Modelos para dados de contagem: parte 2

Prof. Caio Azevedo

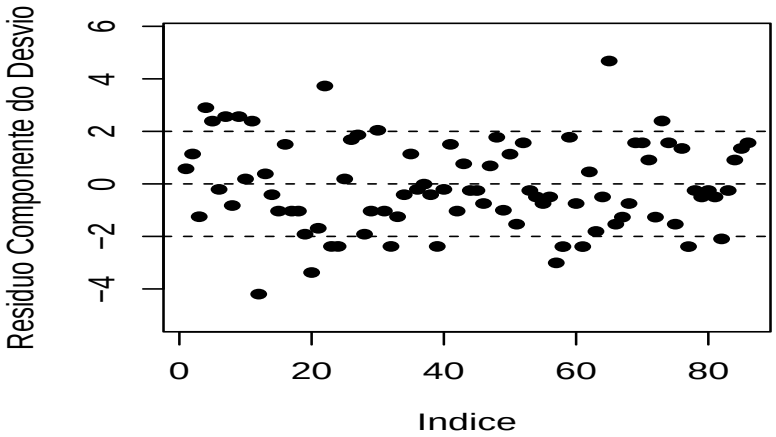
Exemplo 4: comparação do número de acidentes

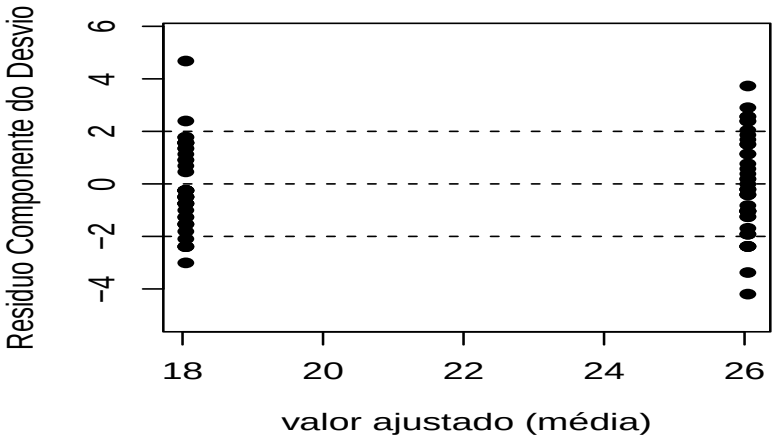
Modelo

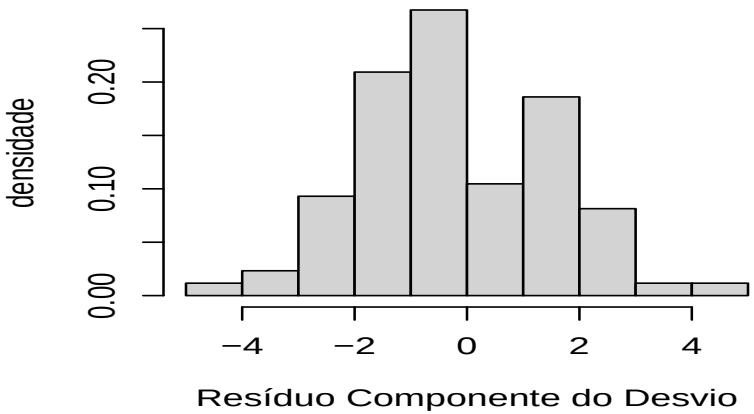
$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_i), i = 1\,(\text{ano de } 1961), 2\,(\text{ano de } 1962), j = 1, ..., 43 \\
\ln \mu_i &= \mu + \alpha_i, \alpha_1 = 0
\end{aligned}
$$

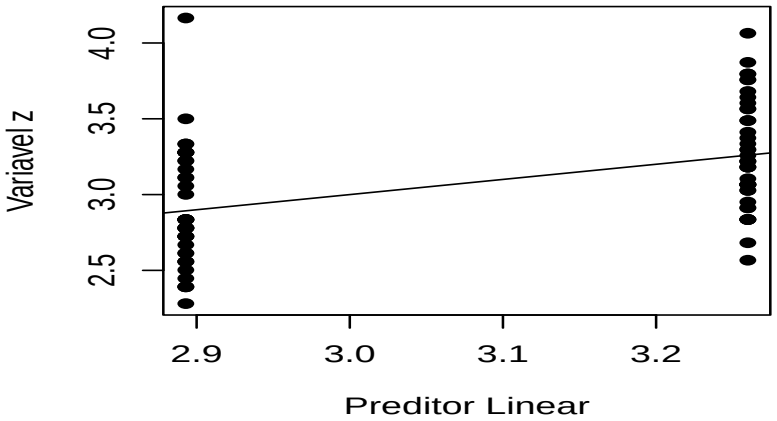
- $\mathcal{E}(Y_{ij}) = \mu_i = e^{\mu+\alpha_i}$.
- e^{α_2}: o incremento multiplicativo (positivo ou negativo) da média do ano de 1962 em relação à média do ano de 1961 ($\mu_2 = \mu_1 e^{\alpha_2}$).

