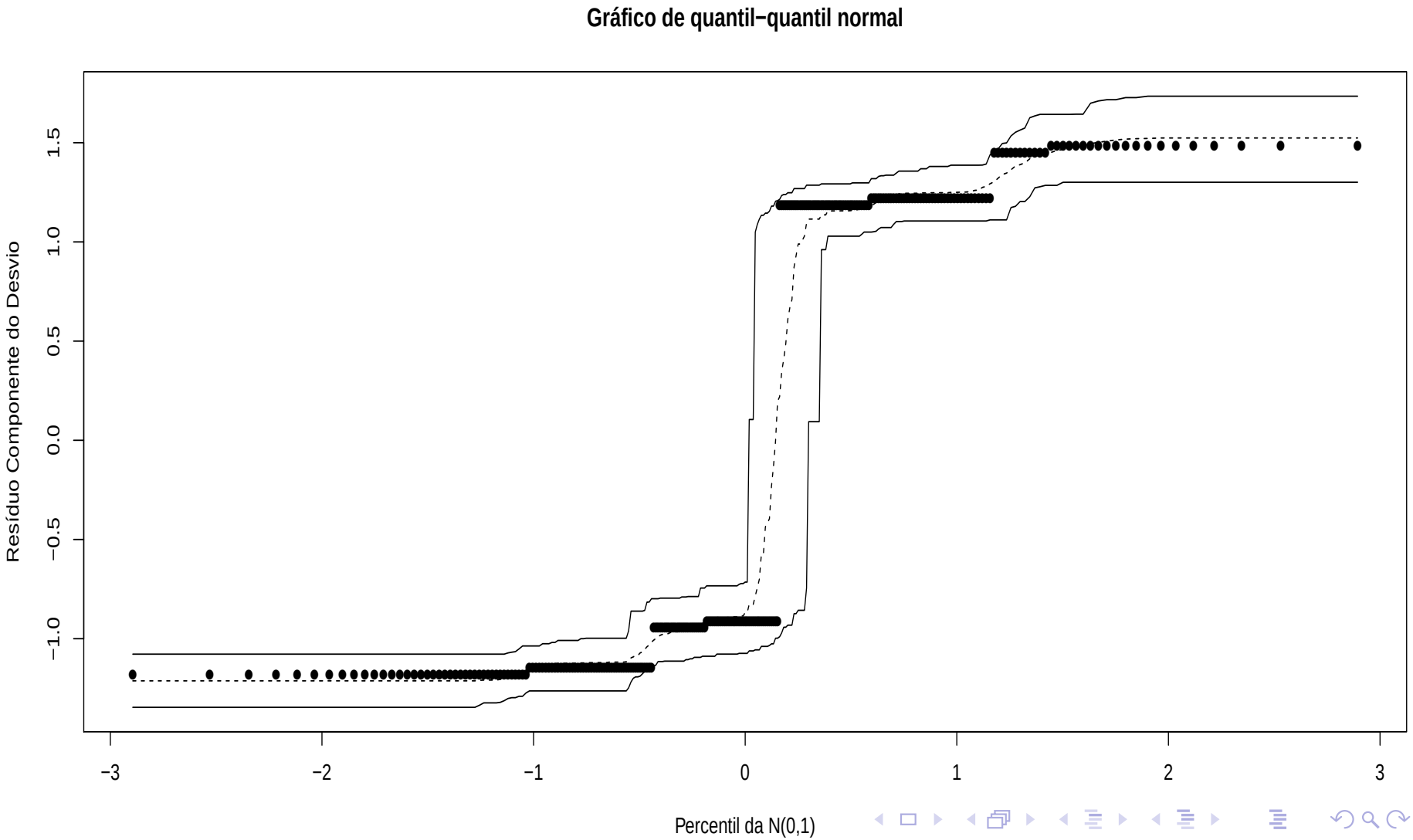
# Gráficos de envelopes para o RCD



# Ajuste do modelo sem interação

| Parâmetro | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\mu$ | -0,085 | 0,196 | -0,434 | 0,6642 |
| $\alpha_2$ | 0,079 | 0,253 | 0,312 | 0,7551 |
| $\beta_2$ | -0,592 | 0,268 | -2,211 | 0,0270 |

