Modelos de regressão para dados discretos (parte 3): dados binários

Prof. Caio Azevedo

Exemplo 9: preferência de consumidores com relação à marcas de carros

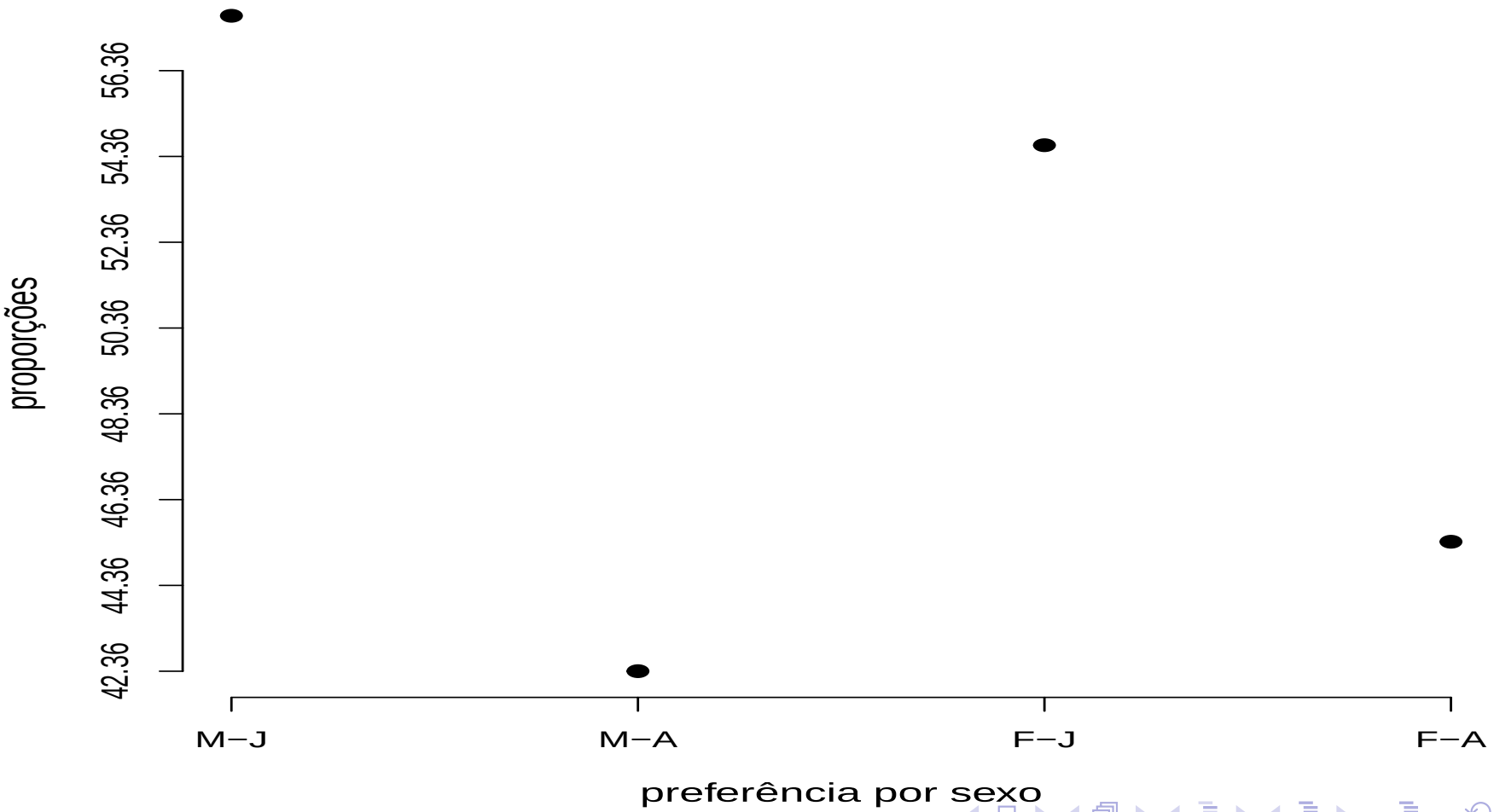
- Uma amostra aleatória de 263 consumidores foi considerada.
- As seguintes variáveis foram observadas para cada comprador: preferência do tipo de automóvel (1: americano, 0: japonês), idade (em anos), sexo (0: masculino; 1: feminino) e estado civil (0: casado, 1: solteiro).
- Variável resposta: preferência do tipo de automóvel. Por enquanto, vamos desconsiderar a variável idade.
- Para maiores detalhes ver Foster, Stine e Waterman (1998, pgs. 338-339) e Paula (2024).

Análise descritiva

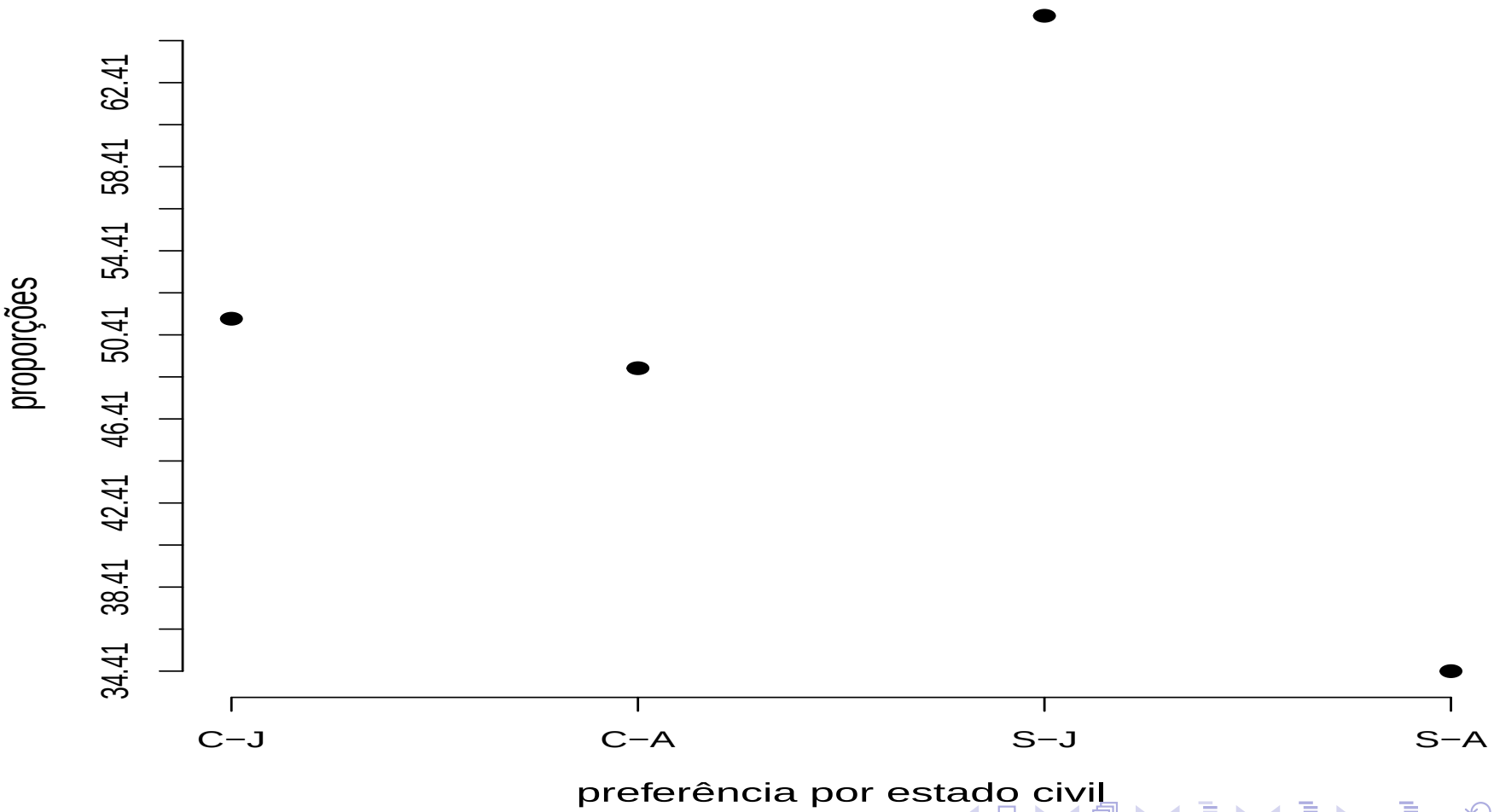
- Os percentuais foram calculados dentro de cada categoria de sexo e estado civil (os percentuais dentro de cada linha somam 100%).

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

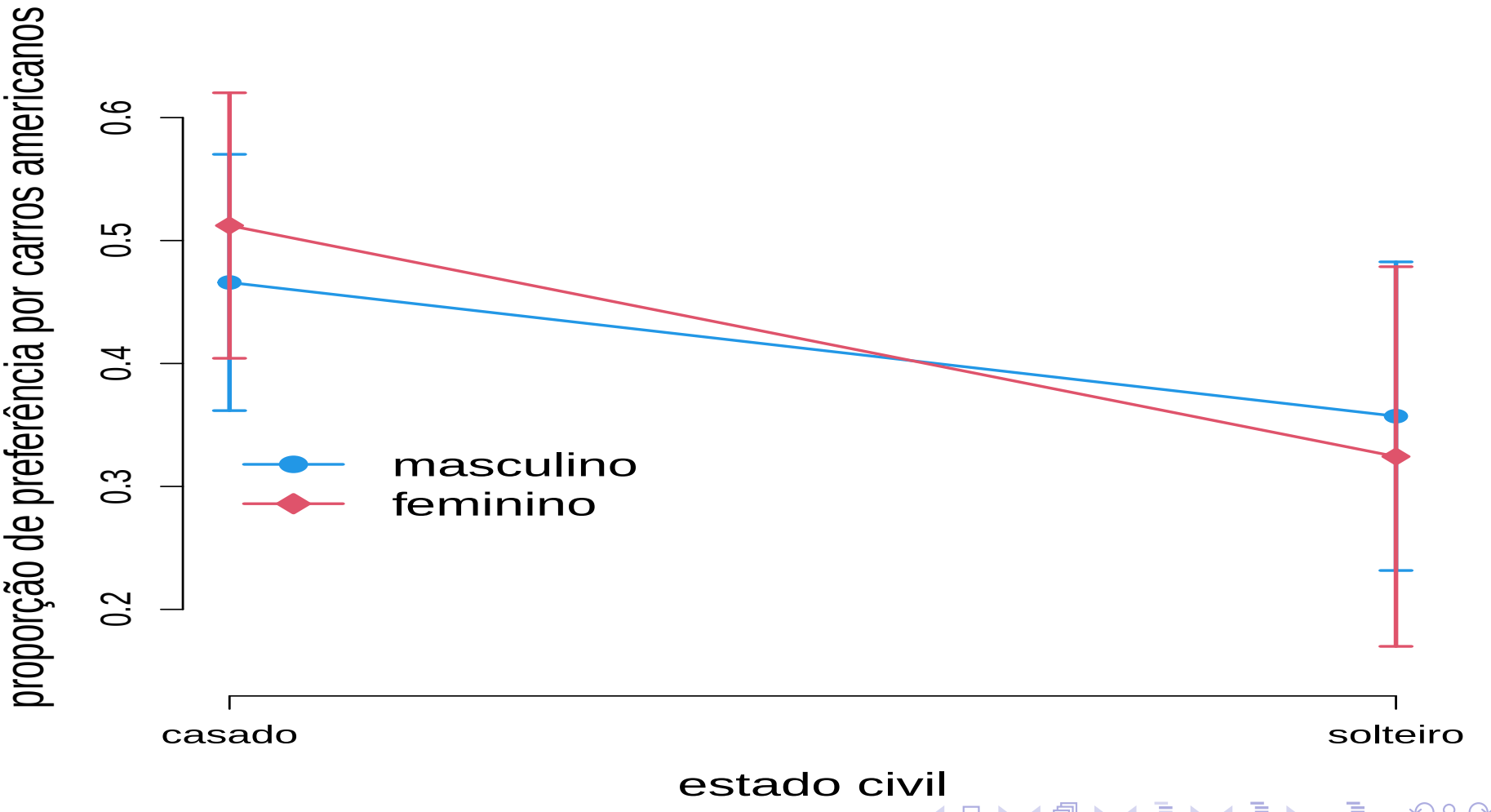
Gráficos de proporções por sexo



Gráficos de proporções por estado civil



Gráficos de perfis médios



Modelo logito

- Modelo

$$\begin{aligned} Y_{ijk} &\overset{ind.}{\sim} \text{Bernoulli}(\mu_{ij}) \\ \ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij}, i = 1,2, j = 1,2, k = 1,2,...,n_{ij} \\ \beta_1 = \gamma_1 &= (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = 0, \forall i, j. \end{aligned}$$

- n_{ij} : número total de consumidores pertecentes ao i-ésimo sexo (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 sexo e ao j-ésimo estado civil prefere carros americanos e 0, caso ele prefira carros japoneses.

Modelo

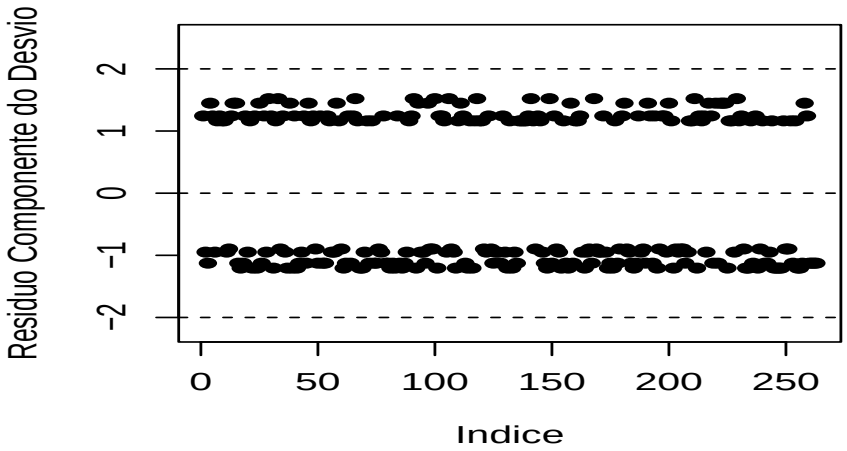
- $\boldsymbol{\beta} = (\alpha, \beta_2, \gamma_2, (\beta\gamma)_{22})'$.
- Logitos

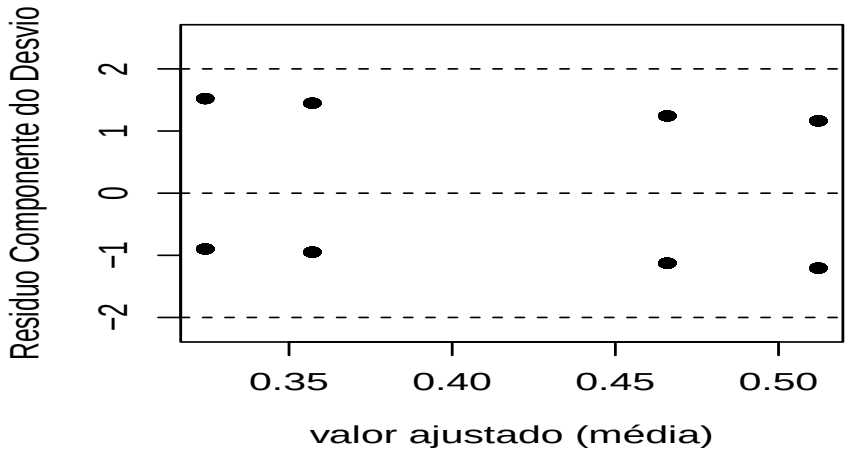
$$
\begin{aligned}
\ln\left(\frac{\mu_{11}}{1-\mu_{11}}\right) &= \alpha \Rightarrow \mu_{11} = \frac{e^{\alpha}}{1+e^{\alpha}}, \\
\ln\left(\frac{\mu_{21}}{1-\mu_{21}}\right) &= \alpha + \beta_2 \Rightarrow \mu_{21} = \frac{e^{\alpha+\beta_2}}{1+e^{\alpha+\beta_2}}, \\
\ln\left(\frac{\mu_{12}}{1-\mu_{12}}\right) &= \alpha + \gamma_2 \Rightarrow \mu_{12} = \frac{e^{\alpha+\gamma_2}}{1+e^{\alpha+\gamma_2}}, \\
\ln\left(\frac{\mu_{22}}{1-\mu_{22}}\right) &= \alpha + \beta_2 + \gamma_2 + (\beta\gamma)_{22}, \\
&\Rightarrow \mu_{22} = \frac{e^{\alpha+\beta_2+\gamma_2+(\alpha\beta)_{22}}}{1+e^{\alpha+\beta_2+\gamma_2+(\beta\gamma)_{22}}}
\end{aligned}
$$