Gráficos de diagnóstico: Exemplo 4

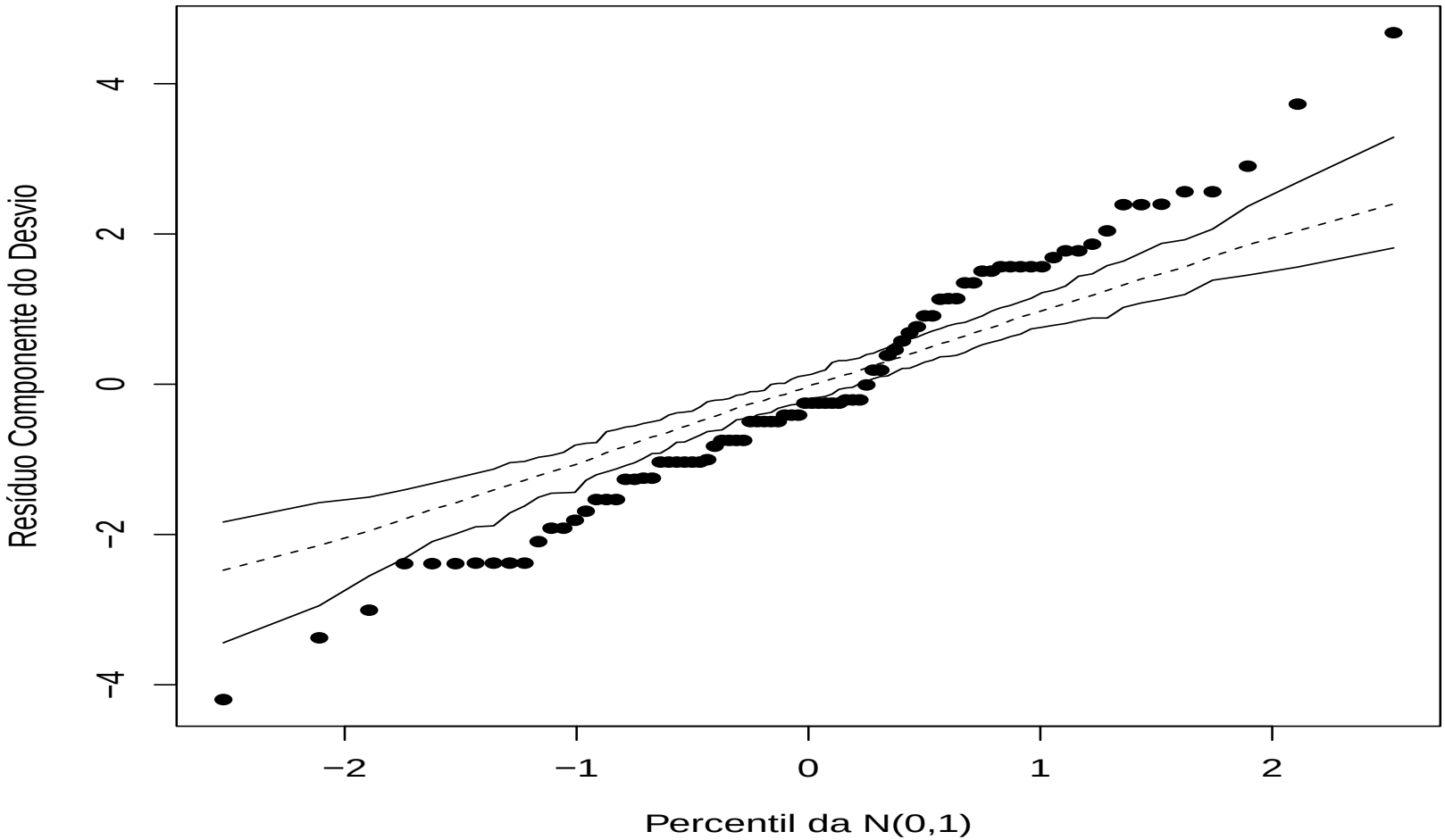




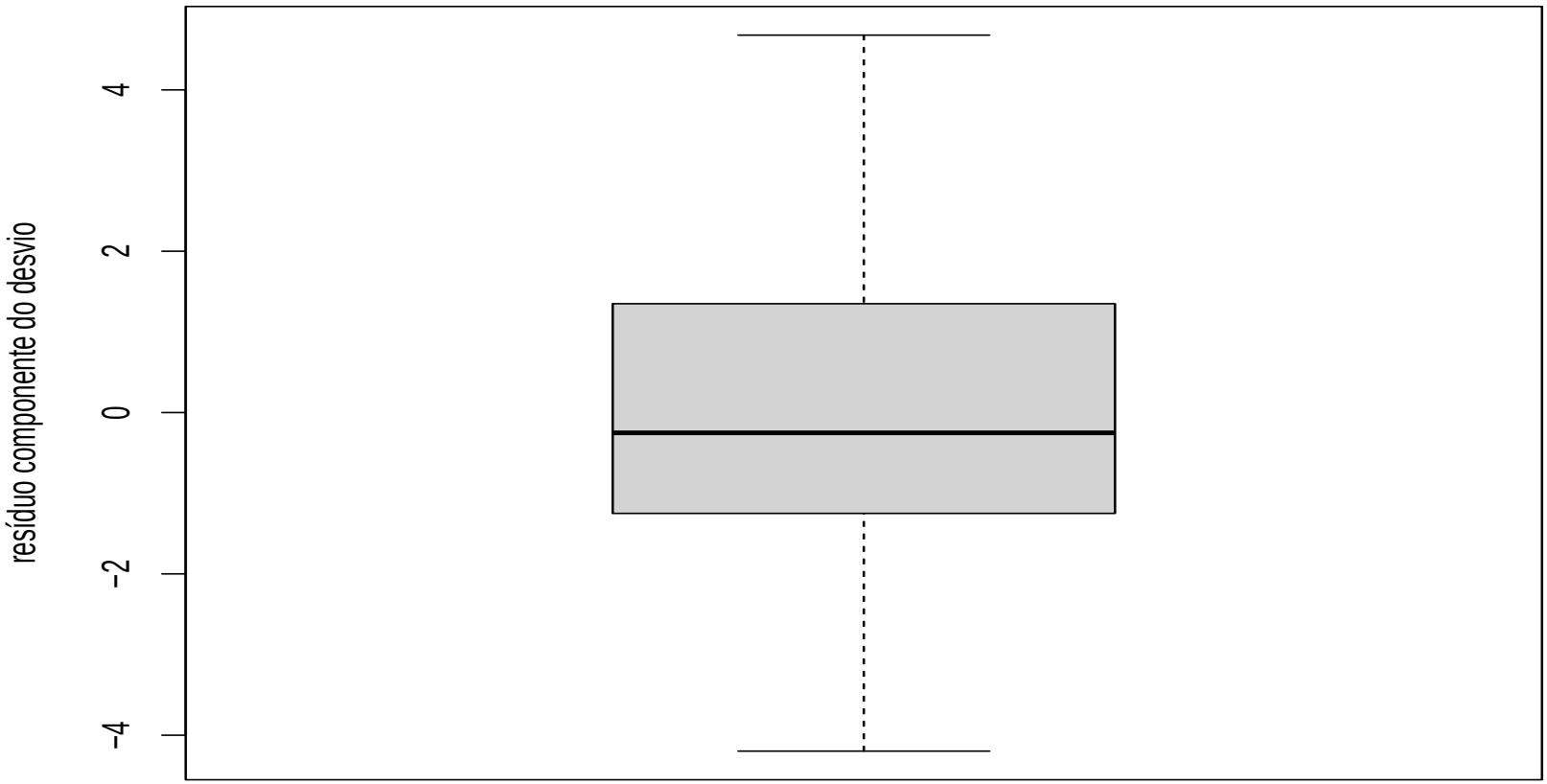




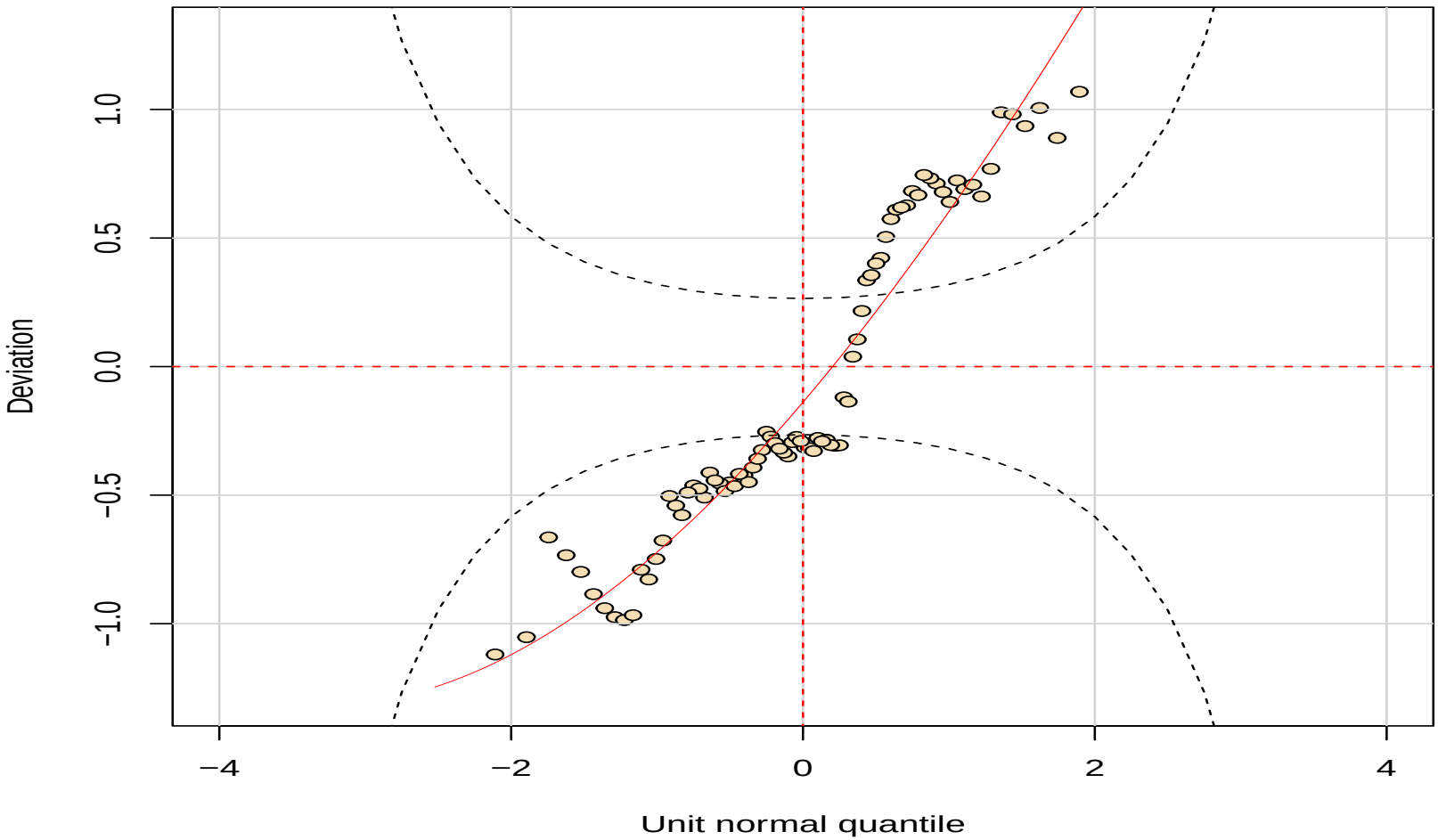
Envelope para os resíduos: Exemplo 4



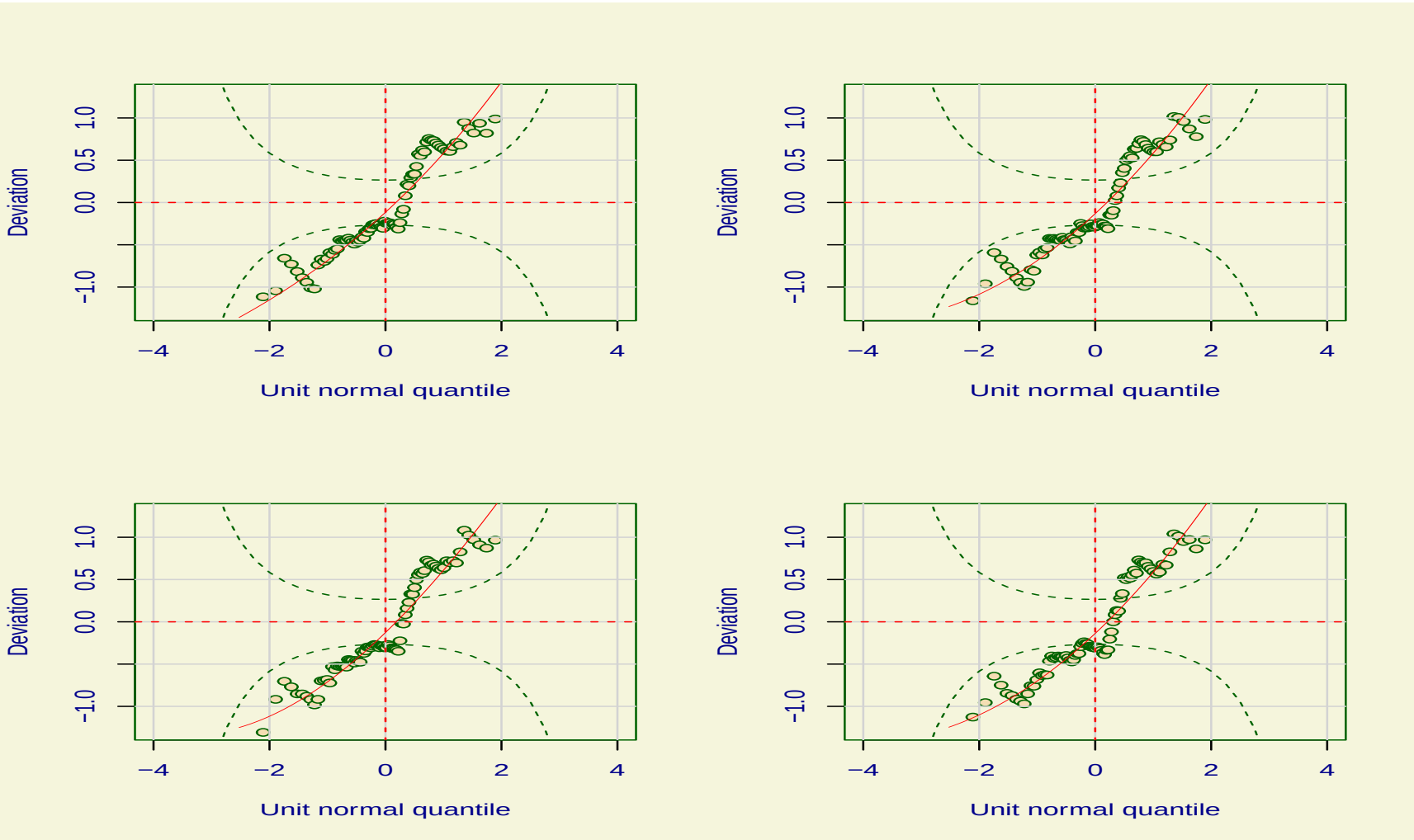
Box plot para os resíduos: Exemplo 4



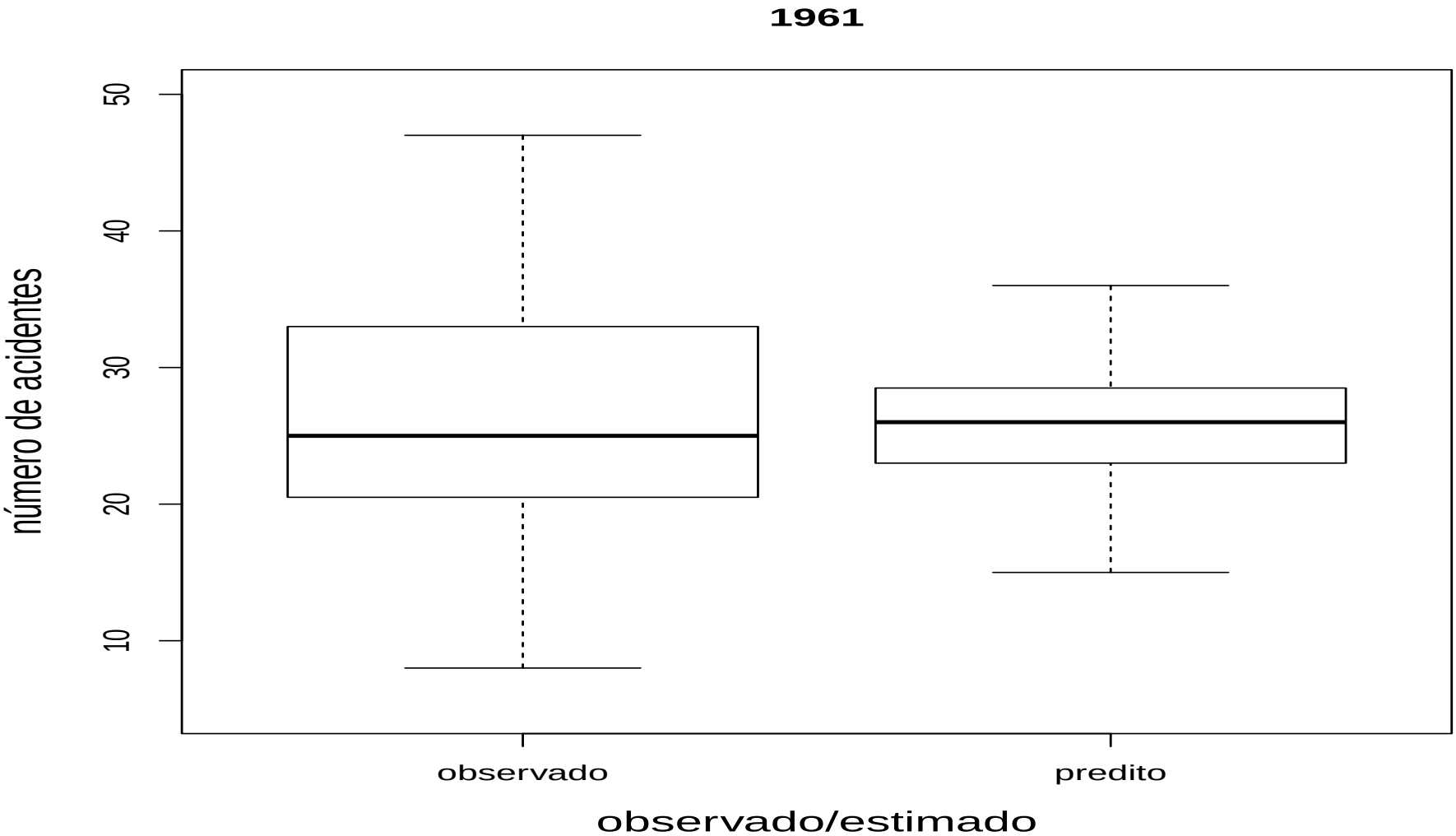
Worm plot para os resíduos: Exemplo 4



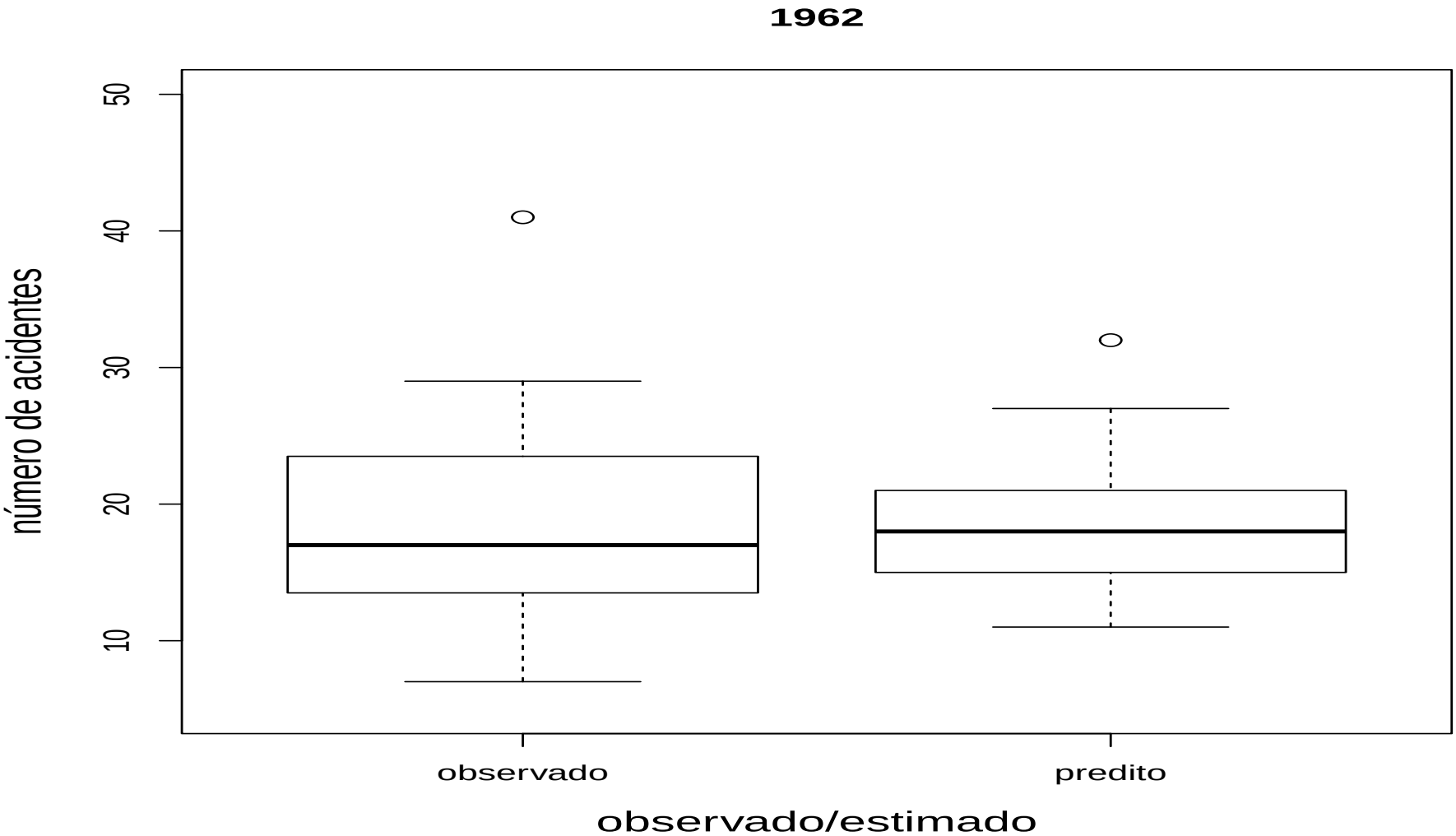
Worm plots para os resíduos: Exemplo 4



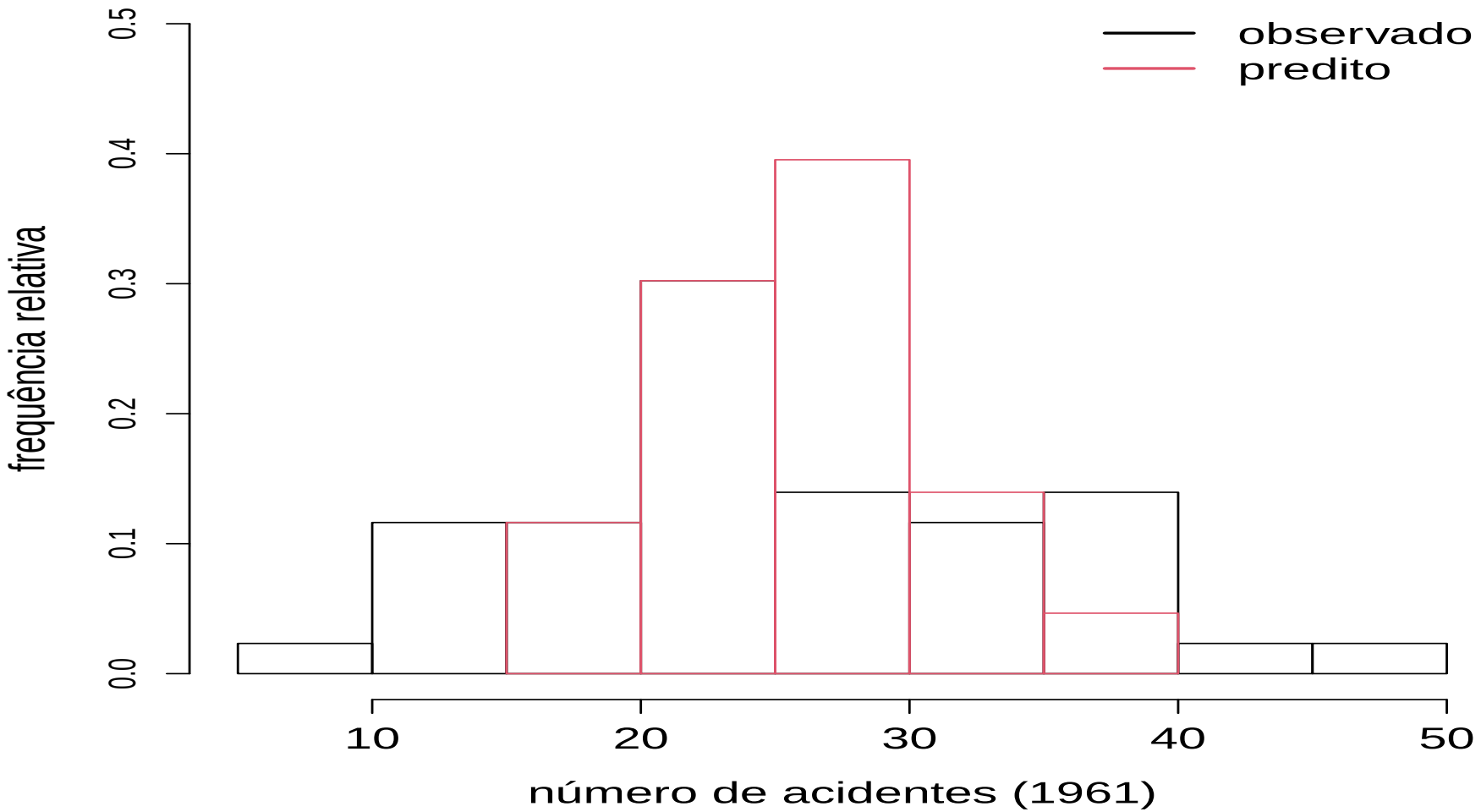