# Gráficos de envelopes para o RCD

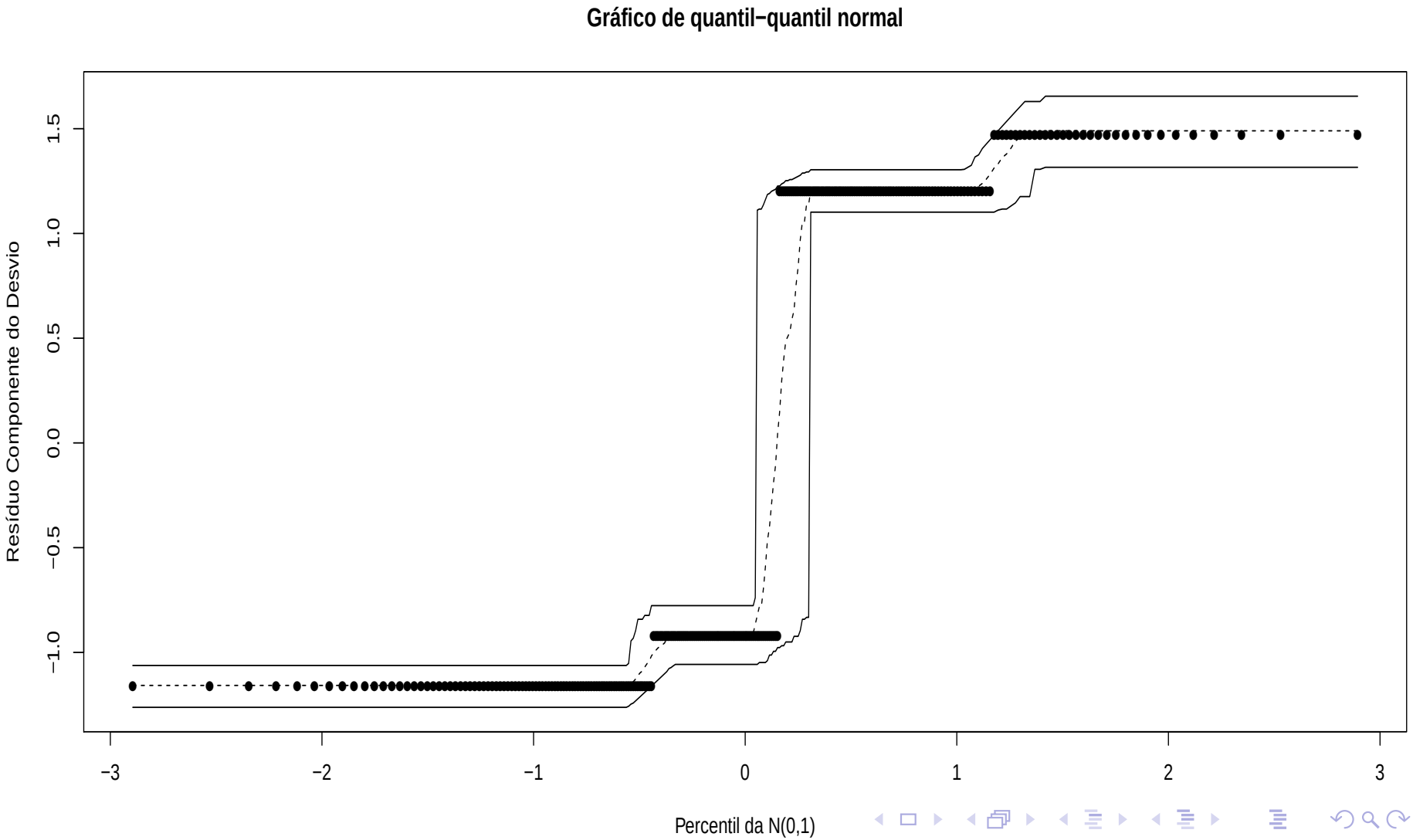


# Ajuste do modelo com somente o fator estado civil

| Parâmetro | Estimativa | EP | Estat. $Z_t$ | p-valor |
|---|---|---|---|---|
| $\mu$ | -0,0471 | 0,153 | -0,307 | 0,7590 |
| $\beta_2$ | -0,5981 | 0,267 | -2,242 | 0,0250 |

Modelo final: fator estado civil paracer ser significativo.