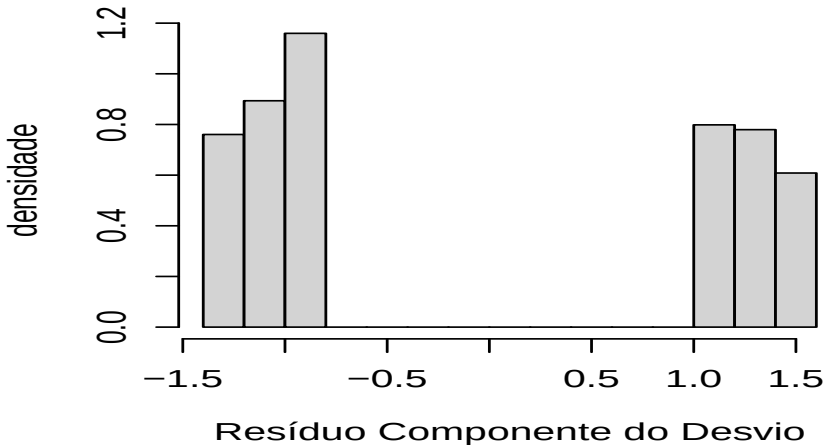
Modelo

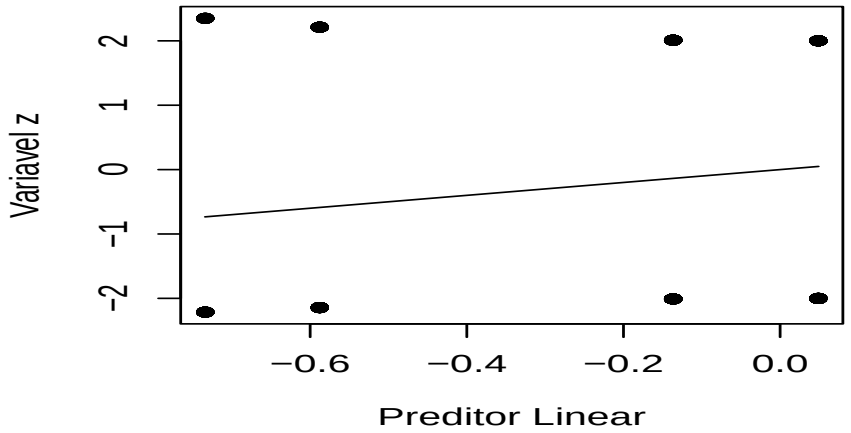
- Os parâmetros seguem as interpretações usuais, mas agora em termos das probabilidades e das razões de chances.
- **Exercício: provar que o parâmetro $(\beta\gamma)_{22}$ está relacionado com a presença de interação entre os fatores.**
- **Exercício: interprete os parâmetros $(\beta_2, \gamma_2)'$ em termos de razões de chances, dado a presença de interação.**
- **Exercício: provar que os parâmetros $(\beta_2, \gamma_2)'$ estão relacionados com a presença dos efeitos dos seus respectivos fatores, dado a ausência de interação.**

Gráficos de diagnóstico para o RCD: modelo logito

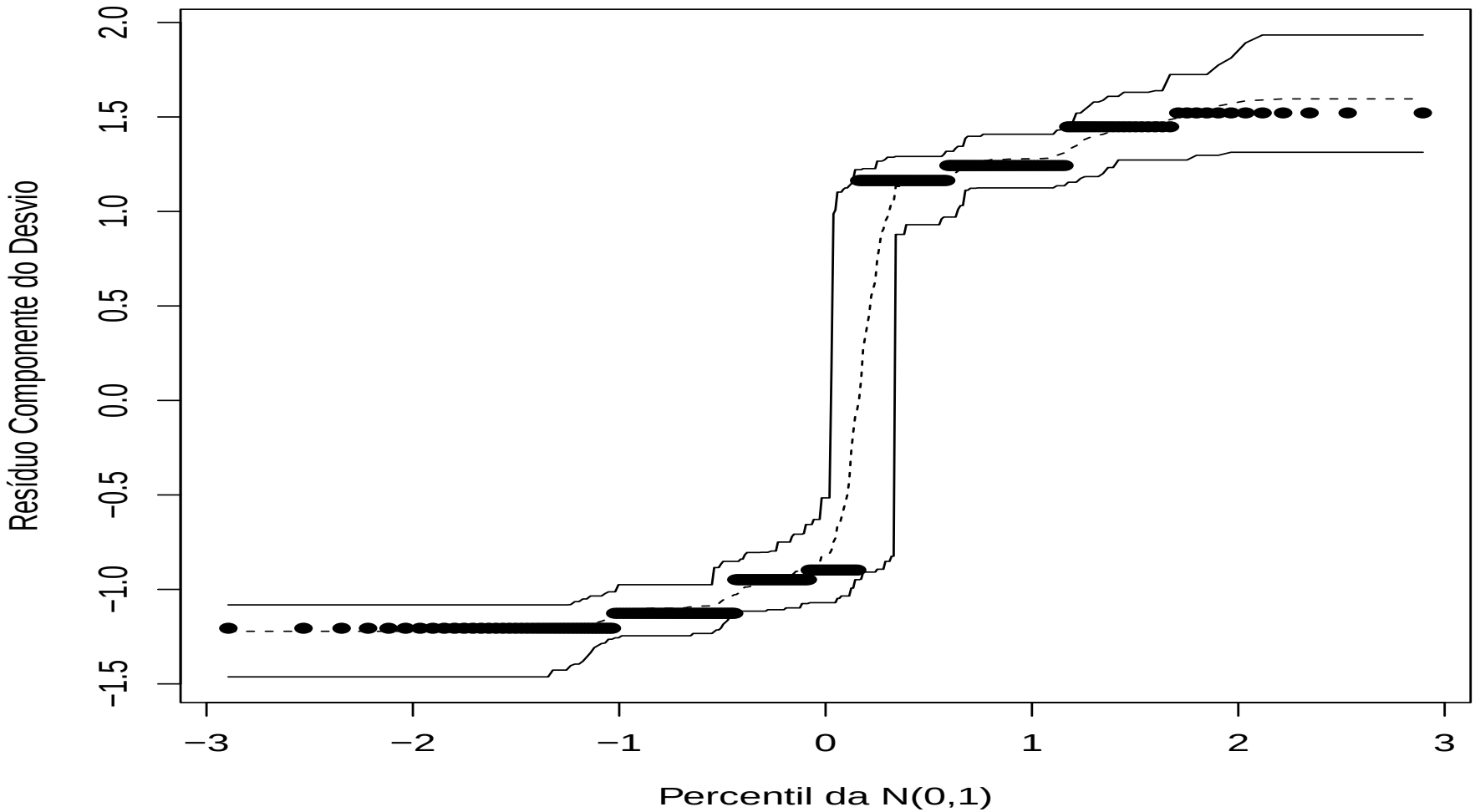




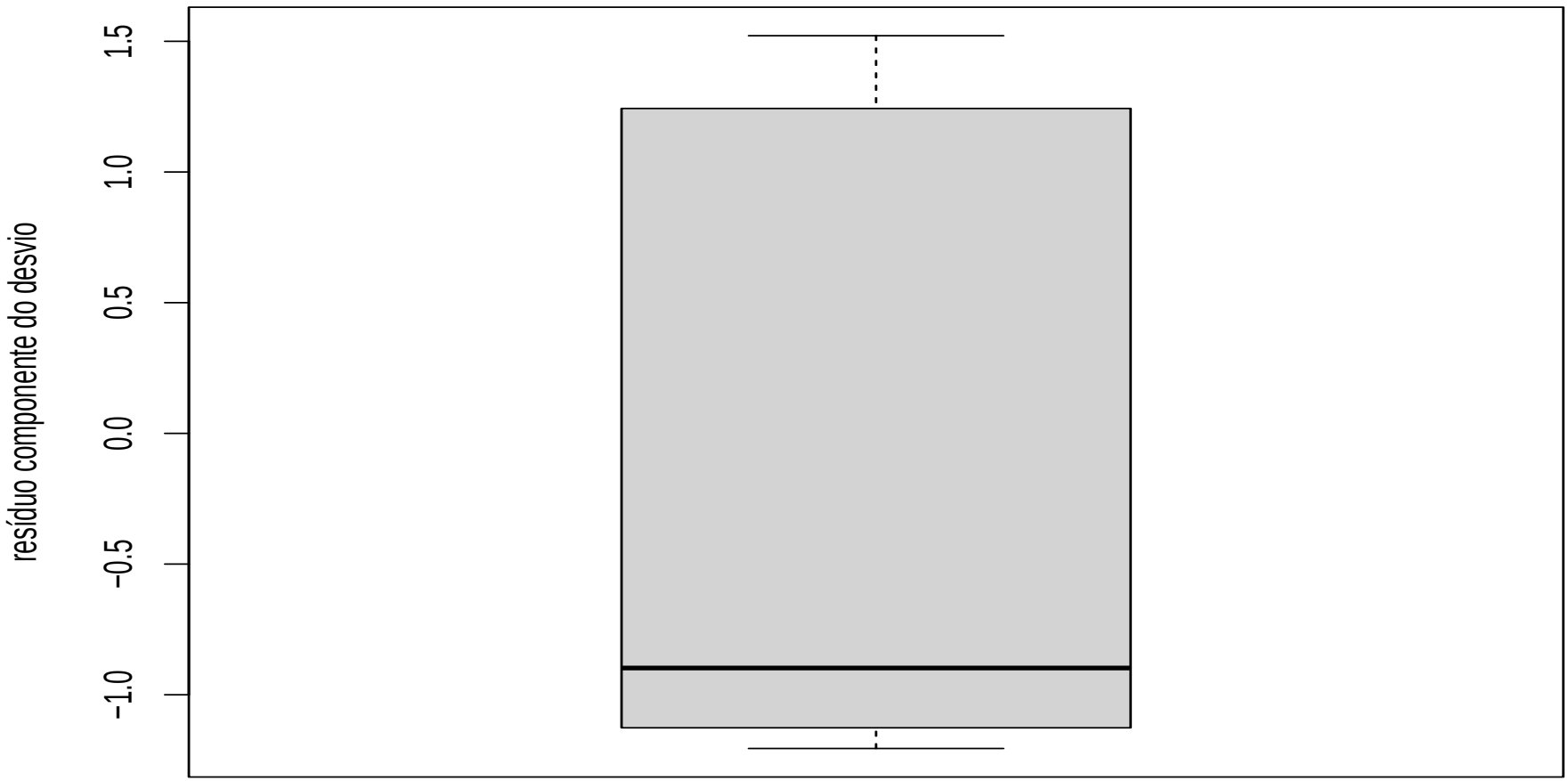




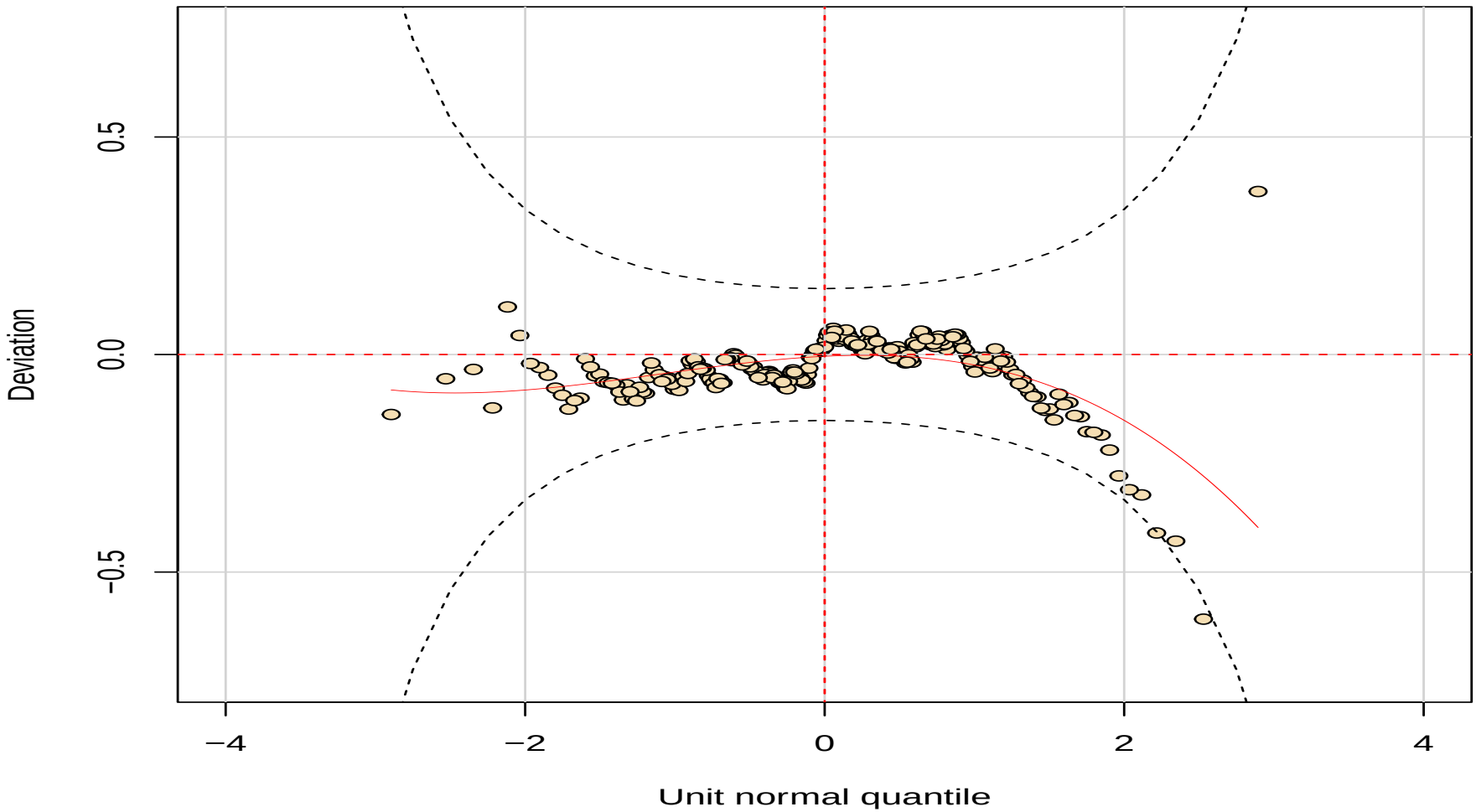
Gráficos de envelopes para o RCD: modelo logito



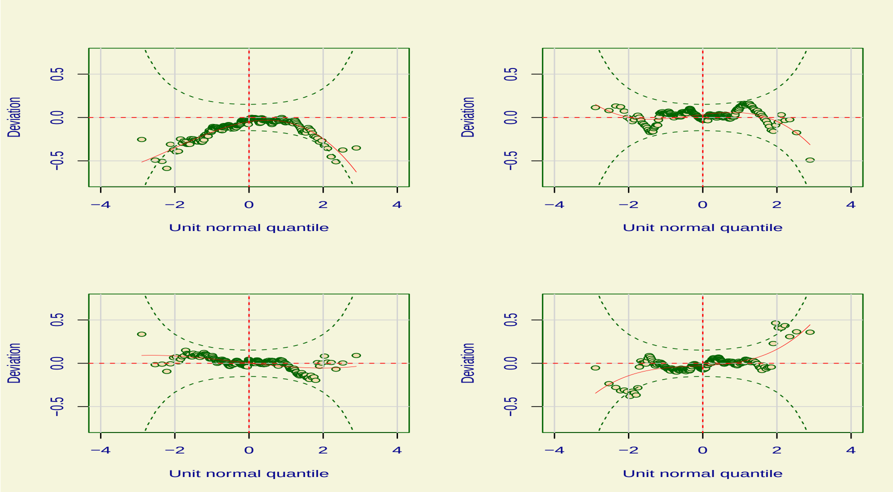
Box plot para o RCD: modelo logito



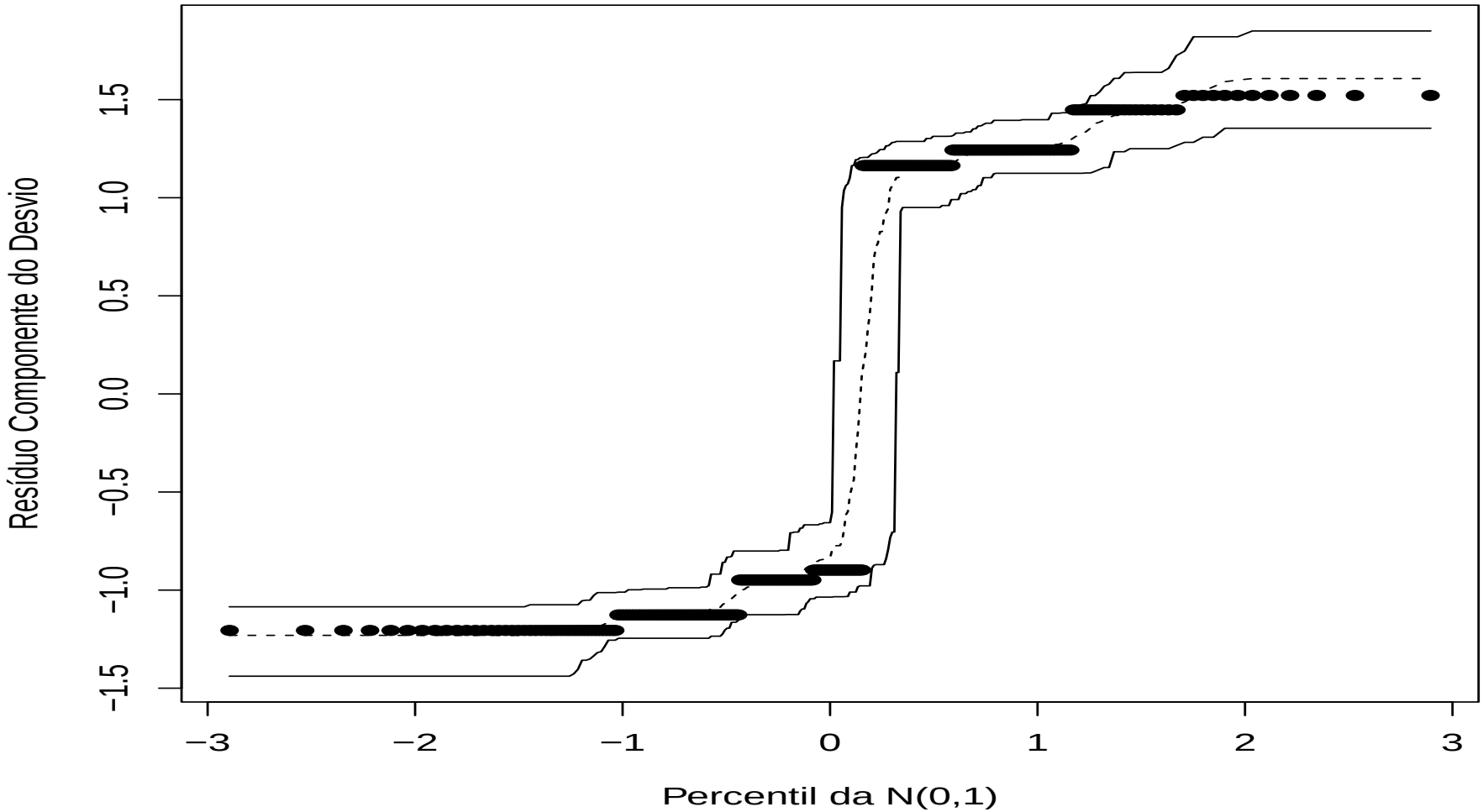
Worm plot para o modelo RQA: modelo logito