Distribuições preditas e observadas (boxplot)



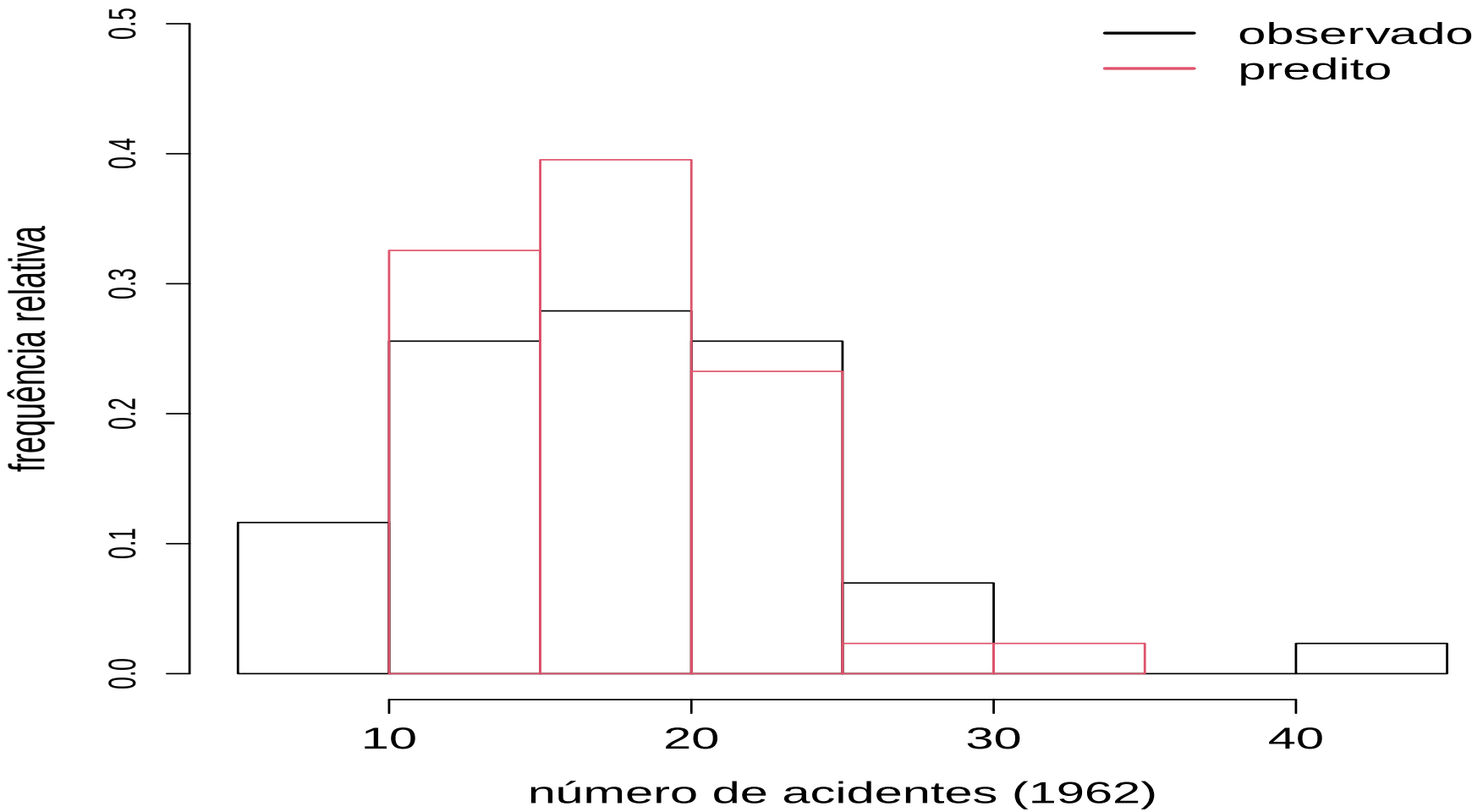
Distribuições preditas e observadas (boxplot)



Distribuições preditas e observadas (histograma)



Distribuições preditas e observadas (histograma)



Comentários

- Temos indicações de normalidade e homocedasticidade dos RCD.
- Contudo, o gráfico QQ e o worm plot indicam um péssimo ajuste, sugerindo a existência de superdispersão.
- Adicionalmente, $D(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}) = 235,17$ (p $= <$ 