# Gráficos de envelopes para o RCD

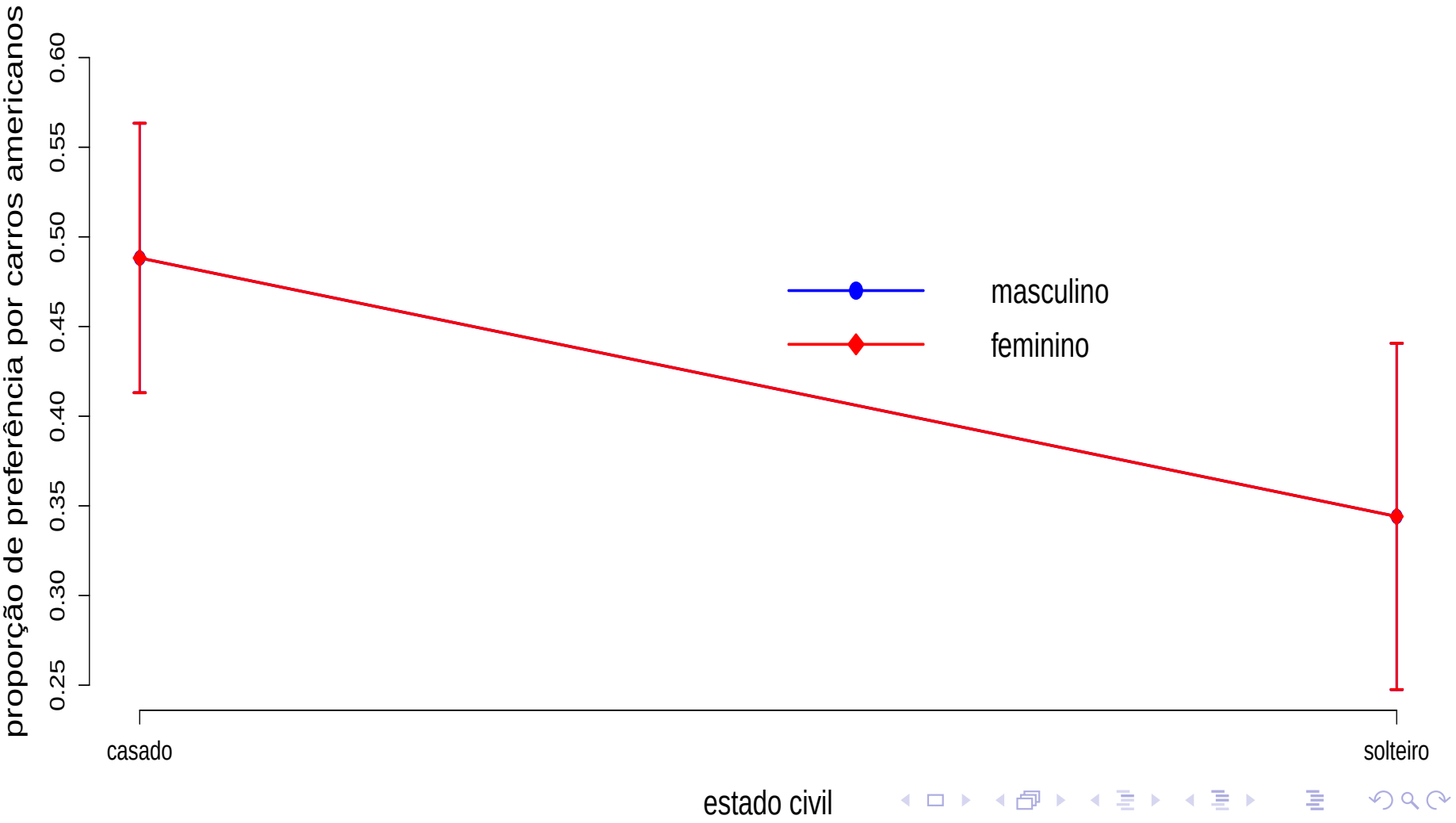


# Percentuais preditos pelo modelo final (através do método delta)

| Estado civil | Gênero | Estimativa | EP | IC(95%) |
|---|---|---|---|---|
| Casado | Masculino | 48,82 | 3,83 | [41,31 ;56,34] |
| Solteiro | Masculino | 34,41 | 4,93 | [24,75 ; 44,06] |
| Casado | Feminino | 48,82 | 3,83 | [41,31 ; 56,34] |
| Solteiro | Feminino | 34,41 | 4,93 | [24,75 ; 44,06] |

**Exercício: obter os resultados acima aplicando o método delta.**

# Proporções preditas pelo modelo final



# Seleção de modelos

- Vimos como verificar se um determinado modelo se ajusta adequadamente aos dados.
- Uma outra questão de interesse surge quando se dispõe de diversos modelos (que se ajustam adequadamente aos dados) e respondem às perguntas de interesse, e queremos escolher um como o "mais apropriado".
- Há diversas técnicas disponíveis para este fim.
- Veremos técnicas baseadas em testes de hipótese e comparação de estatísticas de qualidade de ajuste.

# Teste da razão de verossimilhanças

- Sejam $M_1$ e $M_2$ dois modelos, em que $M_1$ está encaixado em $M_2$, ou seja, o modelo $M_1$ é um caso particular de $M_2$.
- Por exemplo, $M_1$ é um modelo linear obtido de $M_2$, o qual é um modelo quadrático.
- Neste caso temos que