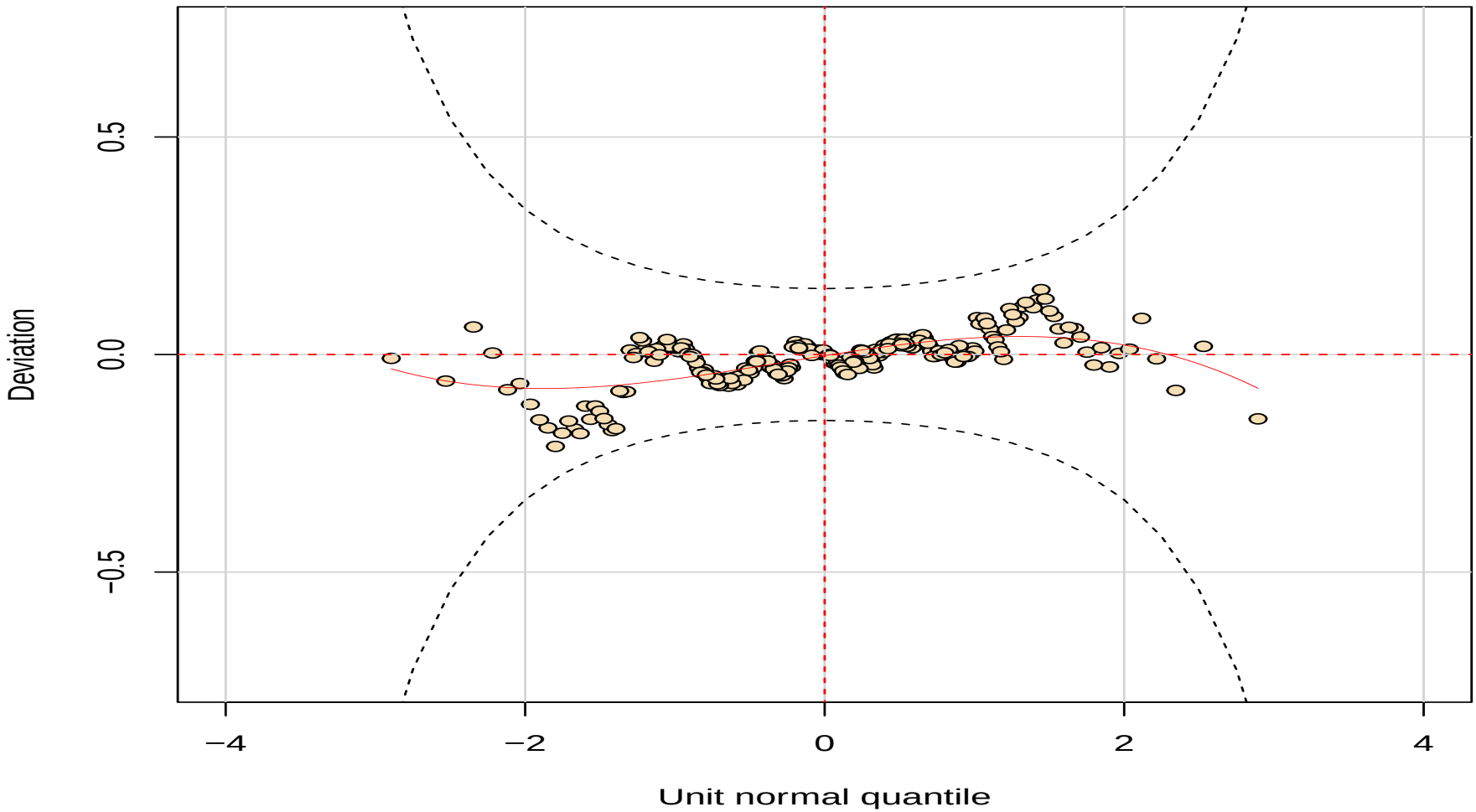
Worm plots para o modelo o RAQ: modelo logito



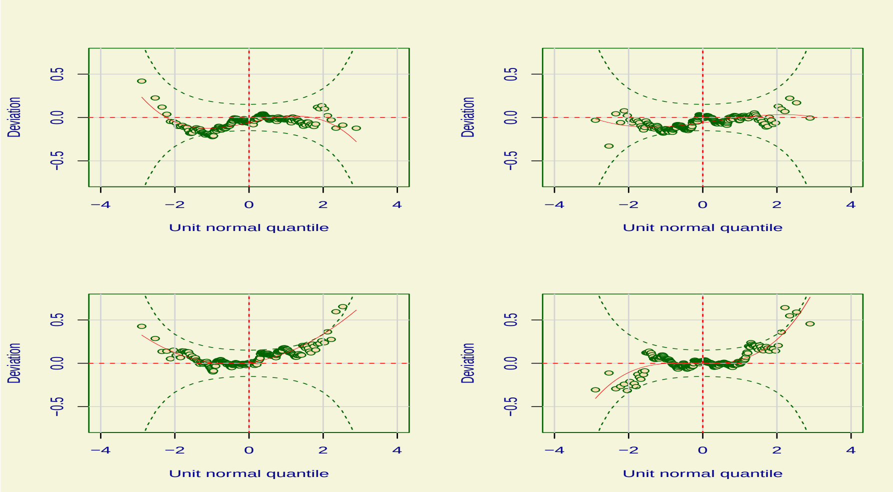
Gráficos de envelopes para o RCD: modelo probito



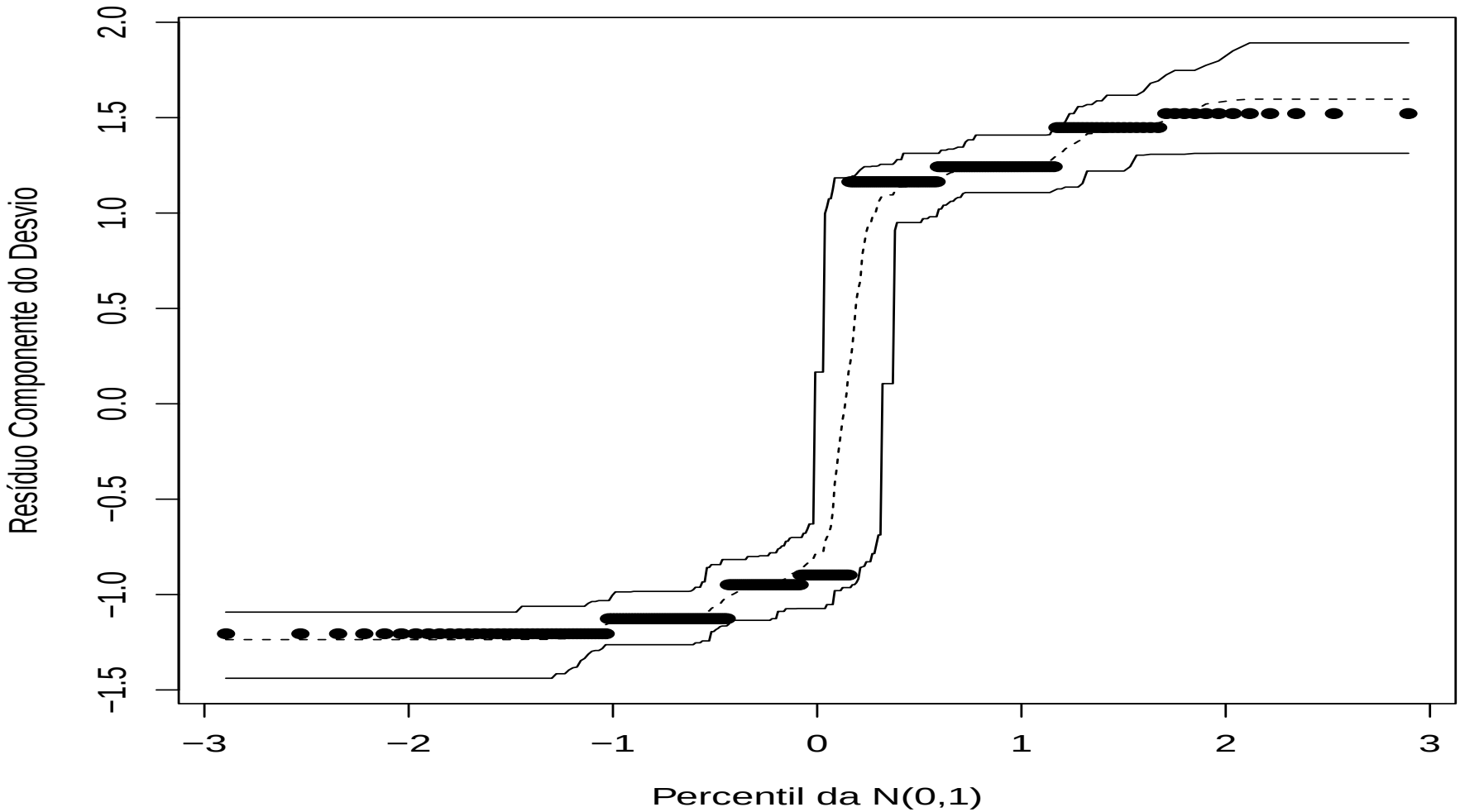
Worm plot para o modelo RQA: modelo probito



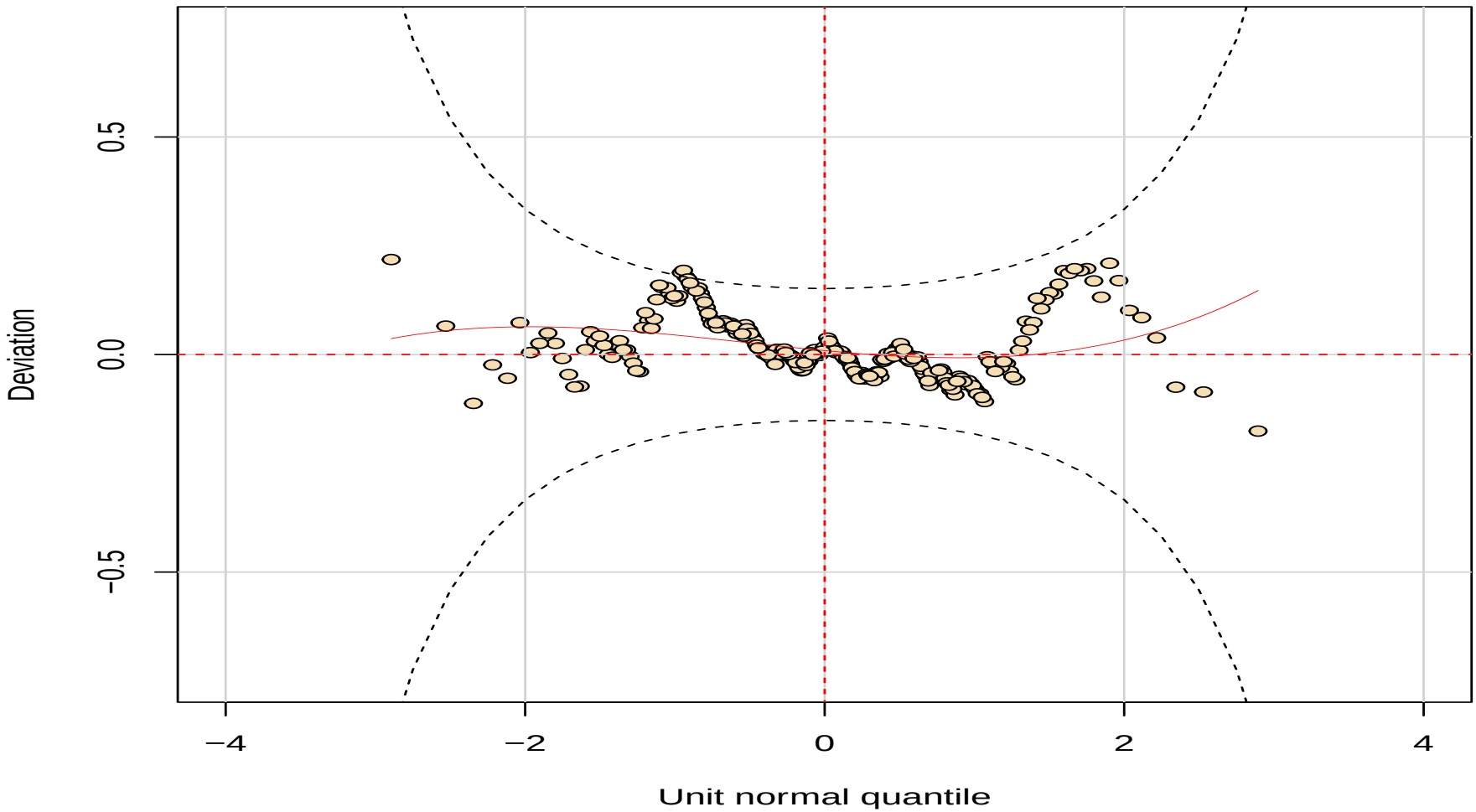
Worm plots para o modelo o RAQ: modelo probito



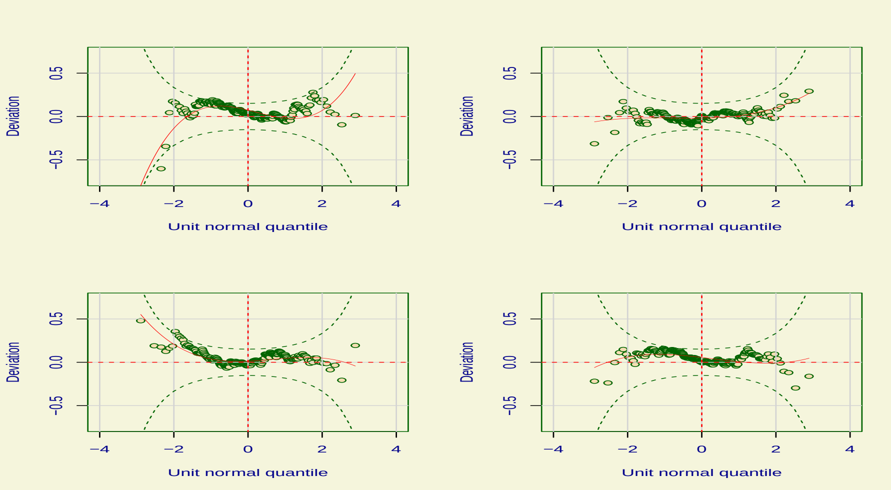
Gráficos de envelopes para o RCD: modelo cauchito



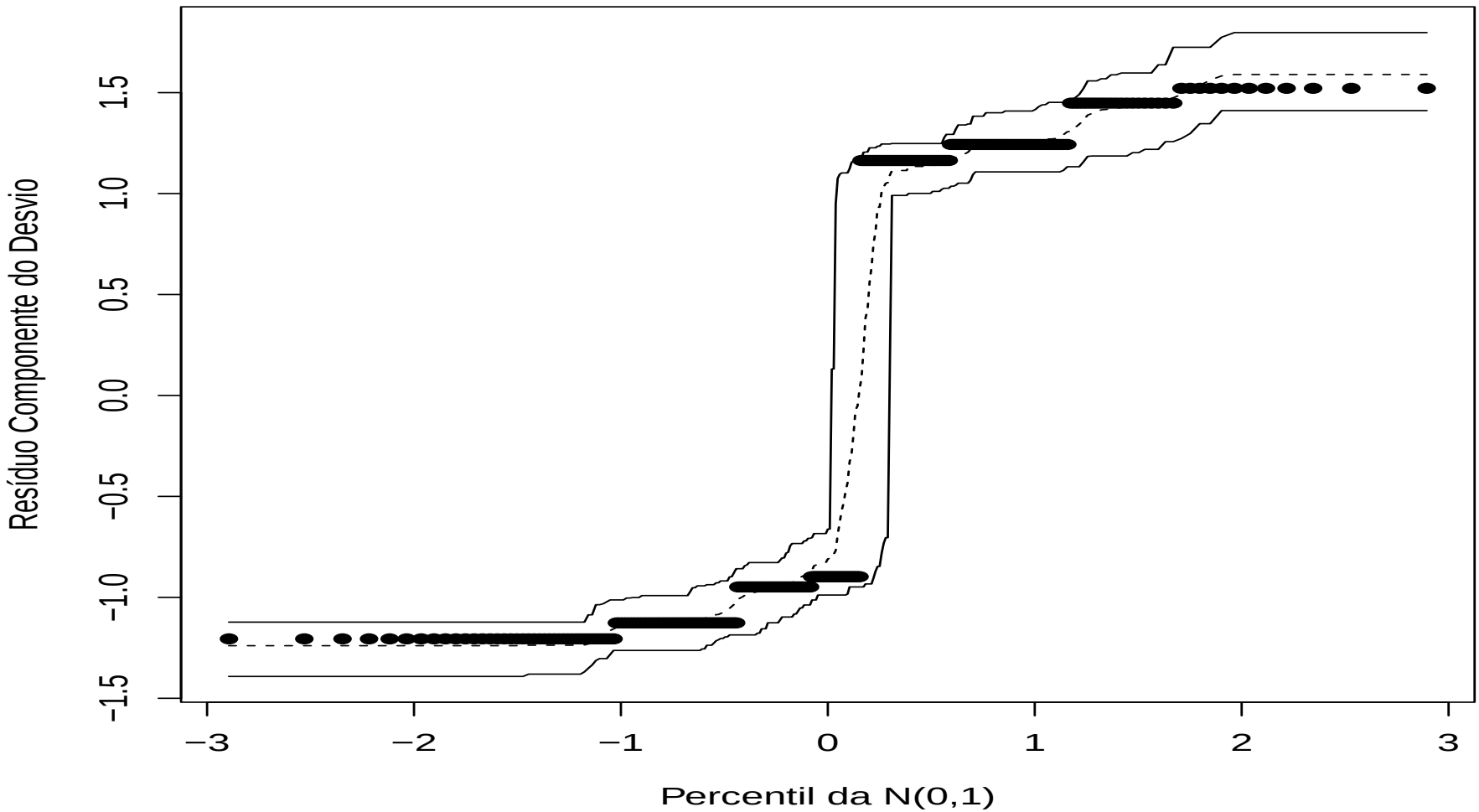
Worm plot para o modelo RQA: modelo cauchito