0,0001) (considerando-se a aproximação pela distribuição $\chi^2_{(84)}$ adequada), o que indica que o modelo não se ajustou bem aos dados.
- Se o problema, em relação ao mal ajuste, estiver sendo causado por superdispersão (o que parece ser o caso), os erros-padrão estão sendo submestimados.

Estimativas dos parâmetros dos modelos

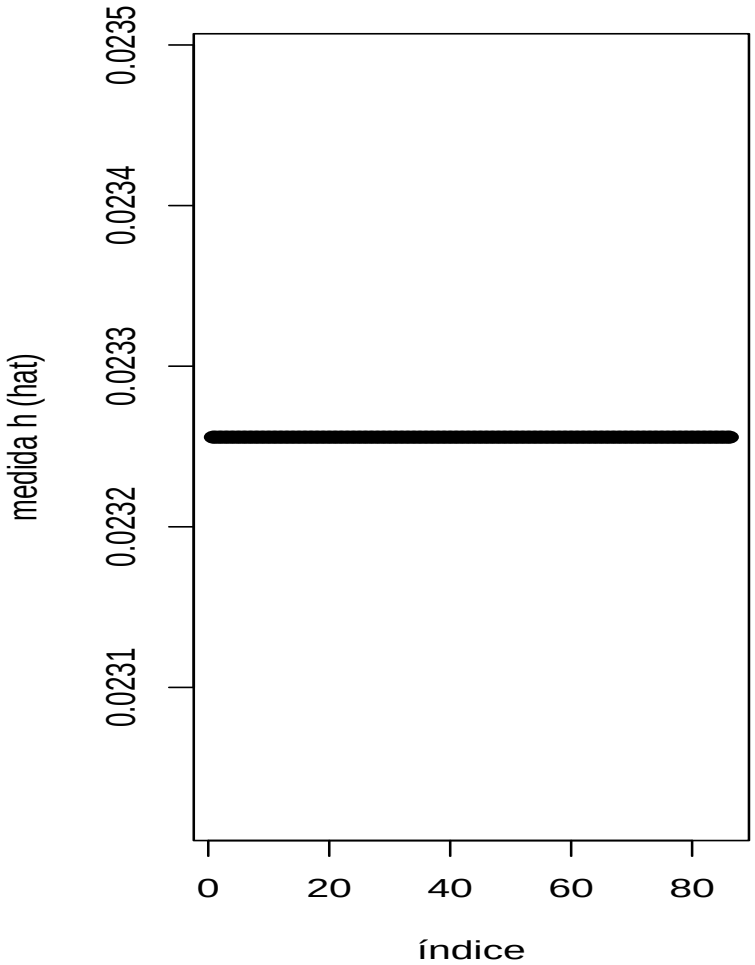
Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
μ	3,26	0,03	[3,20 ; 3,32]	109,10	$<$ 0,0001
α_2	-0,37	0,05	[-0,46 ; -0,28]	-7,86	$<$ 0,0001