  $H_0$ : **o modelo** $M_1$ **é preferível ao modelo** $M_2$ **vs** $H_1$ : **o modelo** $M_2$ **é preferível ao modelo** $M_1$.

# Teste da razão de verossimilhanças (cont.)

- Seja $\widehat{\boldsymbol{\theta}}_i$ o estimador de máxima verossimilhança obtido sob o modelo $i$ e $\widetilde{\boldsymbol{\theta}}_i$ sua respectiva estimativa.
- Denote por $L_i(\widehat{\boldsymbol{\theta}})$ e $l_i(\widehat{\boldsymbol{\theta}})$ o máximo da verossimilhança e da log-verossimilhança do modelo $i$, respectivamente, em relação aos estimadores enquanto que $L_i(\widetilde{\boldsymbol{\theta}})$ e $l_i(\widetilde{\boldsymbol{\theta}})$ são os respectivos máximos avaliados nas estimativas.

# Teste da razão de verossimilhanças (cont.)

- A estatística do TRV é dada por $\Delta = \frac{L_1(\widehat{\boldsymbol{\theta}}_1)}{L_2(\widehat{\boldsymbol{\theta}}_2)}$.
- Rejeita-se $H_0$ se $\Delta \geq \delta_c$, em que $\delta_c$ é um valor crítico adequado.
- Alternativamente, rejeitamos $H_0$ se

$$\Lambda = -2ln(\Delta) = -2\left(l_1(\widehat{\boldsymbol{\theta}}_1) - l_2(\widehat{\boldsymbol{\theta}}_2)\right) \geq \lambda_c,$$

em que $P(Q \geq \lambda_c) = \alpha$, $Q \approx \chi^2_{(\gamma)}$ e

$\gamma =$ número de parâmetros do modelo $M_2$ - número de parâmetros do modelo $M_1$.