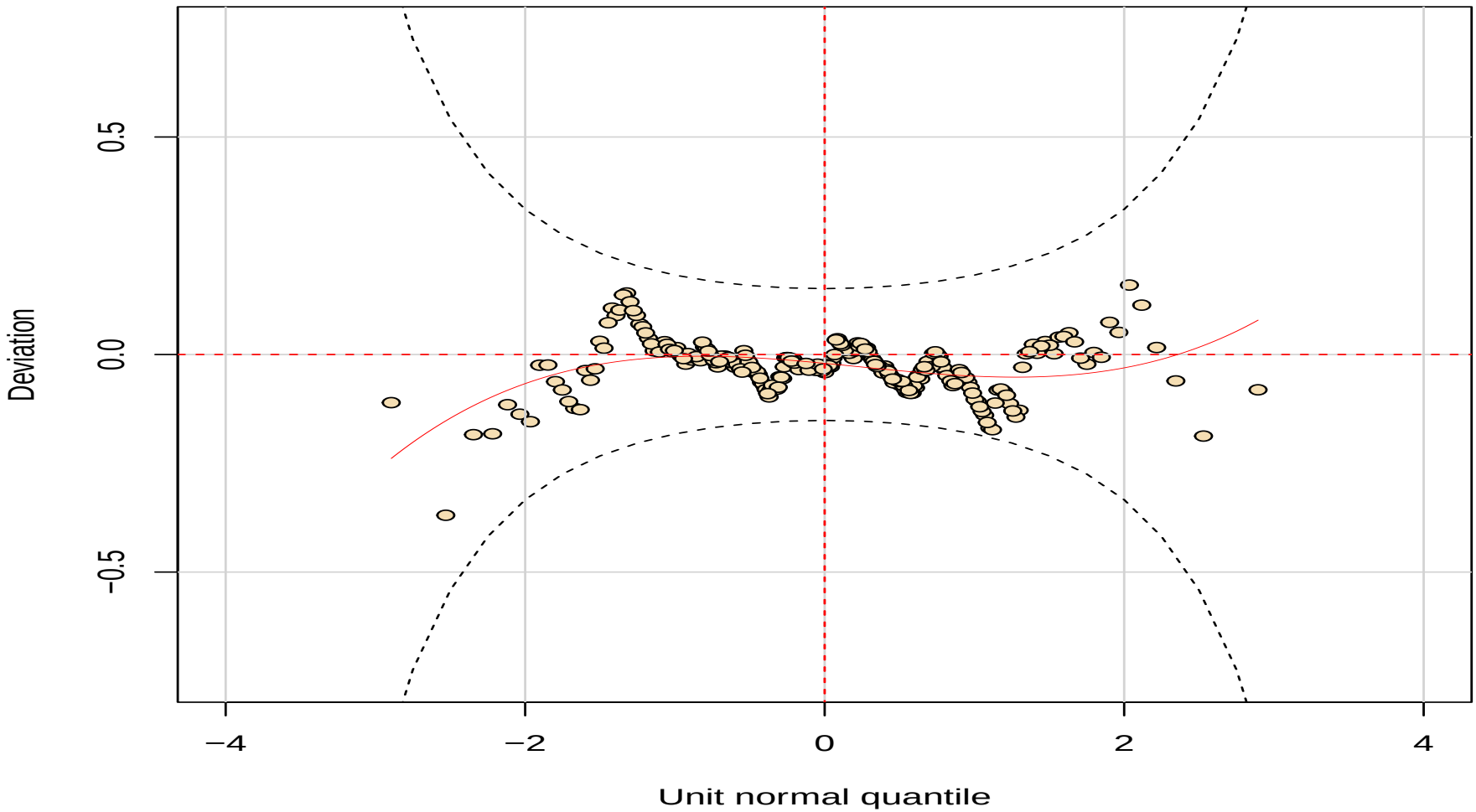
Worm plots para o modelo o RAQ: modelo cauchito



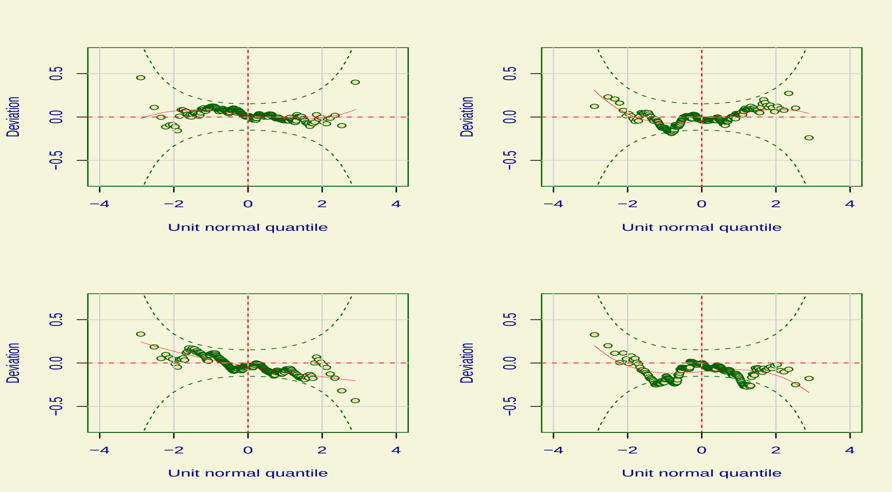
Gráficos de envelopes para o RCD: modelo cloglog



Worm plot para o modelo RQA: modelo cloglog



Worm plots para o modelo o RAQ: modelo cloglog



Comparação com outros modelos (funções de ligação)

Modelo	AIC	BIC	AICc	SABIC	HQCIC	CAIC	DABM
Logito	362,83	377,12	362,99	364,44	368,58	381,12	0,48
Probito	362,83	377,12	362,99	364,44	368,58	381,12	0,48
Cauchito	362,83	377,12	362,99	364,44	368,58	381,12	0,48
Cloglog	362,83	377,12	362,99	364,44	368,58	381,12	0,48

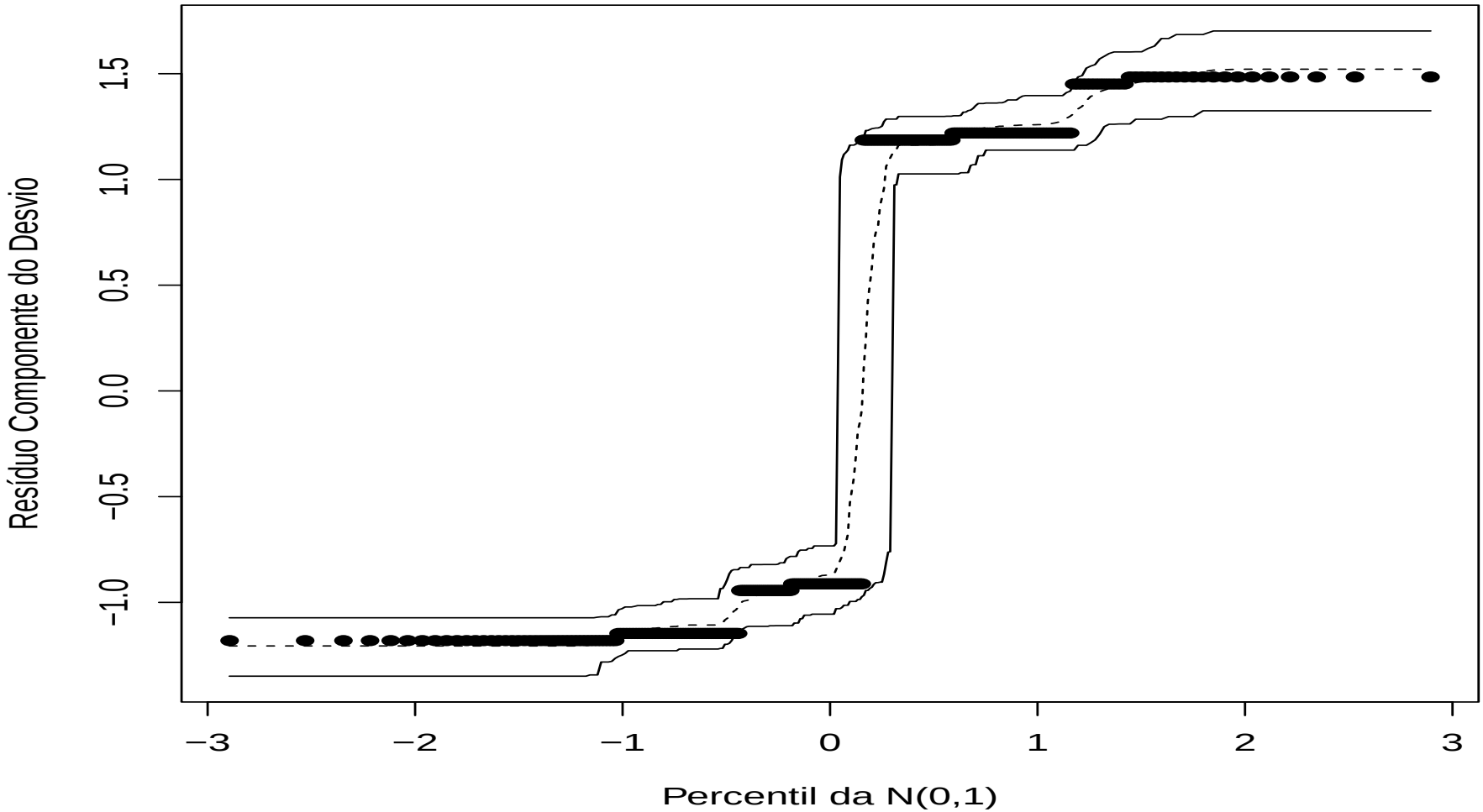
$DABM = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n_{ij}}|y_{ijk} - \widetilde{\mu}_{ij}|$. Lembrando que: $\mu_{ij} = \Phi(\eta_{ij})$ (probito), $\mu_{ij} = \frac{1}{\pi}\arctan(\eta_{ij}) + \frac{1}{2}$ (cauchito) e $\mu_{ij} = 1 - e^{-e^{\eta_{ij}}}$ (cloglog) e $\Phi(.)$ é a fda da norma padrão e $n = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$. Vamos utilizar o modelo logito.

Ajuste do modelo completo

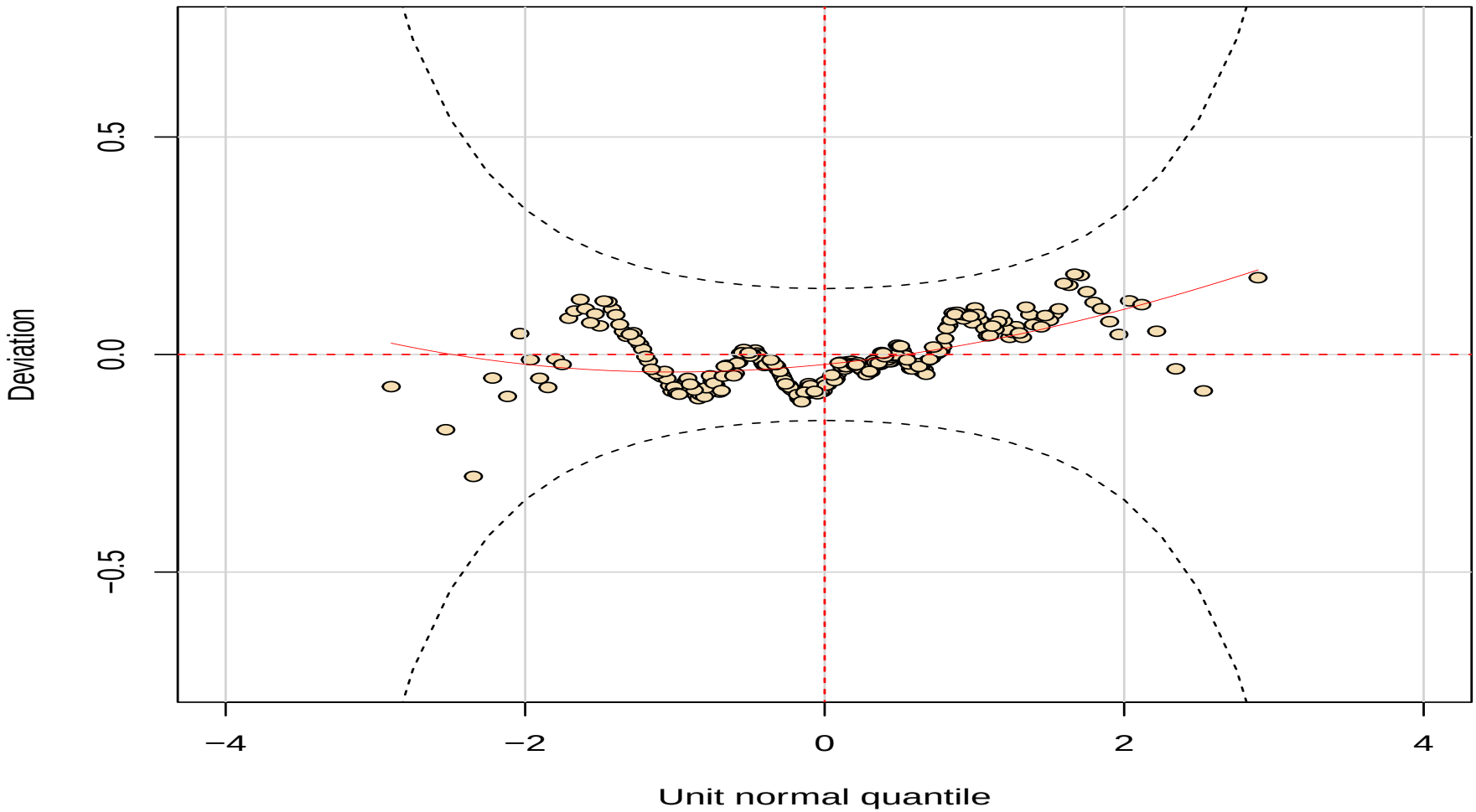
Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
μ	-0,14	0,21	[-0,56 ; 0,28]	-0,64	0,5228
β_2	0,19	0,31	[-0,42 ; 0,79]	0,60	0,5465
γ_2	-0,45	0,35	[-1,14 ; 0,24]	-1,28	0,1991
$(\beta\gamma)_{22}$	-0,33	0,54	[-1,40 ; 0,73]	-0,61	0,5420

Aparentemente, nenhum coeficiente é significativo. Entretanto, vamos explorar o modelo um pouco melhor, ou seja, ajustando um modelo sem interação.

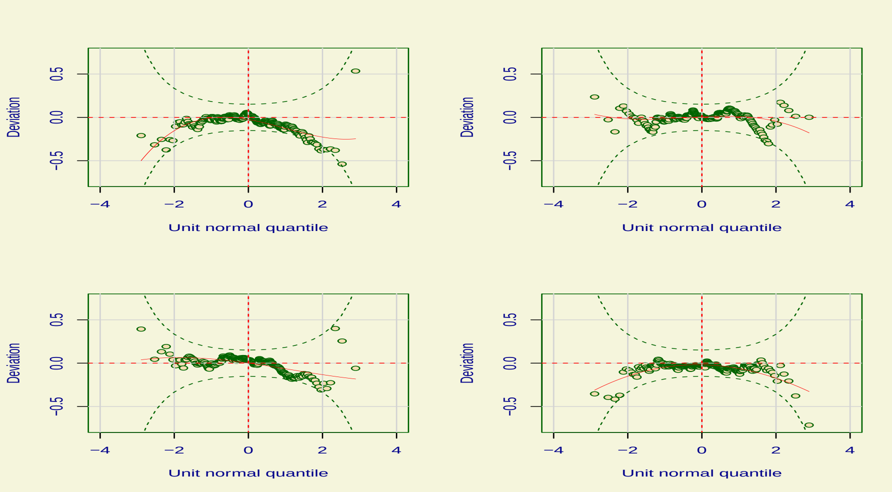
Gráficos de envelopes para o modelo sem interação