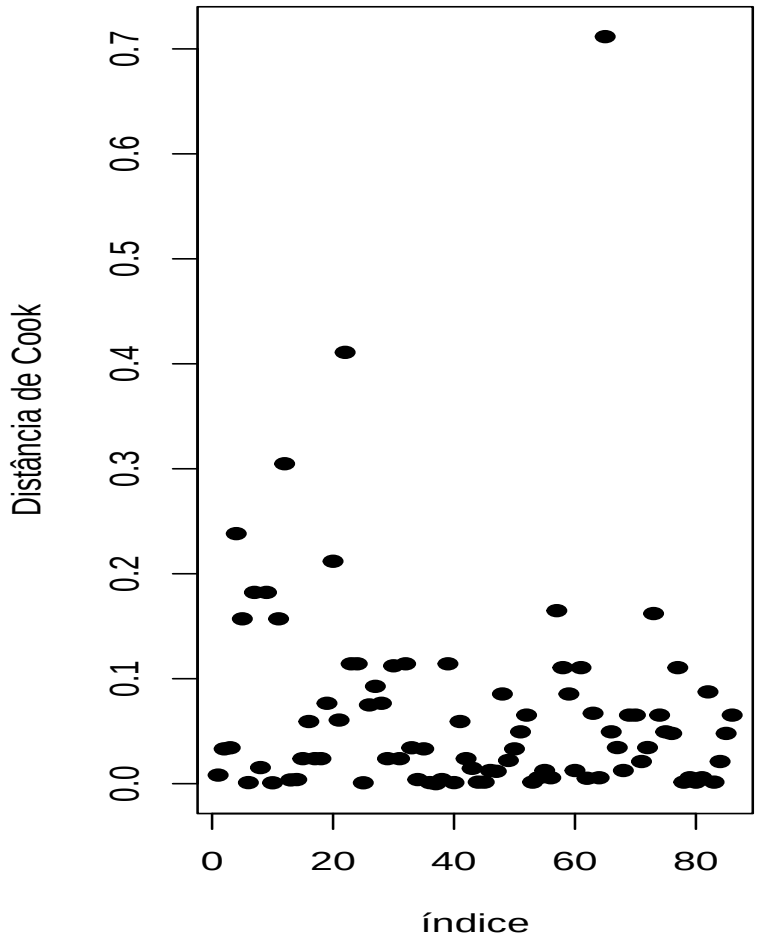
Médias preditas pelo modelo

Par.	Est.	EP	IC(95%)
μ_1	26,05	0,78	[24,52 ; 27,57]
μ_2	18,05	0,65	[16,78 ; 19,32]

- (Relembrando, o modelo não se ajustou bem). Contudo, aparentemente, houve uma redução significativa (do ponto de vista estatístico) no que concerne ao número de acidentes.
- Além disso, as médias preditas são iguais as médias observadas e os respectivos IC's contem os valoes observados.

Pontos alavanca e Distância de Cook





Estimativas dos parâmetros dos modelos

Com todos os pontos

Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
μ	3,26	0,03	[3,20 ; 3,32]	109,10	$<$ 0,0001
α_2	-0,37	0,05	[-0,46 ; -0,28]	-7,86	$<$ 0,0001

Sem a observação # 65

Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
μ	3,26	0,03	[3,20;3,32]	109,10	$<$ 0,0001
α_2	-0,40	0,05	[-0,49;-030]	-8,38	$<$ 0,0001

Exemplo 11: perfil dos clientes de uma loja

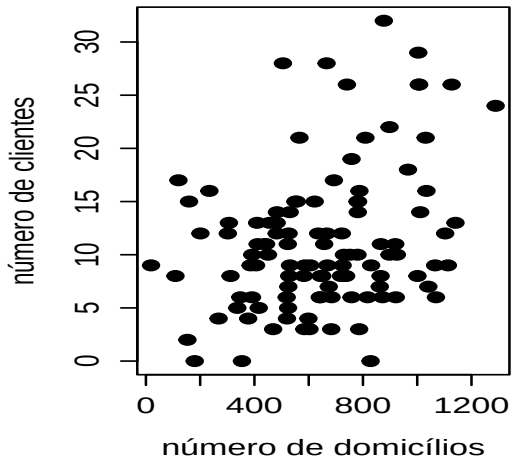
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável resposta).

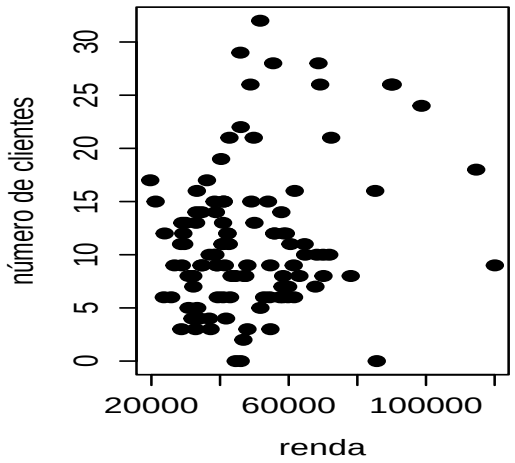
Cont.

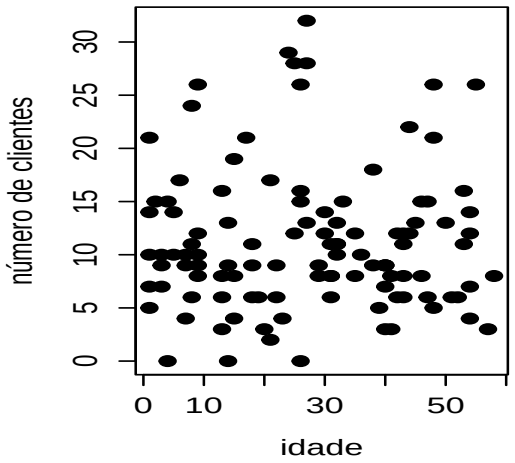
- Variáveis explicativas: número de domicílios (em milhares) (x_1), renda média anual (em milhares de USD) (x_2), idade média dos domicílios (em anos)(x_3), distância ao concorrente mais próximo (em milhas) (x_4) e distância à loja (em milhas) (x_5).
- Variável resposta : número de clientes da referida loja (Y) (contagem).

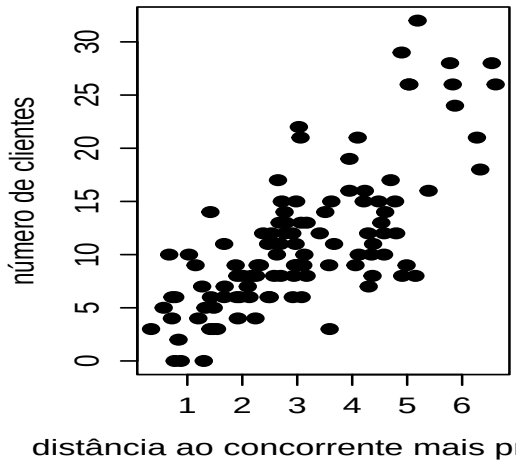
Mapa de Fortaleza (ilustração)

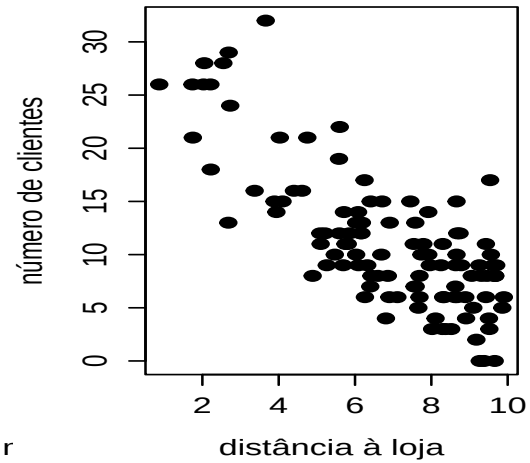
Gráficos de dispersão











Legenda

- ndom - número de domicílios.
- renda - renda média anual.
- idade - idade média dos domicílios.
- disc - distância ao concorrente mais próximo.
- disl - distância à loja

Medidas resumo

	Variável				
Medida-resumo	ndom	renda	idade	dist	disl
Média	647,76	48836,78	27,43	3,07	6,83
DP	263,03	18531,06	16,68	1,50	2,29
CV(%)	40,61	37,94	60,83	49,02	33,54
Mediana	647,00	44564,50	27,00	2,93	7,28
Mínimo	19,00	19673,00	1,00	0,34	0,87
Máximo	1289,00	120065,00	58,00	6,61	9,90
CA	-0,02	1,31	0,03	0,34	-0,68
Curt.	2,55	5,31	1,80	2,41	2,61

Modelo (completo)

$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \text{Poisson}(\mu_i) \\
\ln(\mu_i) &= \beta_0 + \beta_1\left(\frac{x_{1i}-\overline{x}_1}{s_1}\right) + \beta_2\left(\frac{x_{2i}-\overline{x}_2}{s_2}\right) + \beta_3\left(\frac{x_{3i}-\overline{x}_3}{s_3}\right) + \\
&+ \beta_4\left(\frac{x_{4i}-\overline{x}_4}{s_4}\right) + \beta_5\left(\frac{x_{5i}-\overline{x}_5}{s_5}\right), \\
\mu_i &= \exp\left\{\beta_0 + \beta_1\left(\frac{x_{1i}-\overline{x}_1}{s_1}\right) + \beta_2\left(\frac{x_{2i}-\overline{x}_2}{s_2}\right) + \beta_3\left(\frac{x_{3i}-\overline{x}_3}{s_3}\right) + \right. \\
&+ \left. \beta_4\left(\frac{x_{4i}-\overline{x}_4}{s_4}\right) + \beta_5\left(\frac{x_{5i}-\overline{x}_5}{s_5}\right)\right\}, i = 1, 2, ..., 110
\end{aligned}
$$

Modelo (completo)