- Nesse caso, $p - valor \approx P(Q \geq \lambda | H_0)$, em que $\lambda$ é o valor observado da estatística $\Lambda$ e $Q \sim \chi^2_\gamma$. Assim, rejeita-e $H_0$ se $p - valor \leq \alpha$.

# Estatísticas de comparação de modelos

- O TRV é apropriado na comparação somente de modelos encaixados (o modelo com menor número de parâmetros é um caso particular do modelo com maior número de parâmetros).
- Além disso, ele não leva em consideração (diretamente) o número de parâmetros do modelo (somente na distribuição da estatística).
- Existem várias alternativas, em termos de estatísticas para comparar modelos, que "penalizam" a verossimilhança em relação ao número de parâmetros, tamanho da amostra entre outros fatores.
- Veremos o AIC e o BIC.

# Estatísticas de comparação de modelos (cont.)

- O AIC e BIC, para o i-ésimo modelo, são dados, respectivamente, por:

$$\begin{aligned} AIC_i &= -2l_i(\widetilde{\theta}_i) + 2k \\ BIC_i &= -2l_i(\widetilde{\theta}_i) + 2k\ln(n) \end{aligned}$$

que $l_i(\widetilde{\theta}_i)$ denota a log-verossimilhança do i-ésimo modelo avaliada em alguma estimativa (p.e. máxima verossimilhança), $k$ é o número de parâmetros e $n$ é o número de observações.

- Portanto, o modelo que apresentar os menores valores, será o modelo "melhor ajustado" aos dados.

# Métodos de seleção "dinâmico" ou automatizados

- Existem métodos que selecionam modelos, fixados alguns critérios, de modo "dinâmico" (automozatizado).
- Veremos os métodos "forward", "backward" e "stepwise".
- Tais métodos são particularmente úteis quanto se dispões de muitas covariáveis e/ou muitos fatores.
- Sem perda de generalidade, vamos considerar um determinado modelo (normal linear, linear generalizado) tal que o preditor linear é dado por

$$\eta_{ij} = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$$

# Método "forward"

- Primeiramente, ajustamos um modelo com somente o intercepto, ou seja $\eta_{ij} = \beta_0$. Ajustamos então, para cada variável explicativa, um modelo

$$\eta_{ij} = \beta_0 + \beta_j x_{ij}, j = 1, 2, ..., p - 1$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=1,2,...,p-1 (usando-se algum teste como o TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja $P$ o menor nível descritivo entre os $p - 1$ testes. Se $P \leq P_E$ a variável correspondente entra no modelo (caso contrário, o processo é interrompido).

# Métodos "forward" (cont.)

- Vamor supor que a variável $X_1$ foi escolhida. Então, no passo seguinte, ajustamos os modelos

$$\eta_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_j x_{ij}, j = 2, ..., p-1$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=2,...,p-1 (usando-se algum teste como TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja $P$ o menor nível descritivo entre os $p-2$ testes. Se $P \leq P_E$ a variável correspondente entra no modelo. Repetimos o procedimento até que ocorra $P > P_E$.

# Método "backward"

- Primeiramente, ajustamos o seguinte modelo:

$$\eta_{ij} = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=1,2,...,p-1 (usando-se algum teste como o TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja $P$ o maior nível descritivo entre os $p-1$ testes. Se $P > P_S$ a variável correspondente sai do modelo (caso contrário, o processo é interrompido).