Worm plot para o modelo sem interação



Worm plots para o modelo sem interação

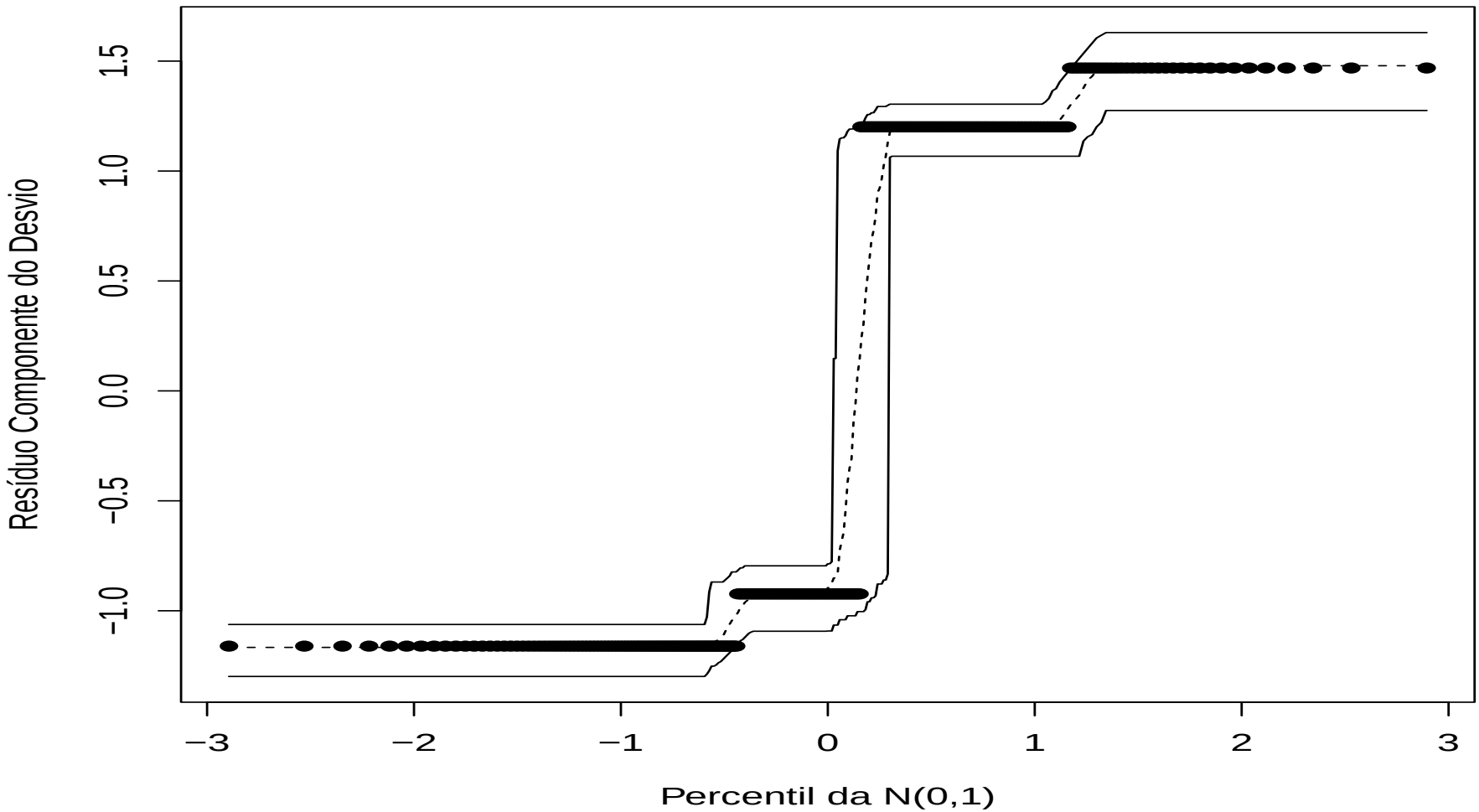


Ajuste do modelo sem interação

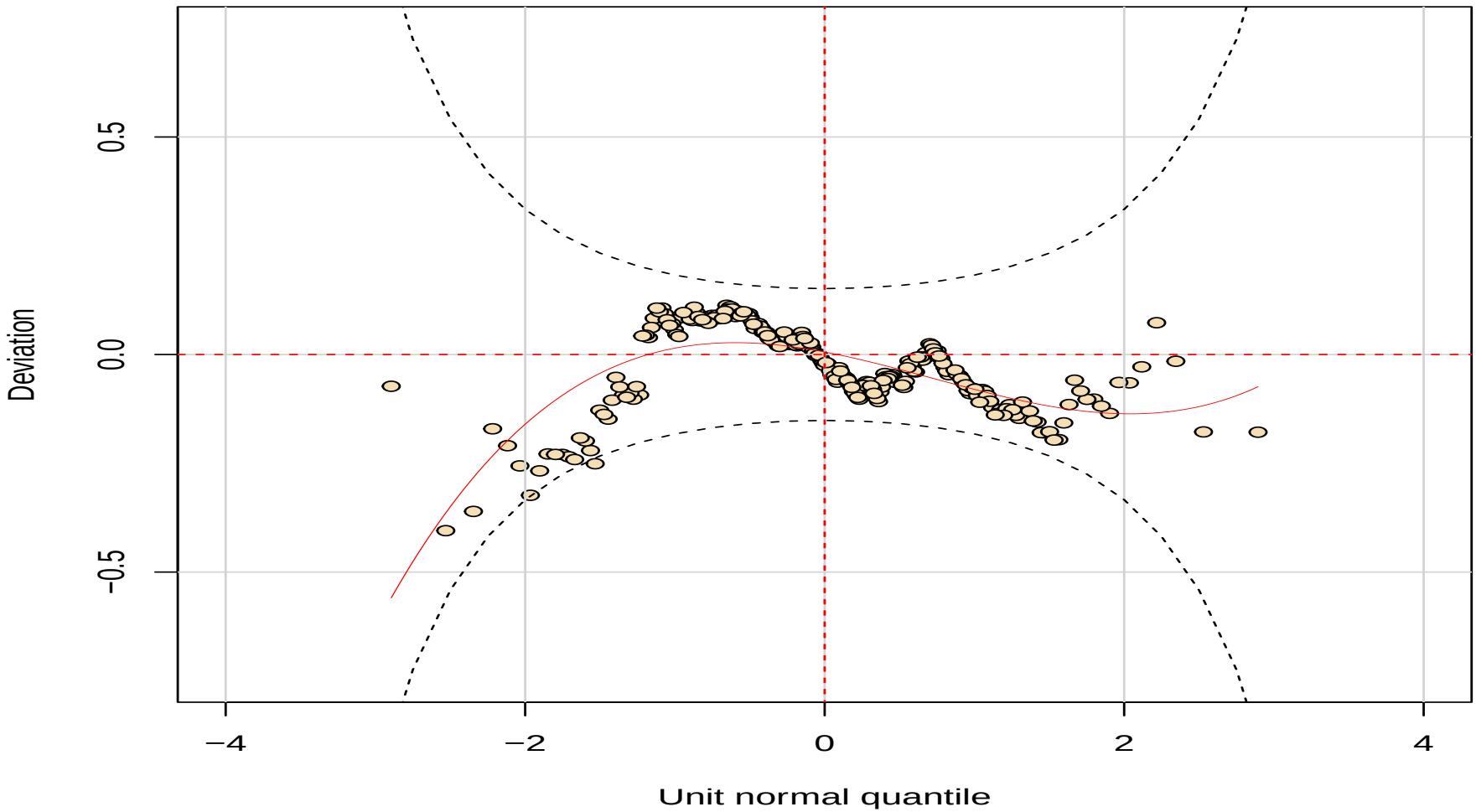
Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
μ	-0,09	0,20	[-0,47 ; 0,30]	-0,43	0,6642
β_2	0,08	0,25	[-0,42 ; 0,57]	0,31	0,7551
γ_2	-0,59	0,27	[-1,12 ; -0,07]	-2,21	0,0270

O fator sexo parece ser não significativo enquanto que o fator estado civil parece ser significativo.

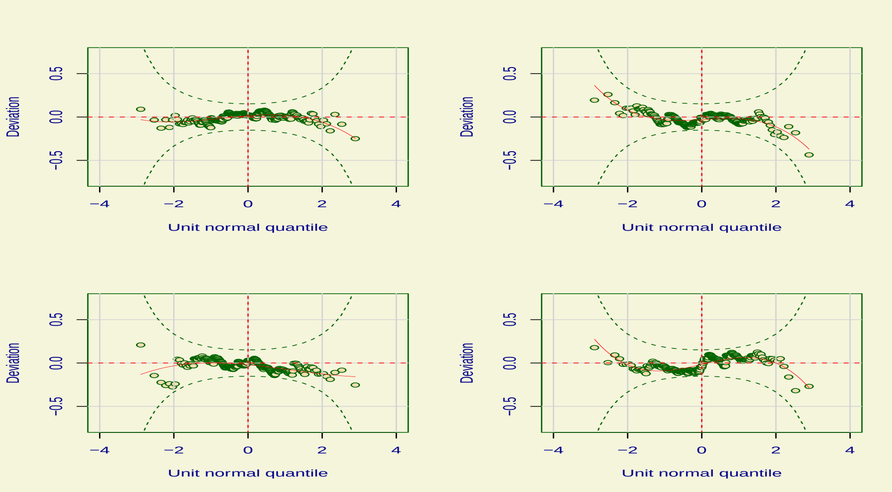
Gráficos de envelopes para o mod. com somente o fator EC



Worm plot para o modelo sem interação



Worm plots para o modelo sem interação



Ajuste do modelo com somente o fator estado civil

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
μ	-0,05	0,15	[-0,35 ; 0,25]	-0,31	0,7590
γ_2	-0,60	0,27	[-1,12 ; -0,08]	-2,24	0,0250

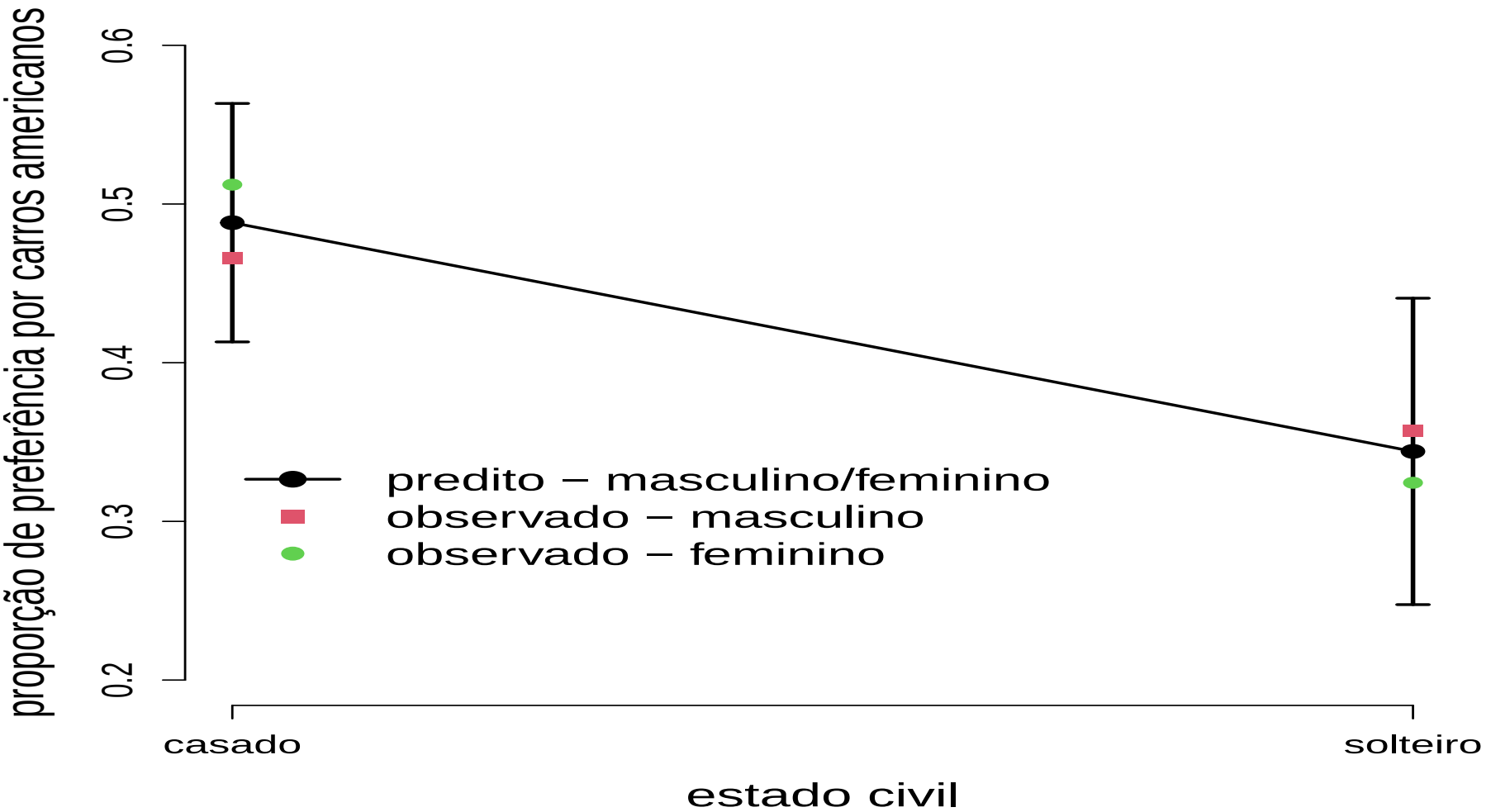
Modelo final: fator estado civil parece ser significativo.

Percentuais preditos pelo modelo final (através do método delta)

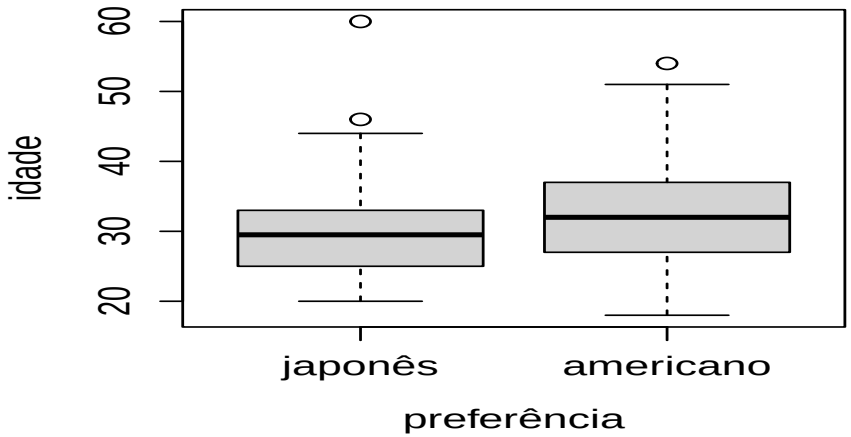
Estado civil	sexo	Estimativa	EP	IC(95%)
Casado	Masculino	48,82	3,83	[41,31 ;56,34]
Solteiro	Masculino	34,41	4,93	[24,75 ; 44,06]
Casado	Feminino	48,82	3,83	[41,31 ; 56,34]
Solteiro	Feminino	34,41	4,93	[24,75 ; 44,06]

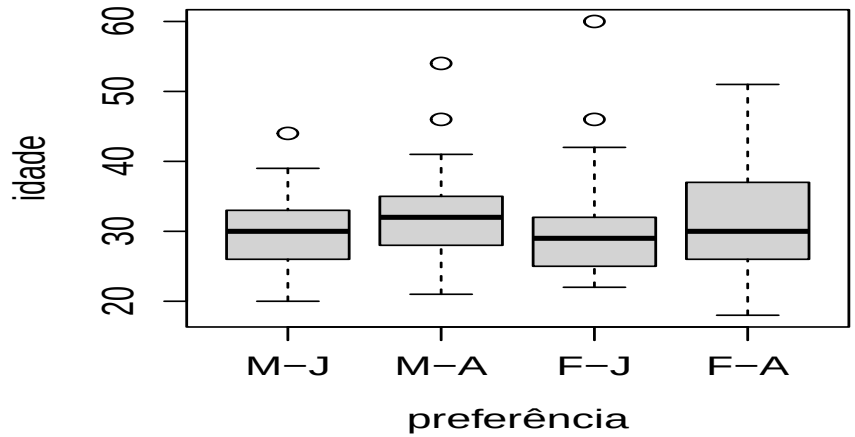
Exercício: obter os resultados acima aplicando o método delta.

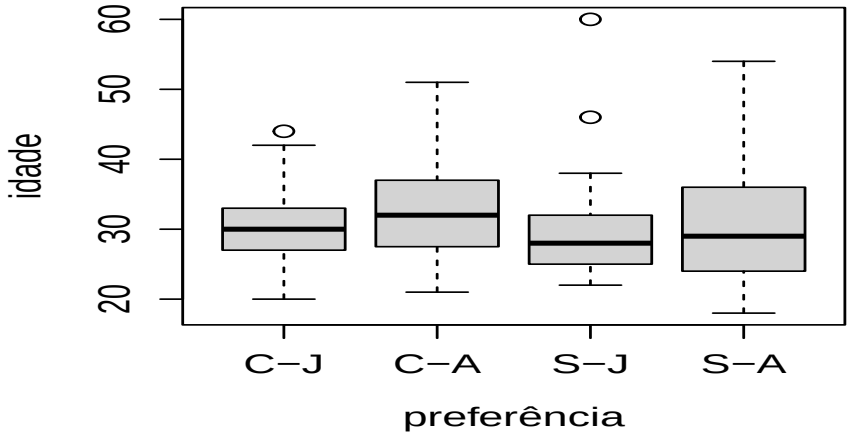
Proporções observadas e preditas pelo modelo final