- x_{ji} : valor da variável explicativa j, associada à área i,

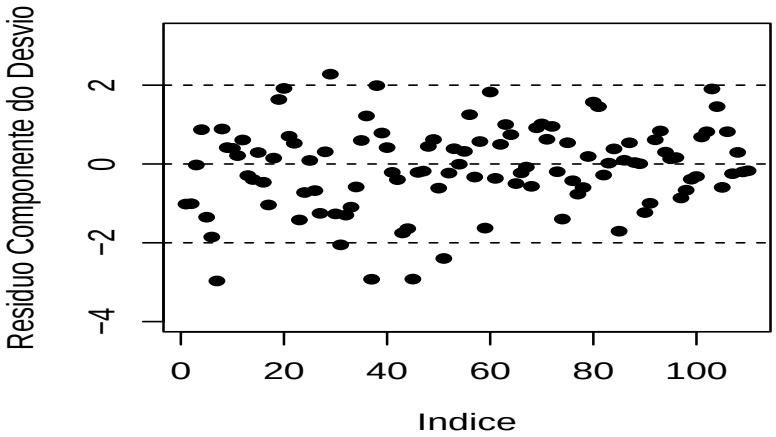
$$\overline{x}_j = \frac{1}{110}\sum_{i=1}^{110} x_{ji}, \text{ e } s_j = \frac{\sum_{i=1}^{n}(x_{ji} - \overline{x}_j)^2}{109} \quad j = 1, 2, ..., 5.$$

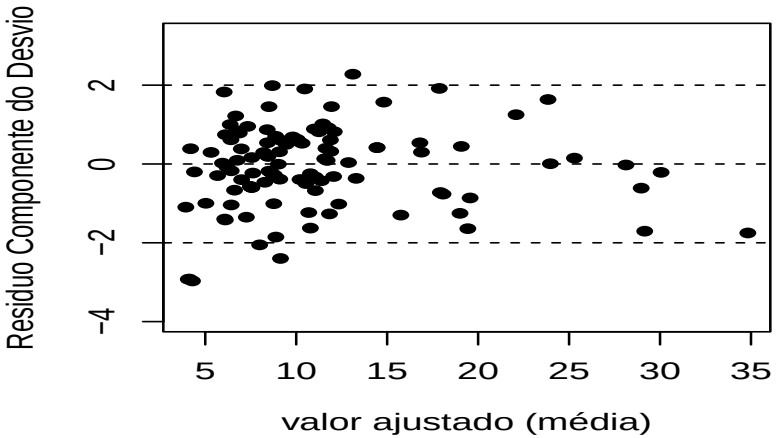
- e^{β_0} : número esperado de clientes para domicílios localizados em áreas com valor médio para cada uma das covariáveis.

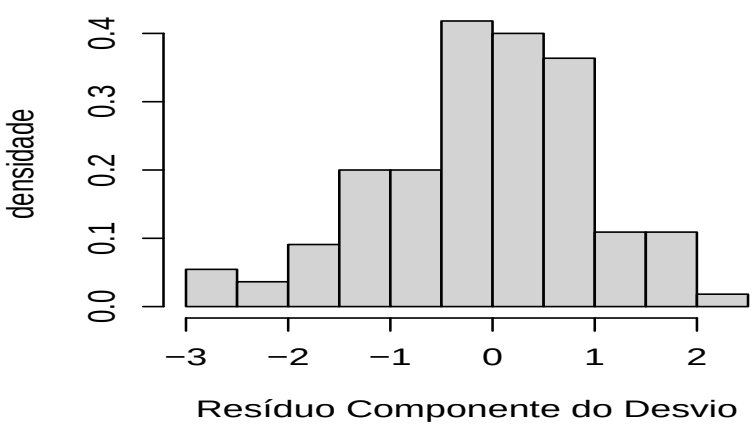
Modelo (completo)

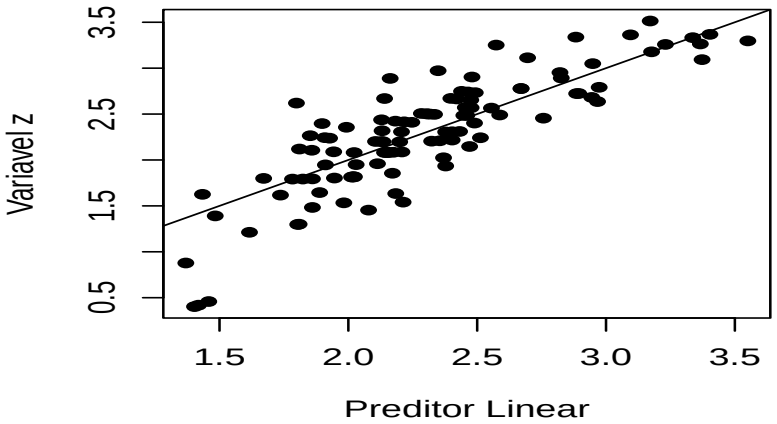
- e^{β_j / s_j} : incremento (positivo ou negativo) no valor esperado do número de clientes, para o aumento em uma unidade no valor da covariável j, mantendo-se todas as outras fixas.
- Uma vez que cada uma das covariáveis está sendo introduzida no modelo com iguais média e variância (e de forma adimensional), as magnitudes dos respectivos coeficientes podem ser diretamente comparadas.