# Método "backward" (cont.)

- Vamos supor que $X_1$ tenha saído do modelo. Então ajustamos o seguinte modelo

$$\eta_{ij} = \beta_0 + \sum_{j=2}^{p-1} \beta_j x_{ij}$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=2,...,p-1 (usando-se algum teste como TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja $P$ o maior nível descritivo entre os $p-2$ testes. Se $P > P_S$ a variável correspondente sai do modelo. Repetimos o procedimento até que ocorra $P \leq P_S$.

# Método "stepwise"

- É uma mistura dos dois procedimentos anteriores.
- Iniciamos o processo com o modelo $\eta_{ij} = \beta_0$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira sai ou não do modelo.
- O processo continua até que nenhuma variável seja incluída ou retirada do modelo.
- Geralmente adotamos $0,15 \leq P_E, P_S \leq 0,25$. Outra possibilidade é usar $P_E = P_S = 0,20$.

# Métodos anteriores usando AIC/BIC

- Para qualquer um dos métodos anteriores, se usarmos alguma estatística de comparação de modelos (como AIC ou BIC), procedemos da seguinte forma
  - Sempre escolhemos o modelo (retirar/incluir a variável) que apresentar o menor valor da estatística.
  - O processo é interrompido quando as estatísticas para todos os modelos possíveis aumentarem em relação ao modelo corrente.
- Observação: as estatísticas AIC e BIC também servem para comparar modelos que difiram em termos da função de ligação e distribuição da variável resposta, entre outras características.

# Aplicação no exemplo 13

- Aplicou-se cada um dos três métodos, forward, backward e stepwise, no exemplo anterior, através da estatística AIC.
- As três abordagens escolheram o modelo que contempla somente o intercepto e o fator estado civil.

# Utilização da variável idade

- Modelo 1:

$$\begin{aligned} \ln\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma(x_{ijk} - \overline{x}), \\ & \quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\ \alpha_1 = \beta_1 &= (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0, \forall i, j. \end{aligned}$$

em que $x_{ijk}$ é a idade do $k$-ésimo indivíduo do gênero $i$ e do estado civil $j$ e $\overline{x} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n_{ij}} x_{ijk}$, $n = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$. Para $x_{ijk} = \overline{x}$ e/ou para indivíduos com a mesma idade, os parâmetros $(\mu, \alpha_2, \beta_2, (\alpha\beta)_{22})'$ possuem a mesma interpretação anterior, enquanto que $\gamma$ é o incremento no logito para o aumento em uma unidade da variável idade.

# Utilização da variável idade

- Modelo 2:

$$\begin{aligned} \ln\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ij}(x_{ijk} - \overline{x}), \\ & \quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\ \alpha_1 = \beta_1 &= (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0, \forall i, j. \end{aligned}$$

Para $x_{ijk} = \overline{x}$ e/ou para indivíduos com a mesma idade e pertencentes ao mesmo grupo, os parâmetros $(\mu, \alpha_2, \beta_2, (\alpha\beta)_{22})'$ possuem a mesma interpretação anterior, enquanto que $\gamma_{ij}$ continua sendo o incremento no logito para o aumento em uma unidade da variável idade, agora para cada grupo.