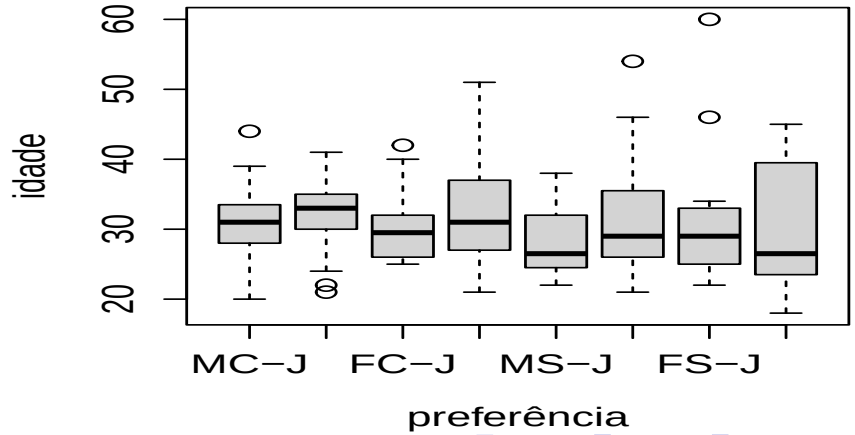


(Levando em consideração a idade) Box-plot da idade

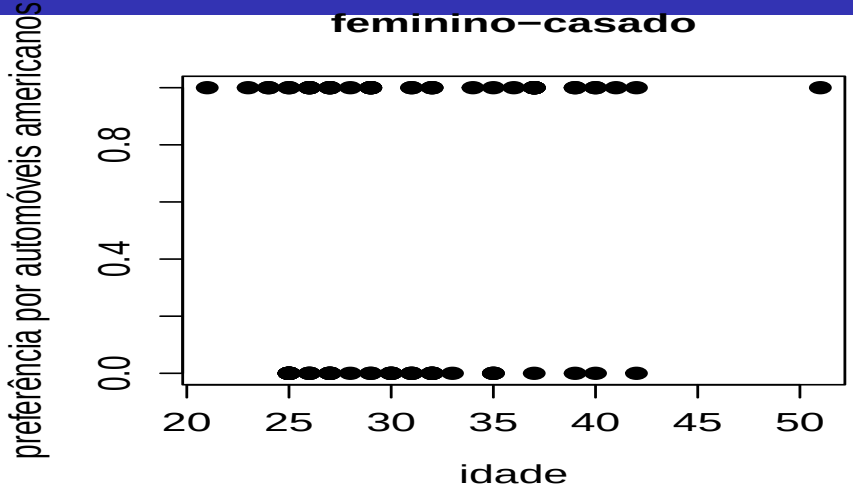


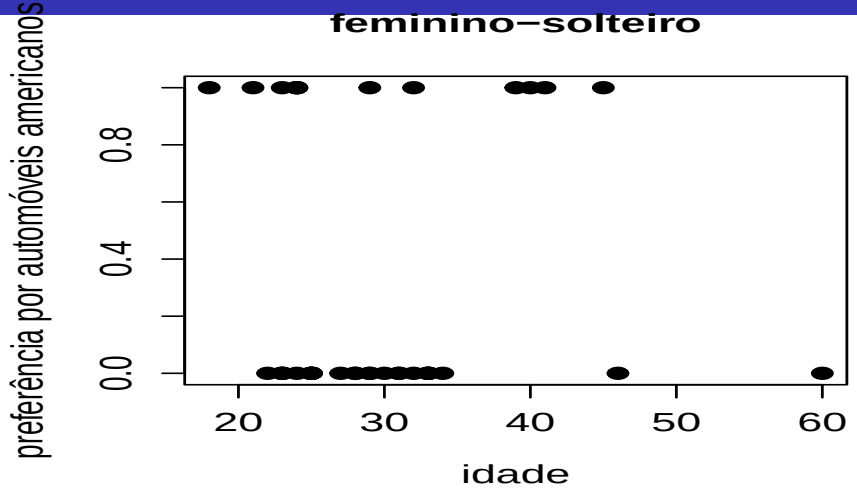


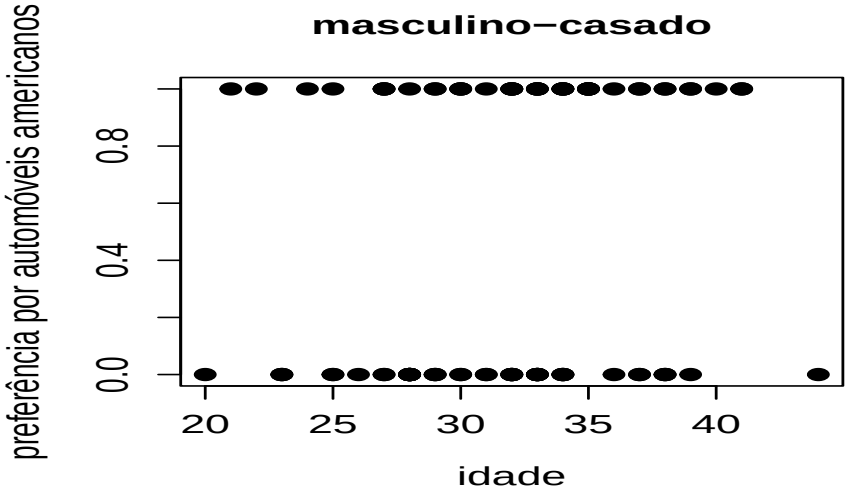


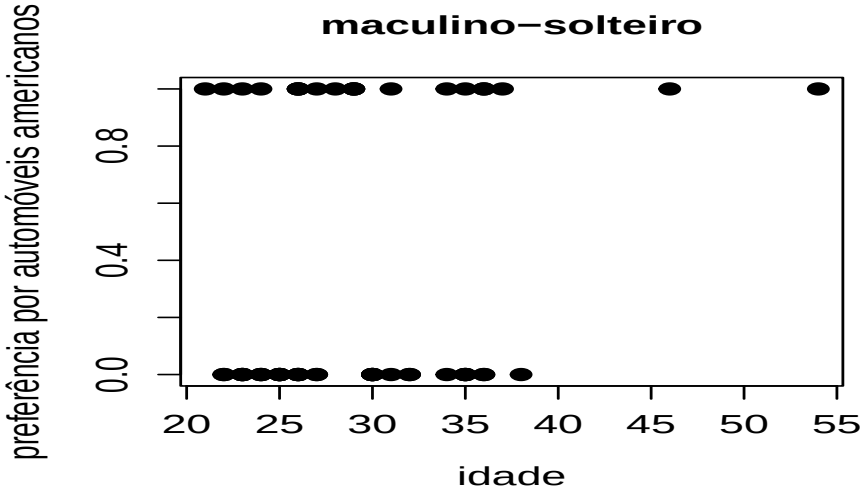


Dispersão: valores observados x idade

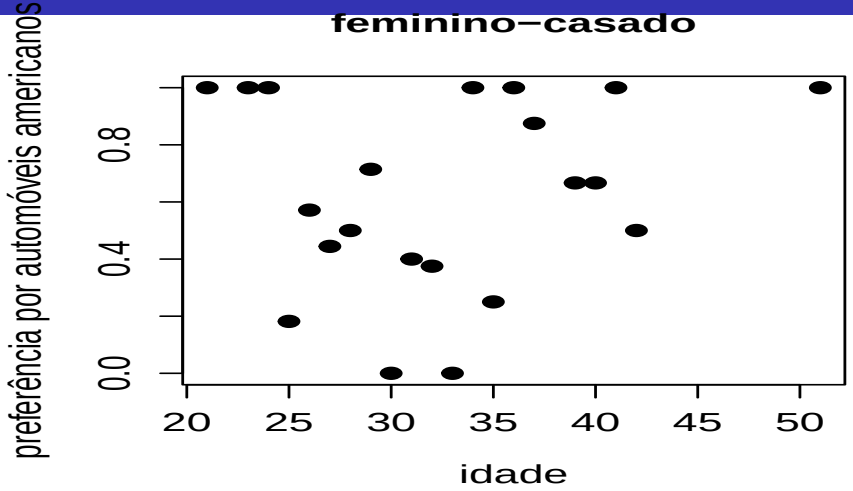


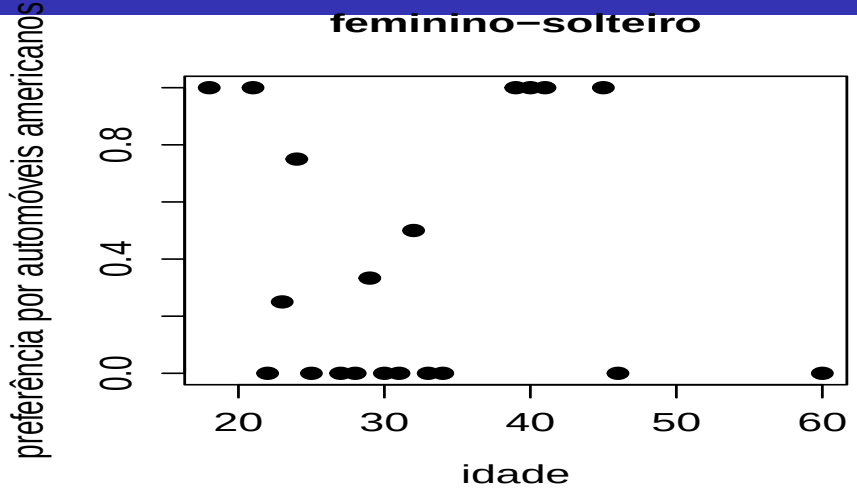


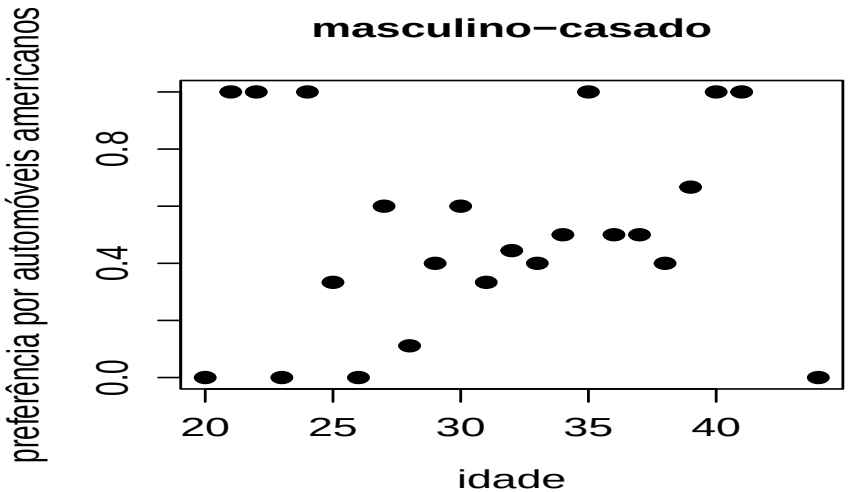


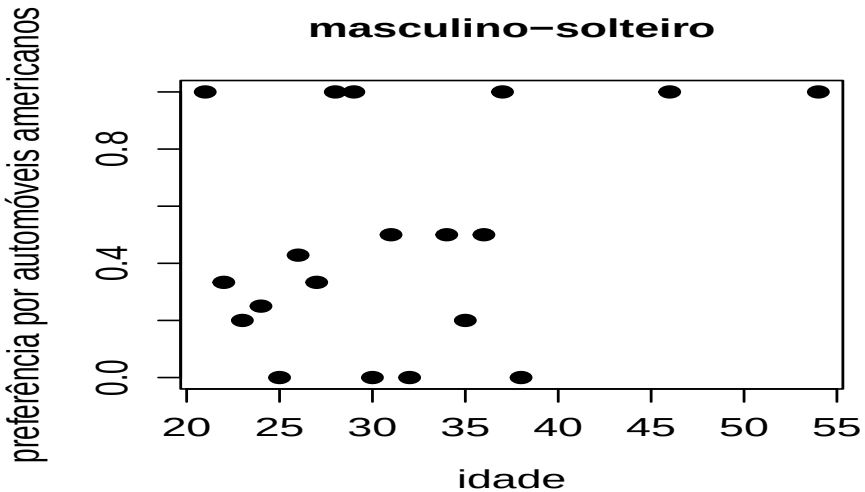


Dispersão: proporções observadas x idade









Utilização da variável idade

- Modelo 1:

$$\begin{aligned} Y_{ijk} &\overset{ind.}{\sim} \text{Bernoulli}(\mu_{ijk}), \\ \ln\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \delta(x_{ijk} - \overline{x}), \\ i &= 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij}, \\ \beta_1 &= \gamma_1 = (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = 0, \forall i, j, \end{aligned}$$

em que

- x_{ijk} é a idade do k-ésimo indivíduo do sexo i e do estado civil j,
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n_{ij}} x_{ijk}$, $n = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$. Para $x_{ijk} = \overline{x}$ e/ou para indivíduos com a mesma idade, os parâmetros $(\alpha, \beta_2, \gamma_2, (\beta\gamma)_{22})'$

Utilização da variável idade

- Para $x_{ijk} = \overline{x}$ e/ou para indivíduos com a mesma idade, os parâmetros $(\alpha, \beta_2, \gamma_2, (\beta\gamma)_{22})'$ possuem a mesma interpretação anterior.
- Por outro lado, δ é o incremento no logito para o aumento em uma unidade da variável idade.

Utilização da variável idade

- Modelo 2 (interação considerada para a idade):

$$
\begin{aligned}
\ln\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
&+ (\delta + \theta_i + \lambda_j + (\theta\lambda)_{ij})(x_{ijk} - \overline{x}), \\
&\quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\
\beta_1 = \gamma_1 &= (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = \theta_1 = \lambda_1 = (\theta\lambda)_{1j} \\
&= (\theta\lambda)_{i1} = 0, \forall i, j.
\end{aligned}
$$

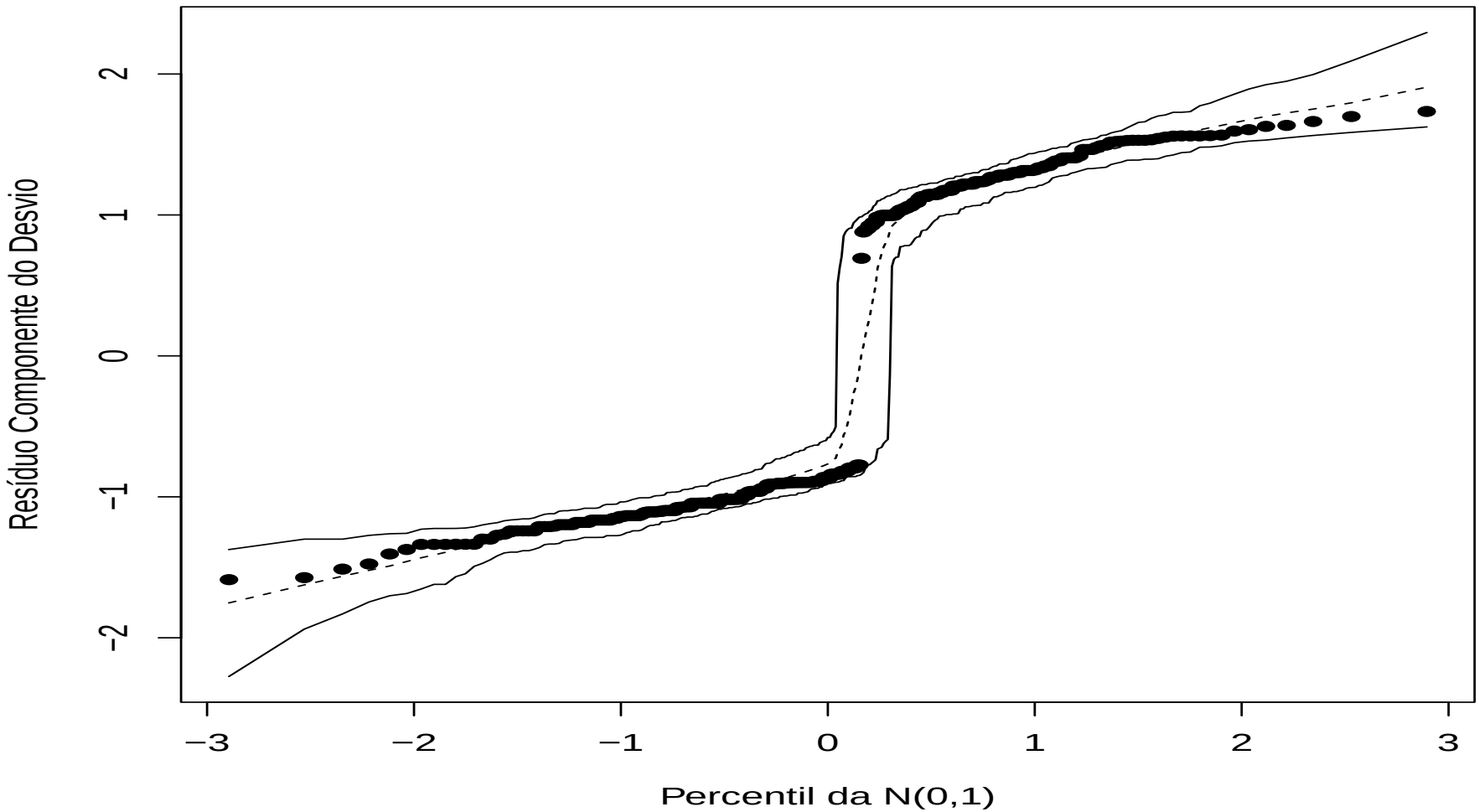
Utilização da variável idade

- (Cont.) Para $x_{ijk} = \overline{x}$ e/ou para indivíduos com a mesma idade e pertencentes ao mesmo grupo, os parâmetros $(\alpha, \beta_2, \gamma_2, (\beta\gamma)_{22})'$ já não mais possuem a mesma interpretação anterior pois o parâmetro $\delta_{ij} = \delta + \theta_i + \lambda_j + (\theta\lambda)_{ij}, i, j = 1, 2$ que continua sendo o incremento no logito para o aumento em uma unidade da variável idade, agora para cada grupo, também influenciará a diferença entre as médias dos grupos.