Gráficos de diagnóstico: Modelo completo

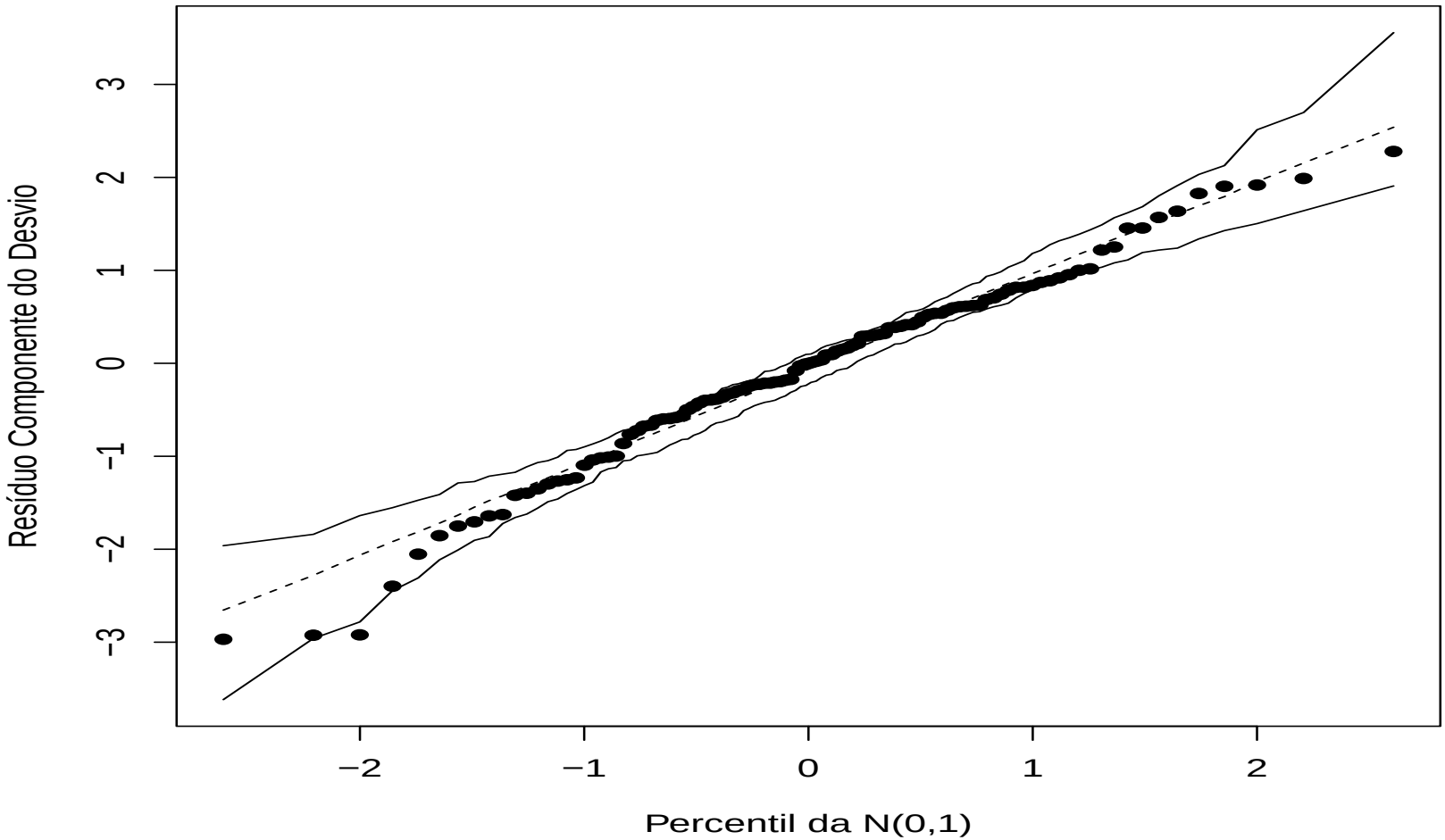




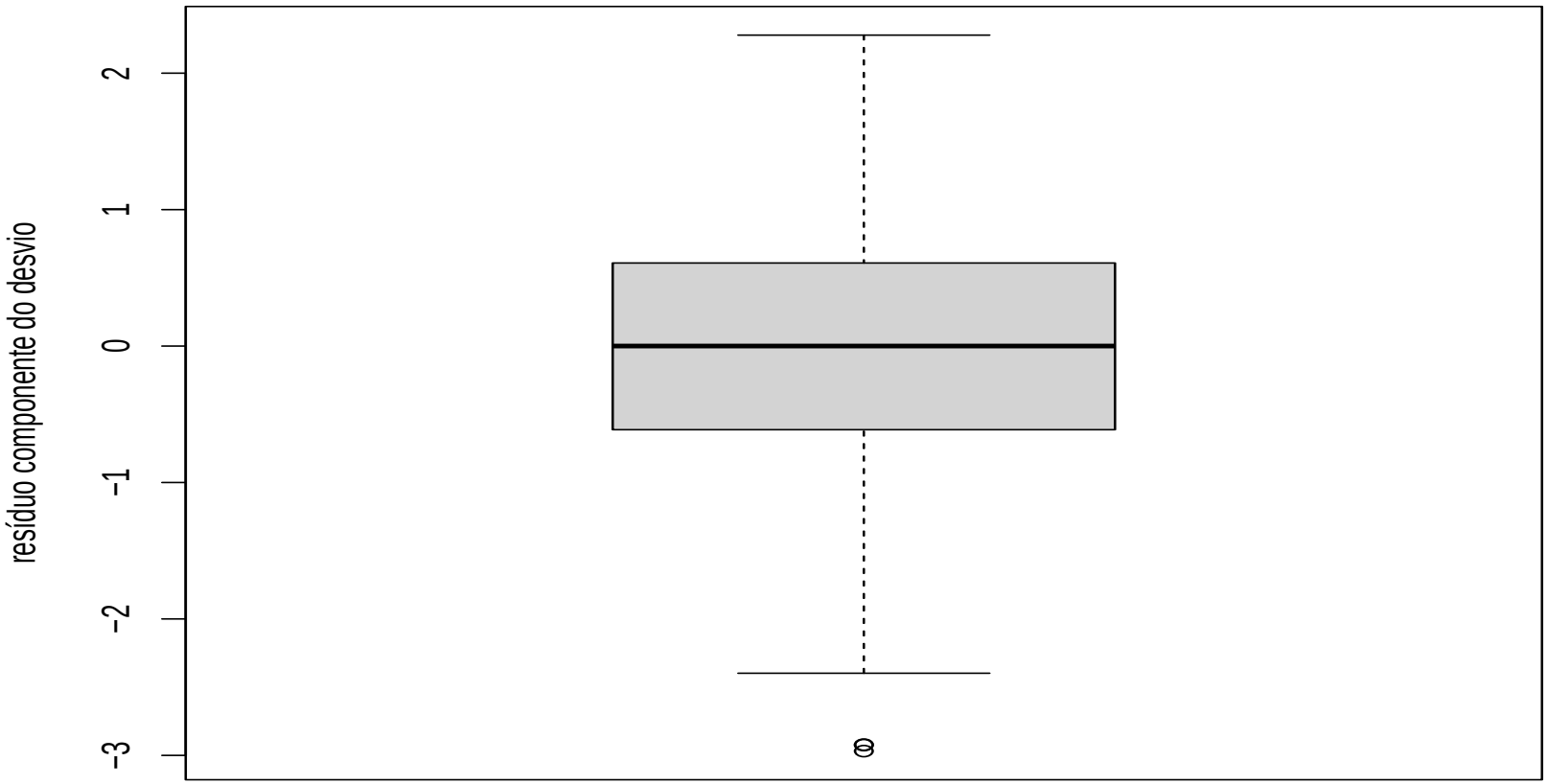




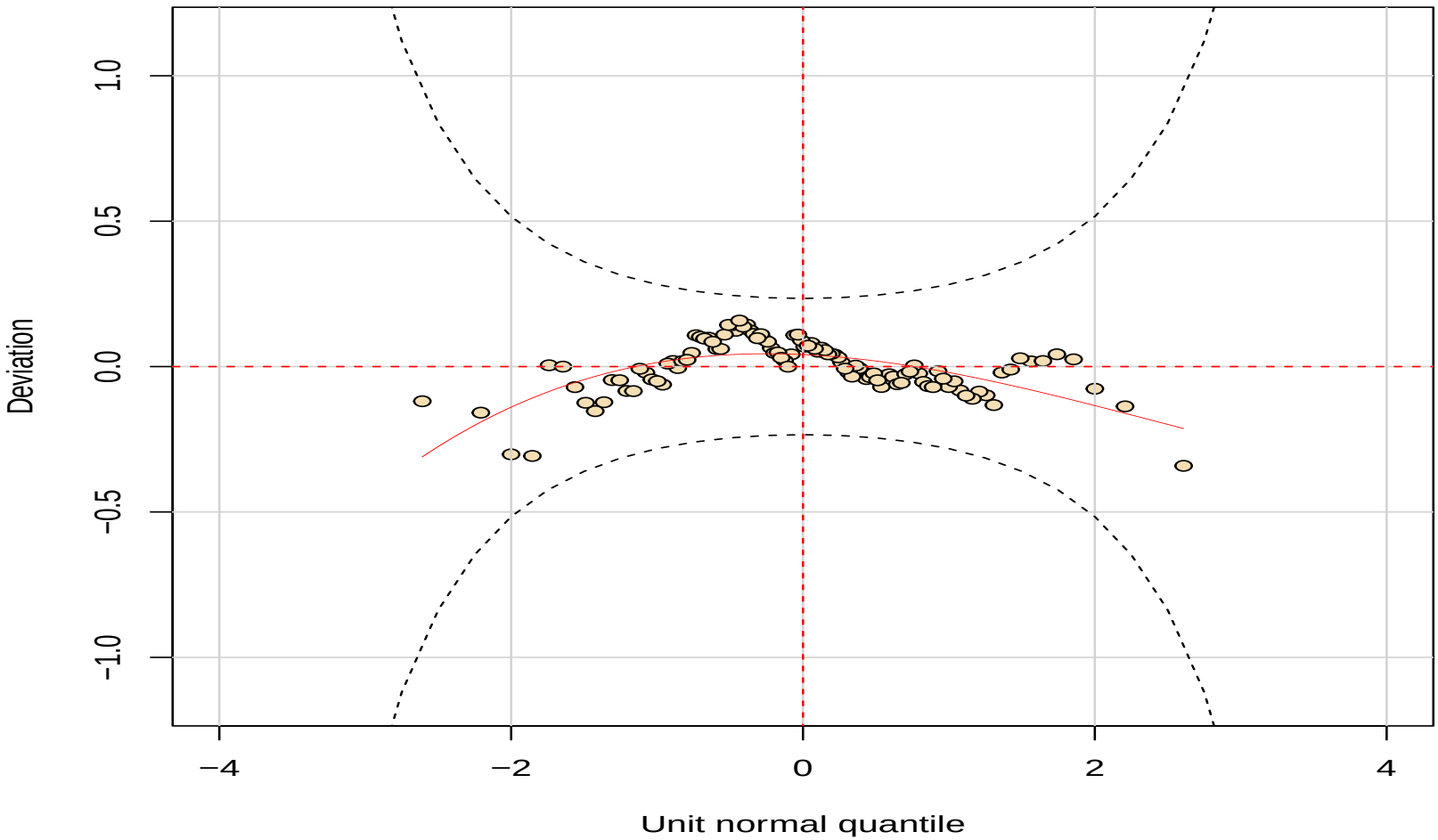
Envelope para os resíduos: Modelo completo



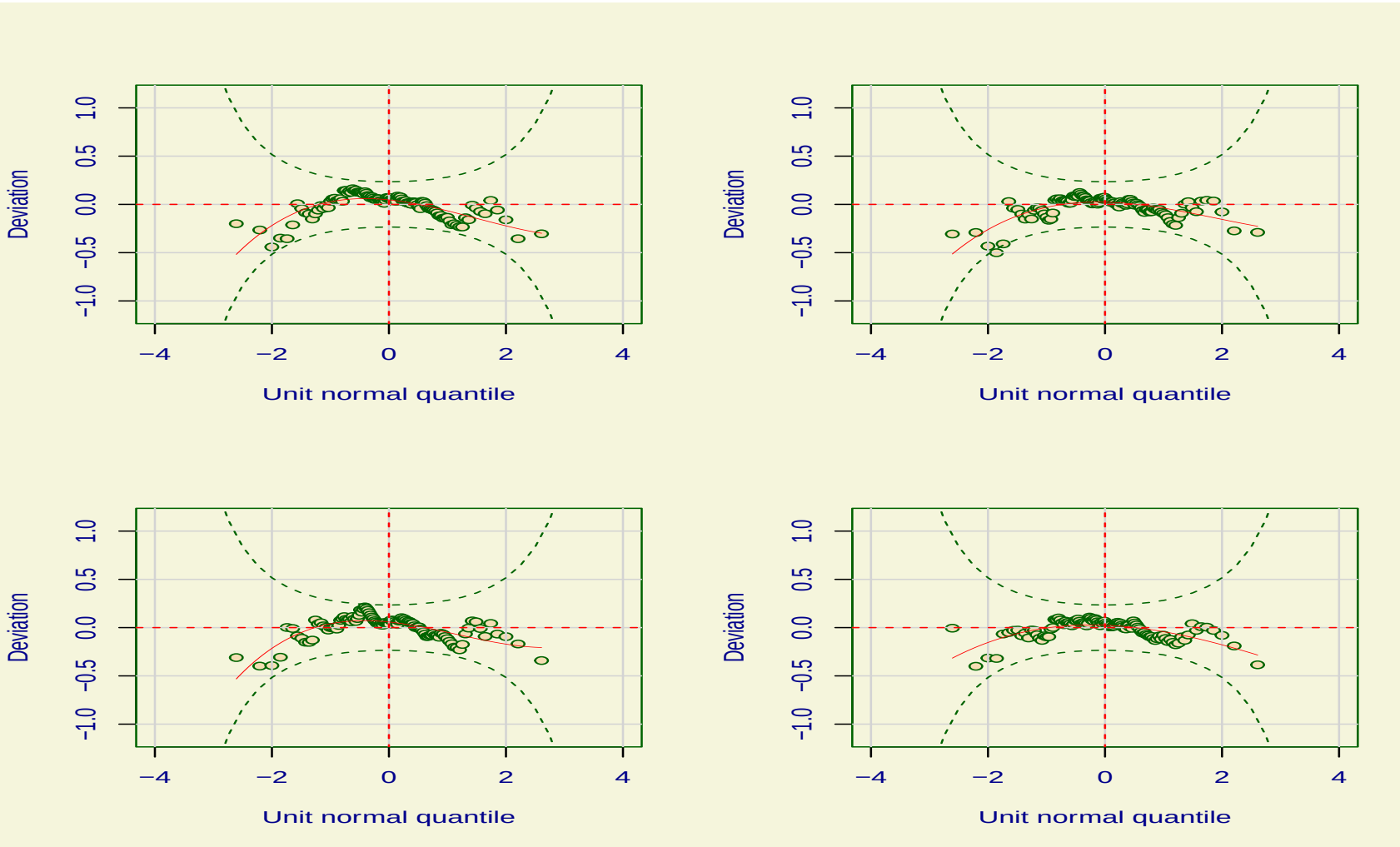
Box plot para os resíduos: Modelo completo