# Modelos finais (usando o método stepwise)

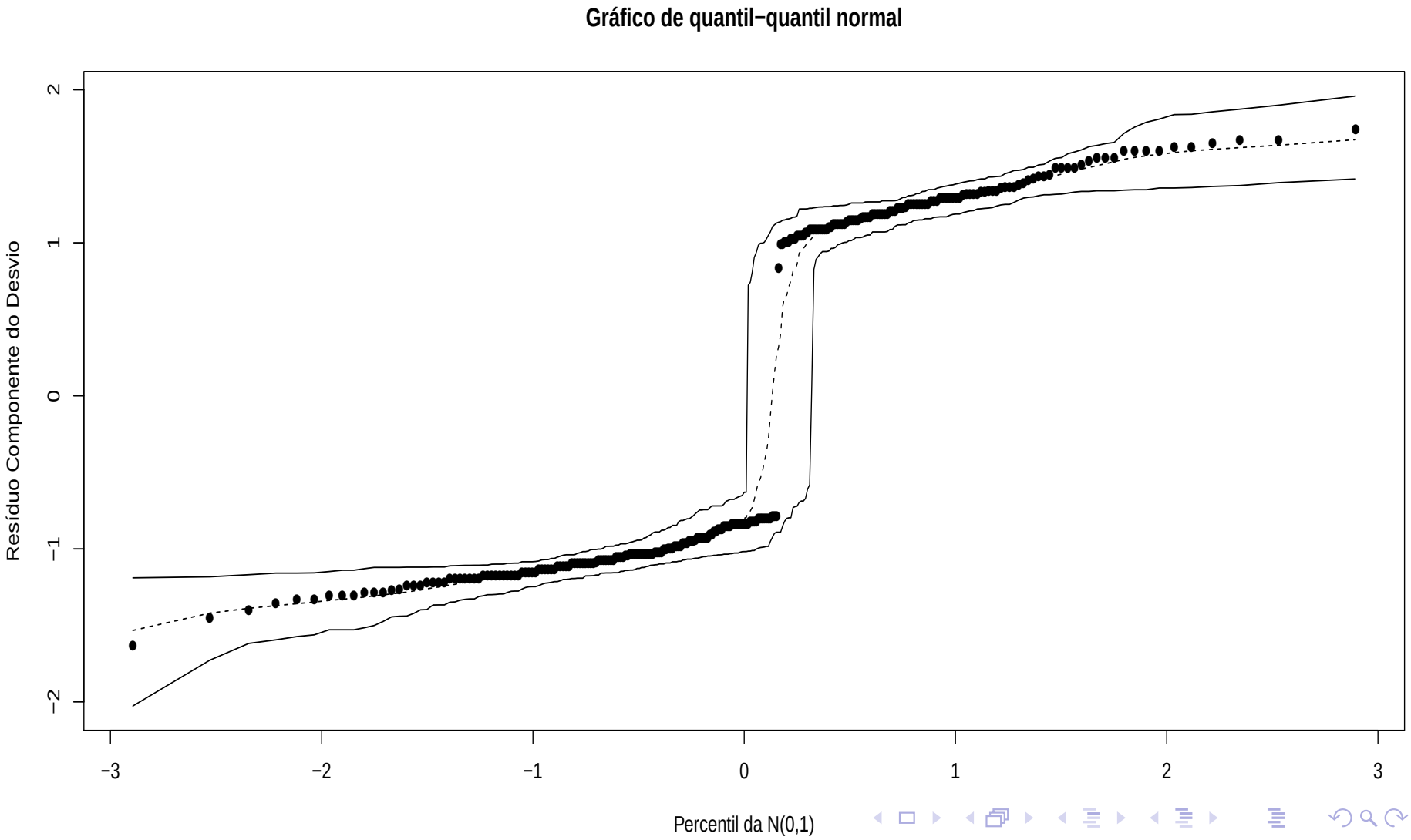
- Modelo 1:

$$\begin{aligned} \ln\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \mu + \beta_j + \gamma(x_{ijk} - \overline{x}), \\ & \quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\ \beta_1 &= 0. \end{aligned}$$