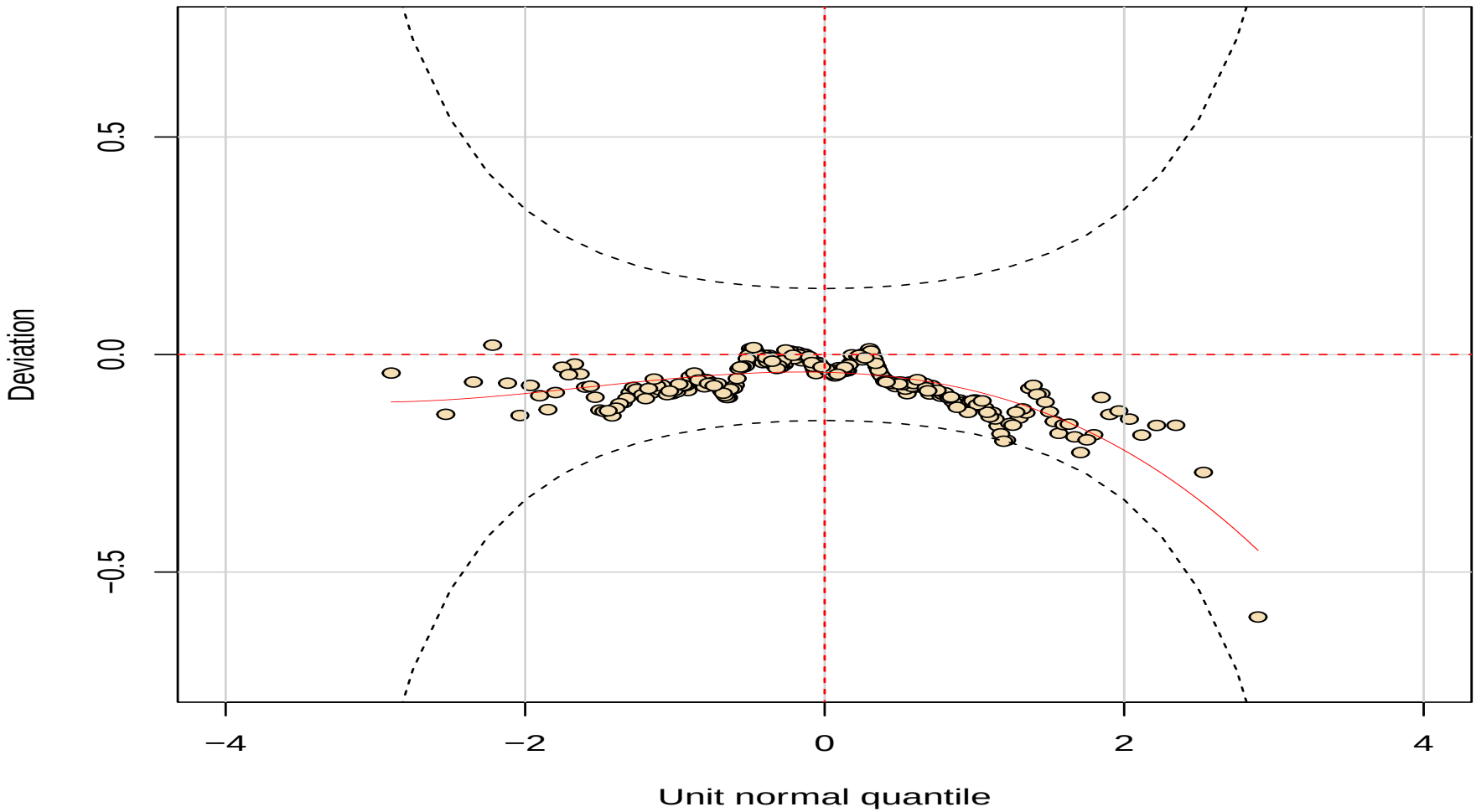
Comparação entre os modelos 1 e 2

- Os gráficos de envelopes para os RCD's indicam que ambos os modelos se ajustam bem (embora os RQA's indiquem que o ajuste não foi razoavelmente satisfatório), veja slides a seguir.
- Análise do desvio para testar $H_0 : \theta_2 = \lambda_2 = (\theta\lambda)_{22} = 0$ vs H_1 : há pelo menos uma diferença. Resultados $f_c(\text{p-valor}) = 0,31(0,9071)$.
- Adicionalmente, através de testes individuais de nulidade conseguimos reduzir o modelo 1, chegando ao modelo 3.

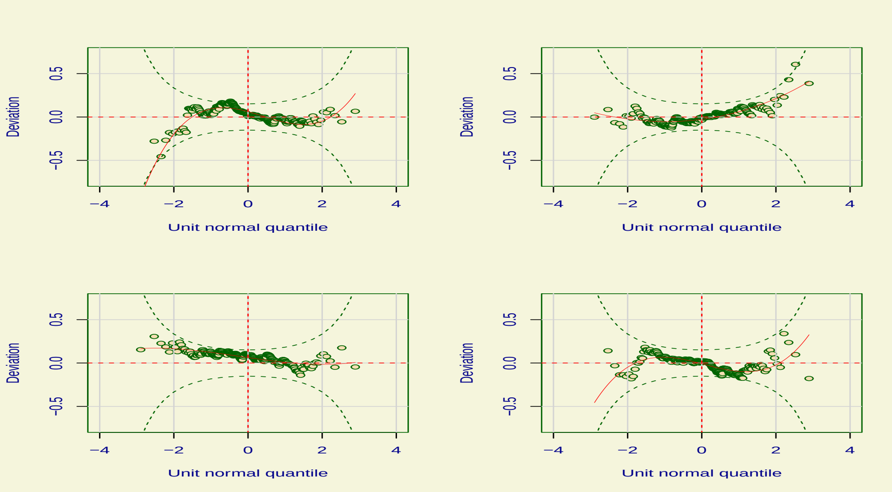
Gráficos de envelopes para o modelo 1



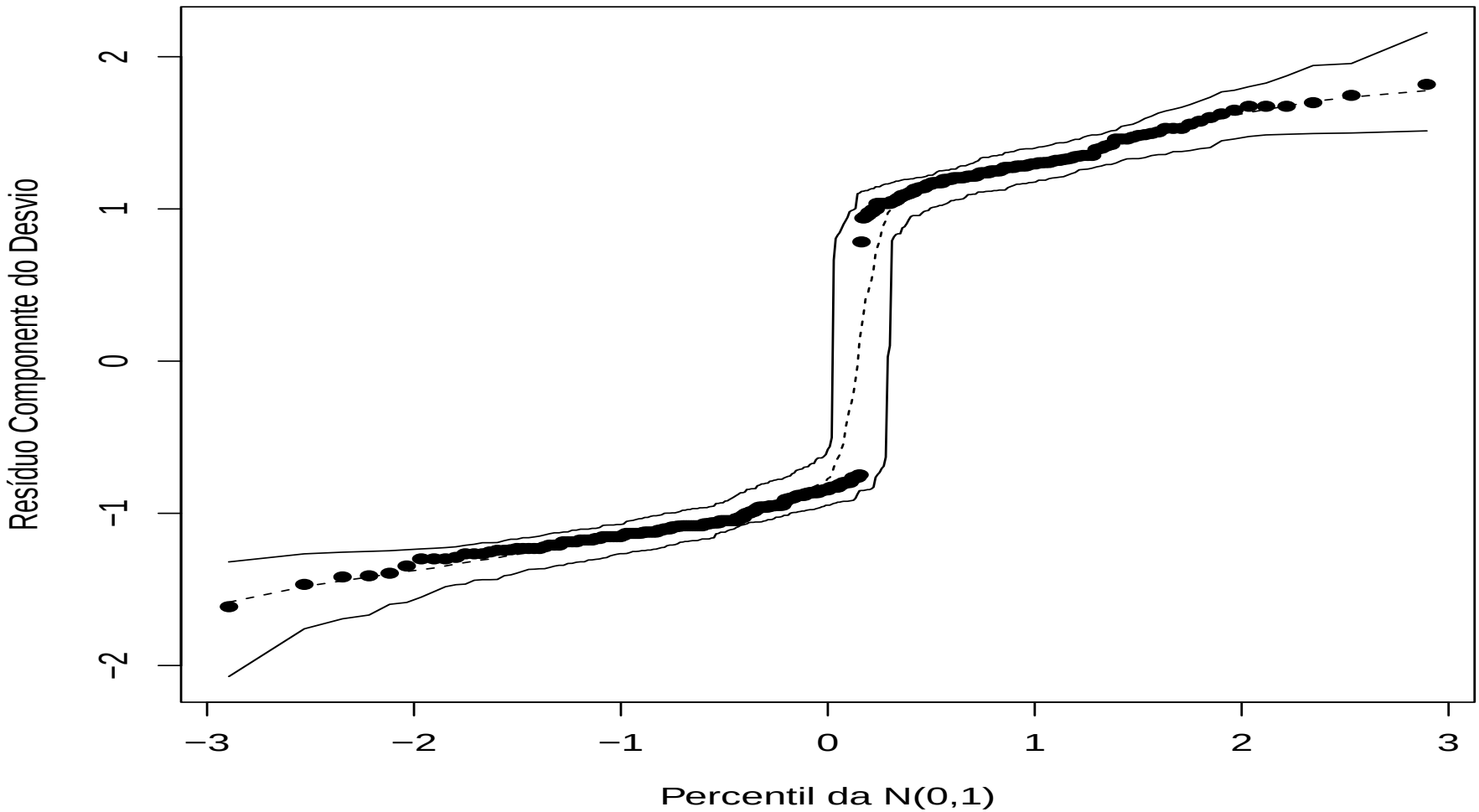
Worm plot para o modelo 1



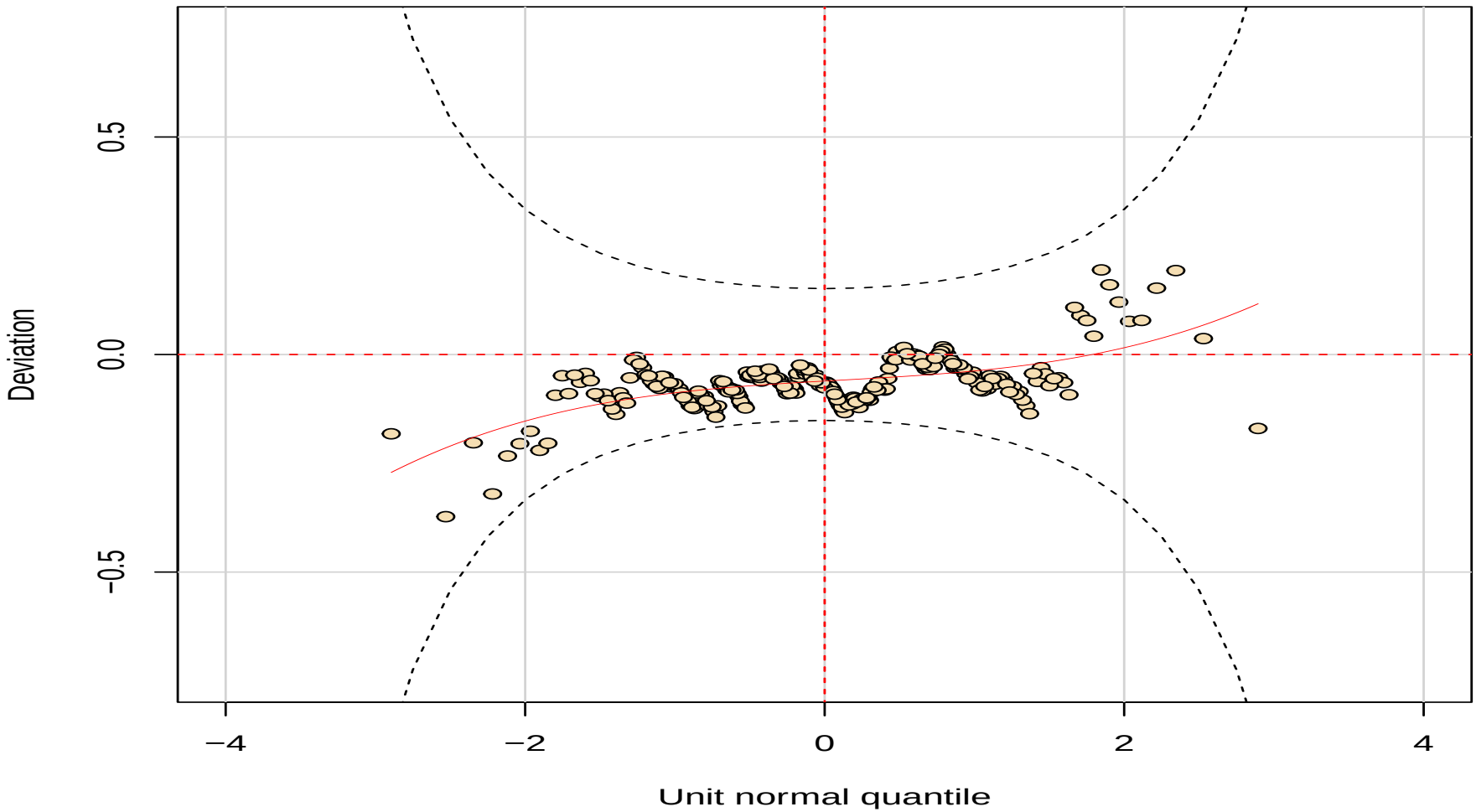
Worm plots para o modelo 1



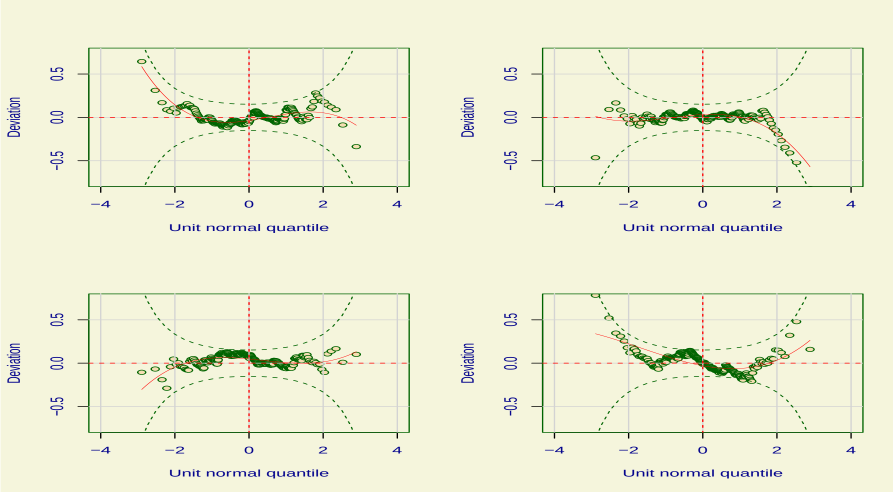
Gráficos de envelopes para o modelo 2



Worm plot para o modelo 2



Worm plots para o modelo 2



Modelo 1

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
α	-0,19	0,22	[-0,61;0,23]	-0,88	0,3806
β_2	0,23	0,31	[-0,38;0,84]	0,74	0,4573
γ_2	-0,34	0,36	[-1,04;0,37]	-0,94	0,3490
$(\beta\gamma)_{22}$	-0,43	0,55	[-1,51;0,66]	-0,78	0,4383
δ	0,05	0,02	[0,01;0,09]	2,36	0,0185

Modelo 1 sem interação sexo $\times$ estado civil

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
α	-0,12	0,20	[-0,51;0,27]	-0,62	0,5346
β_2	0,09	0,26	[-0,41;0,60]	0,37	0,7125
γ_2	-0,52	0,27	[-1,05;0,02]	-1,90	0,0574
δ	0,05	0,02	[0,01;0,09]	2,31	0,0209

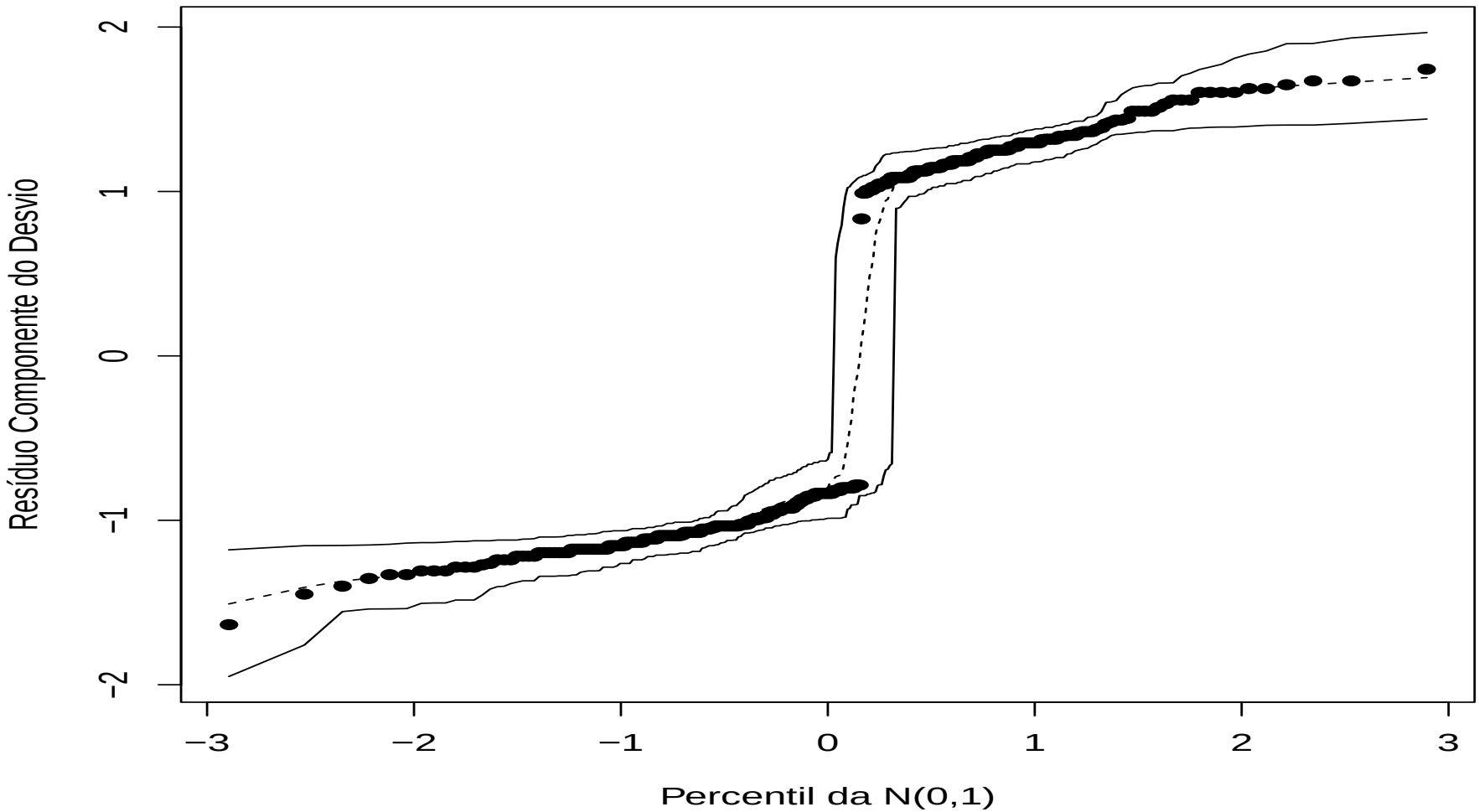
Utilização da variável idade (cont.)

- Modelo 3:

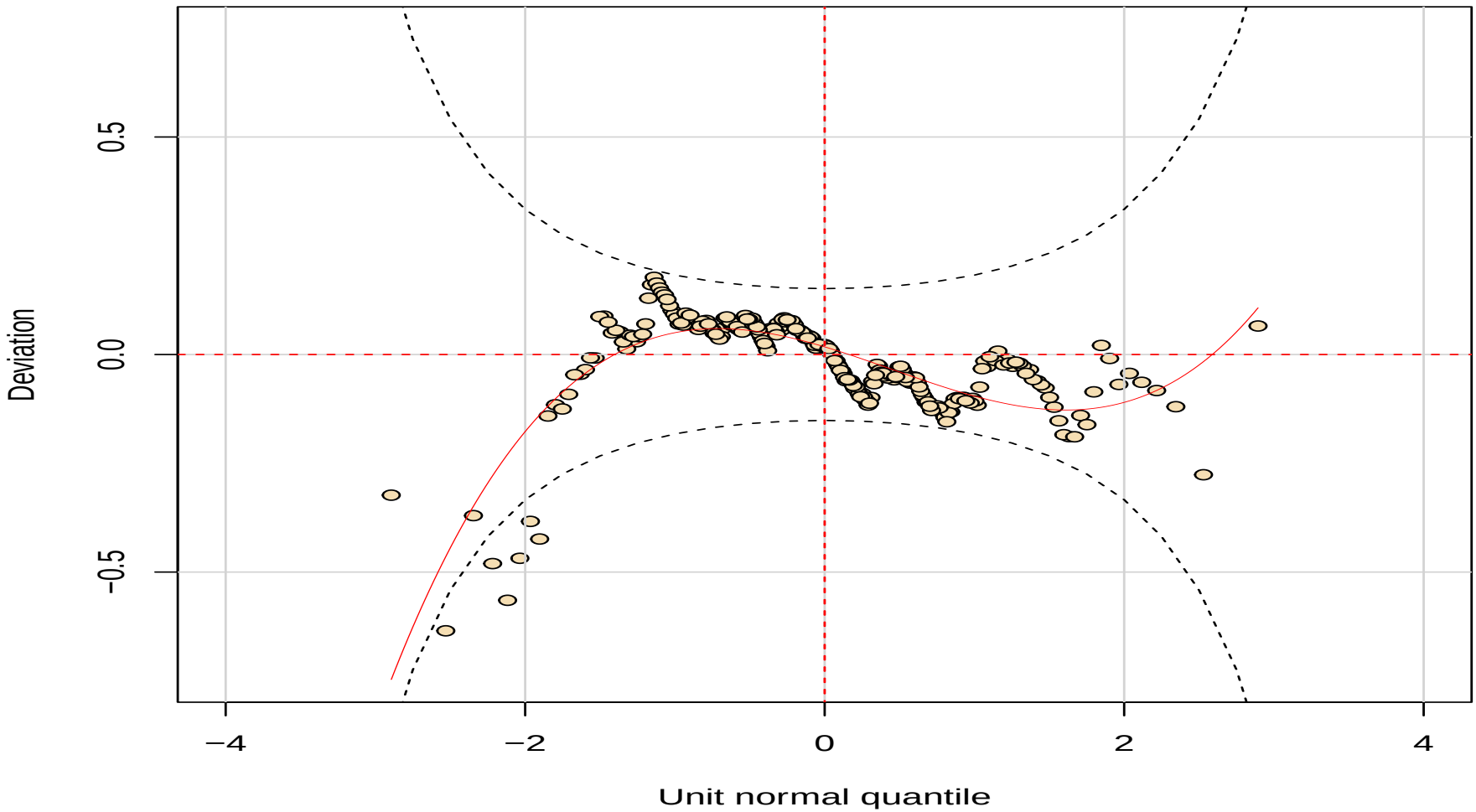
$$\begin{aligned} \ln\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) &= \alpha + \gamma_j + \delta(x_{ijk} - \overline{x}), \\ & \quad i = 1, 2, j = 1, 2, k = 1, 2, ..., n_{ij} \\ \beta_1 = \gamma_1 &= (\beta\gamma)_{1j} = (\beta\gamma)_{i1} = \theta_1 = \lambda_1 \\ &= (\theta\lambda)_{1j} = (\theta\lambda)_{i1} = 0, \forall i, j. \end{aligned}$$

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

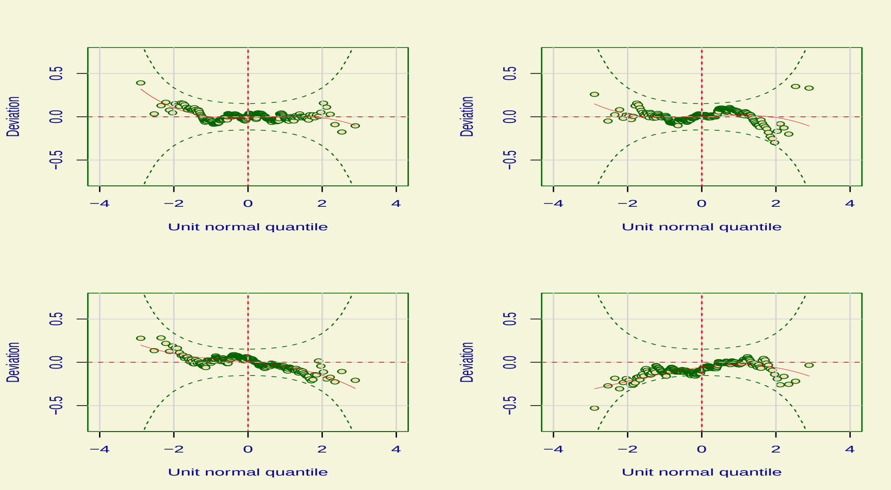
Gráficos de envelopes para o modelo 3: RCD



Wrom plot de envelopes para o modelo 3



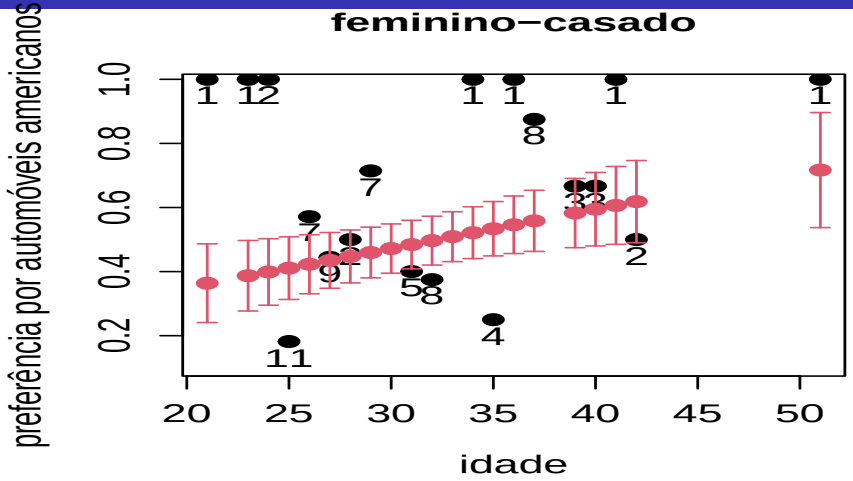
Worm plots de envelopes para o modelo 3

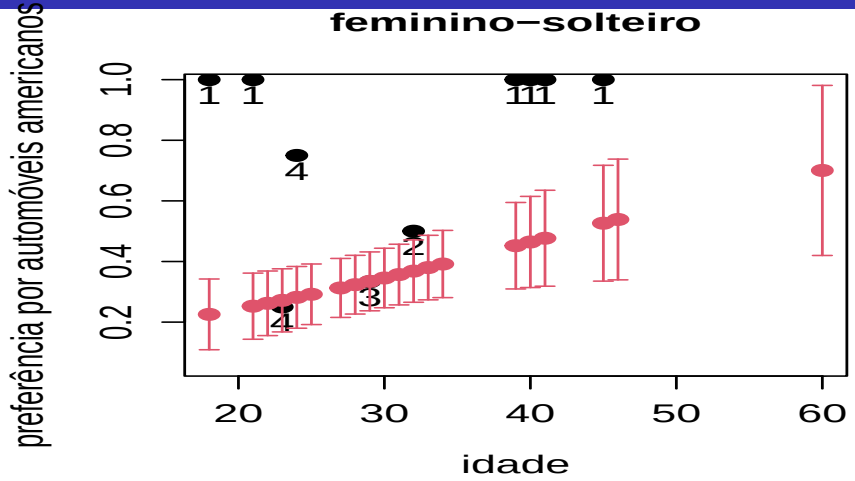


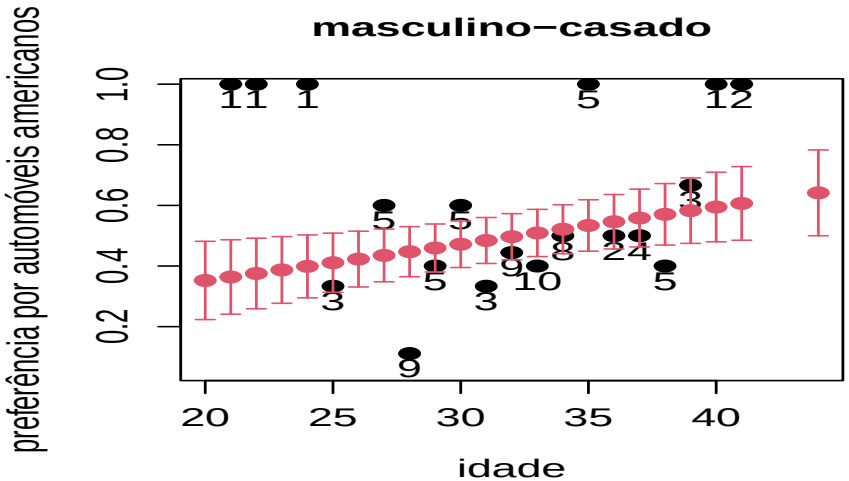
Ajuste do modelo 3

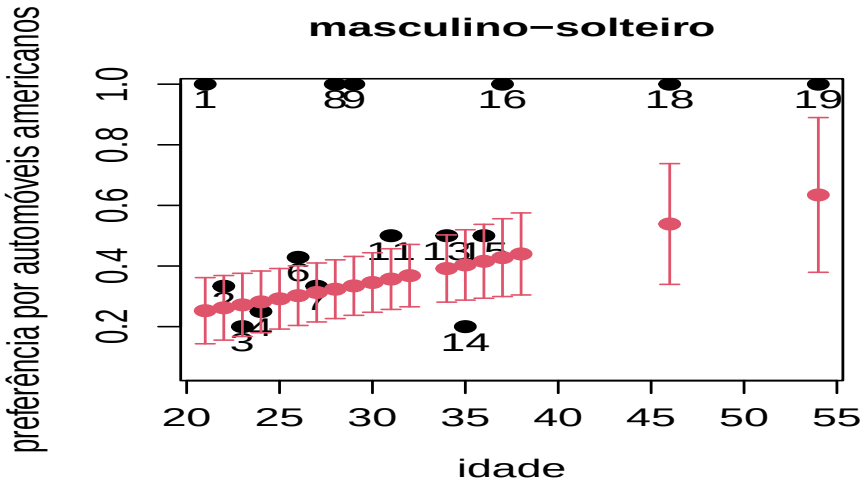
Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
α	-0,08	0,16	[-0,38 ; 0,23]	-0,50	0,6167
γ_2	-0,53	0,27	[-1,06 ;0,01]	-1,94	0,0529
δ	0,05	0,02	[0,01 ; 0,09]	2,30	0,0213

Proporções observadas e preditas pelo modelo 3

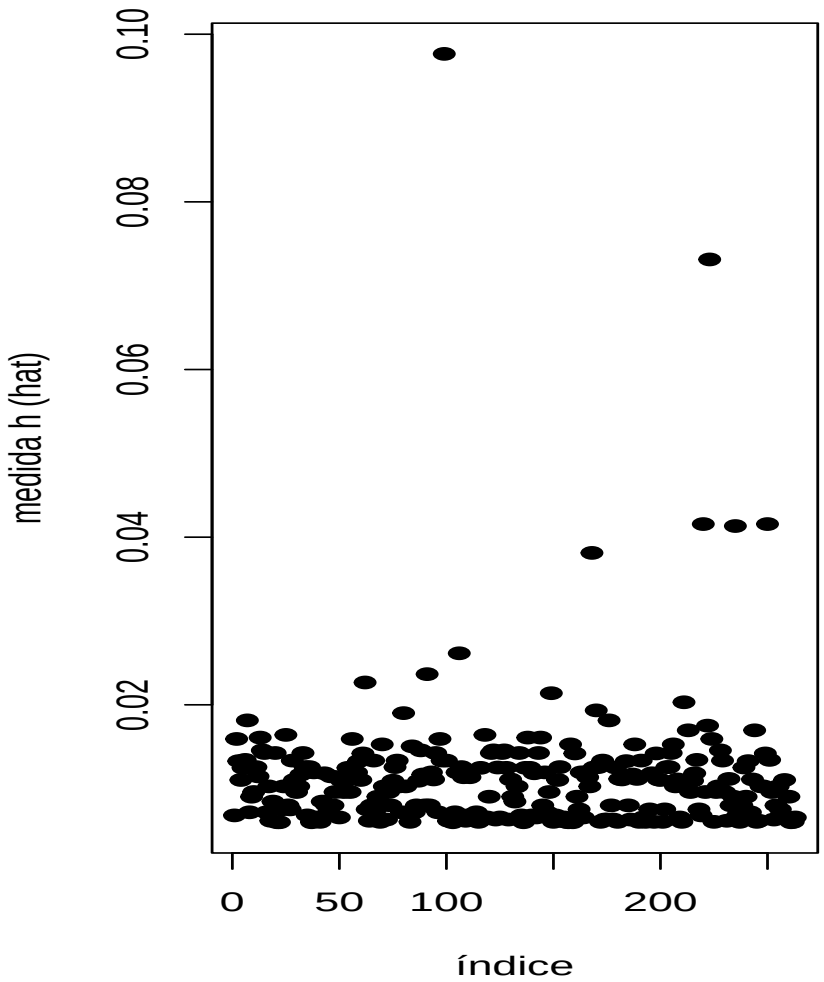


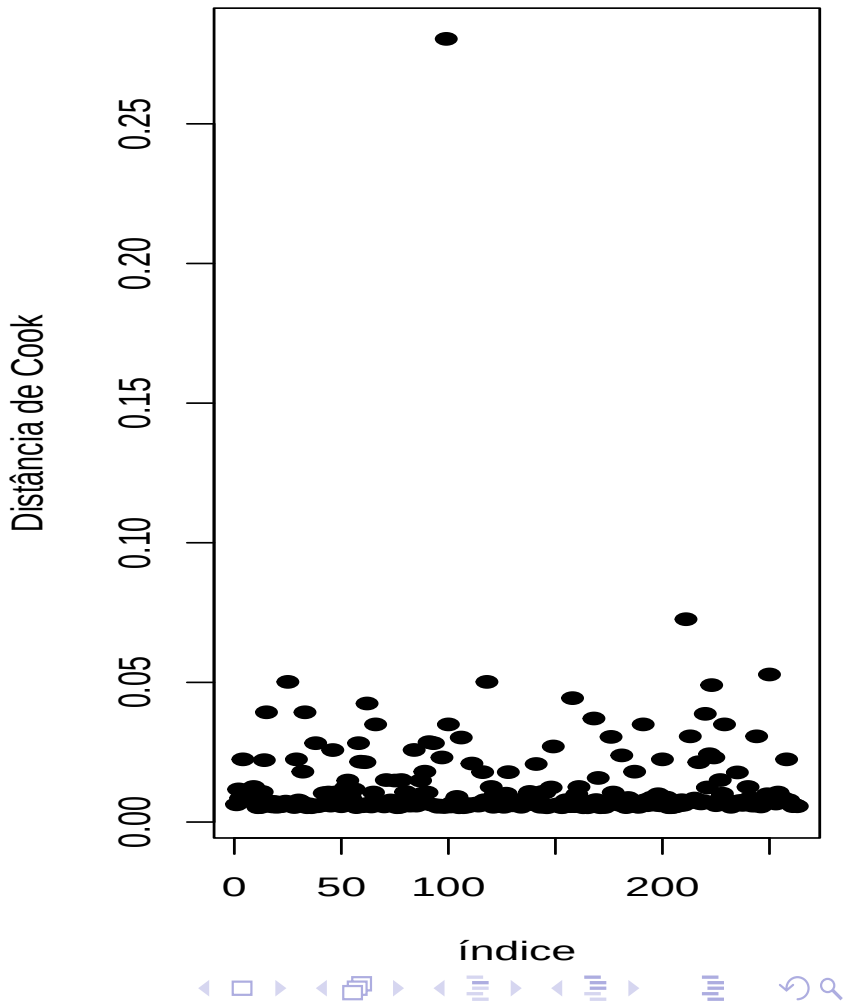






Análise de influência





Análise de Sensibilidade

Parâmetros α

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
todas	-0,08	0,16	[-0,38;0,23]	-0,50	0,6167
-# 99	-0,08	0,16	[-0,39;0,22]	-0,54	0,5886
-# 223	-0,08	0,16	[-0,39;0,22]	-0,54	0,5886

Análise de Sensibilidade

Parâmetros γ_2

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
todas	-0,53	0,27	[-1,06;0,01]	-1,94	0,0529
-# 99	-0,55	0,27	[-1,09;-0,01]	-2,01	0,0440
-# 223	-0,55	0,27	[-1,09;-0,01]	-2,01	0,0440

Análise de Sensibilidade

Parâmetros δ

Parâmetro	Estimativa	EP	IC(95%)	Estat. Z_t	p-valor
todas	0,05	0,02	[0,01;0,09]	2,30	0,0213
-# 99	0,05	0,02	[0,00;0,09]	2,04	0,0409
-# 223	0,05	0,02	[0,00;0,09]	2,04	0,0409