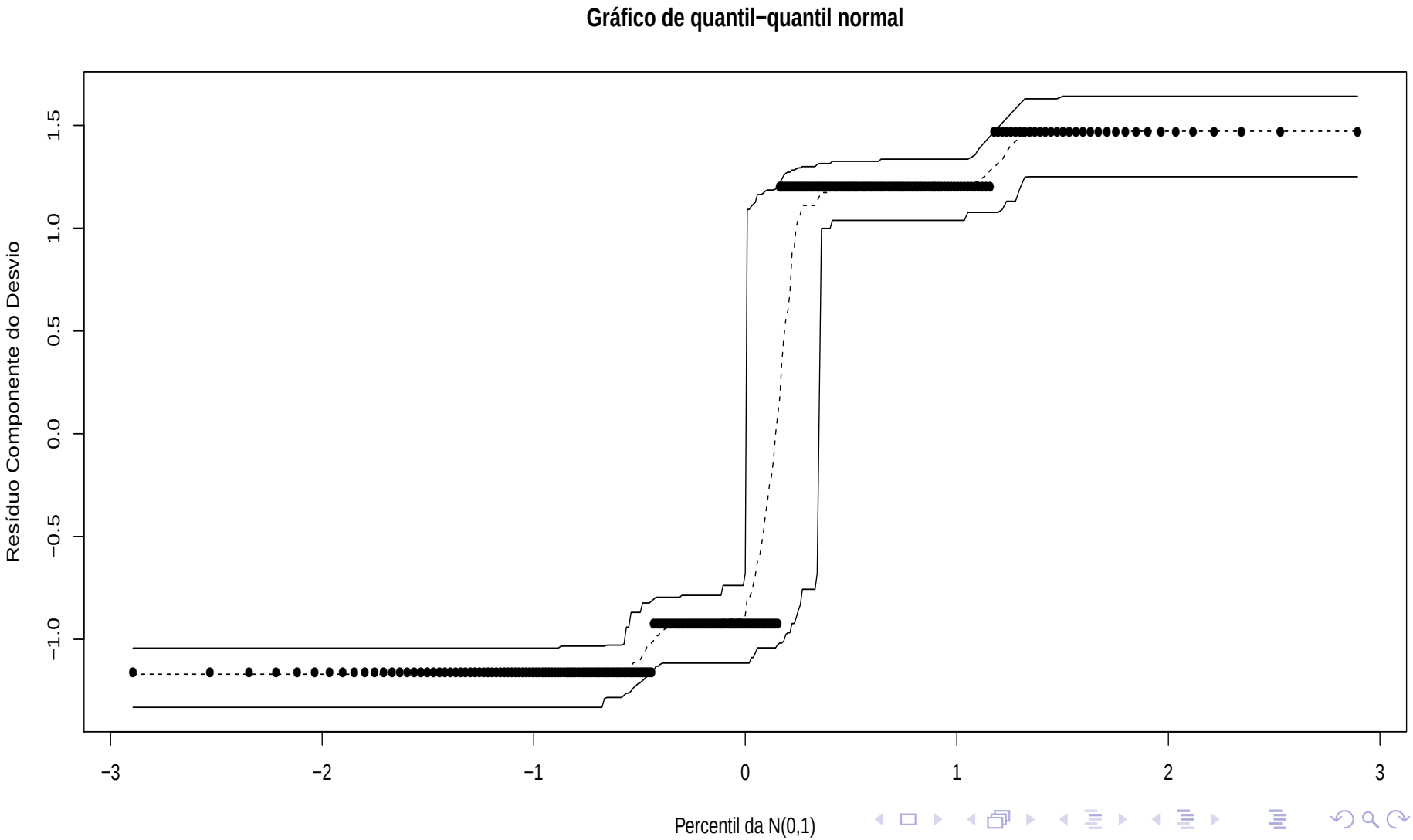
- Modelo 2:

$$\begin{aligned} \ln\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right) &= \mu + \beta_j \\ & \quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\ \beta_1 &= 0. \end{aligned}$$

# Gráficos de envelopes para os RCD's do modelo final 1



# Gráficos de envelopes para os RCD's do modelo final 2



# Comentários

- As análises de diagnósticos indicaram que os modelos se ajustam bem aos dados.
- Em relação aos dois últimos modelos, através de algum deles, devemos apresentar as estimativas pontuais e intervalares de probabilidades de interesse (em função dos fatores e/ou da idade) e dos parâmetros do modelo.