Worm plot para os resíduos: Modelo completo



Worm plots para os resíduos: Modelo completo



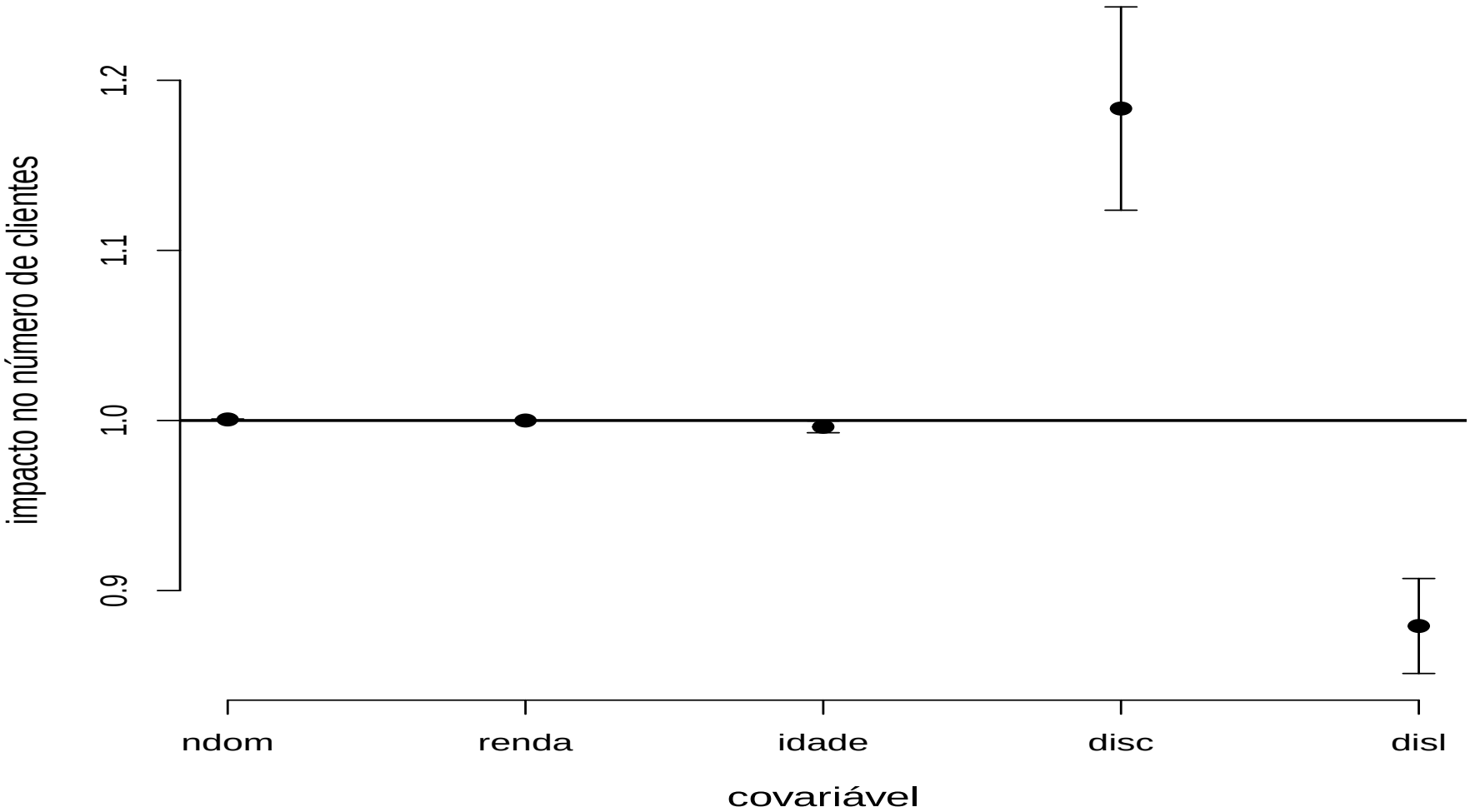
Estimativas dos parâmetros

- A análise residual indica um bom ajuste, apesar de leve assimetria negativa dos RCD.
- $D(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}) = 114,95$ (p = 0,2170) (considerando a aproximação pela distribuição $\chi^2_{(104)}$ adequada), o que indica que o modelo se adequou bem aos dados.
- Isso sugere que o ajuste do modelo pode ser melhorado contudo, este é aceitável.

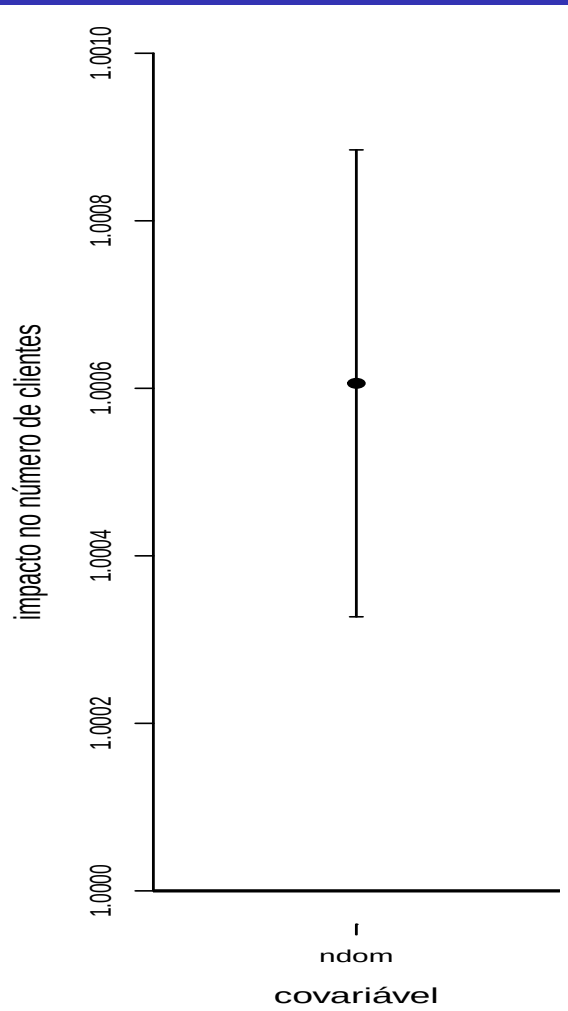
Estimativas dos parâmetros

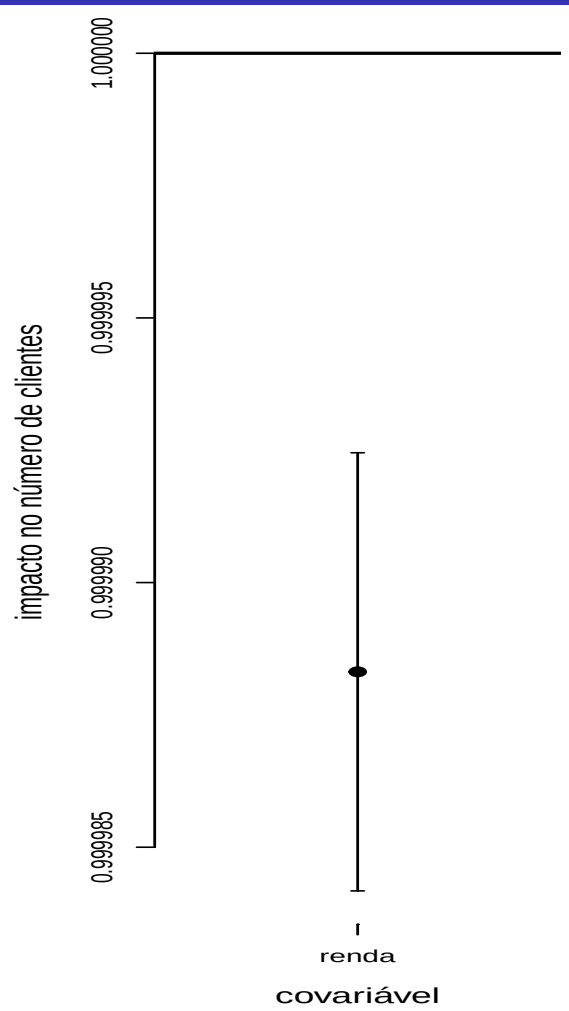
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	2,30	0,03	[2,24 ; 2,36]	72,92	$<$ 0,0001
β_1	0,16	0,04	[0,09 ; 0,23]	4,26	$<$ 0,0001
β_2	-0,22	0,04	[-0,29 ; -0,14]	-5,53	$<$ 0,0001
β_3	-0,062	0,030	[-0,120 ; -0,003]	-2,091	0,0365
β_4	0,25	0,04	[0,18 ; 0,33]	6,53	$<$ 0,0001
β_5	-0,30	0,04	[-0,37 ; -0,22]	-7,95	$<$ 0,0001

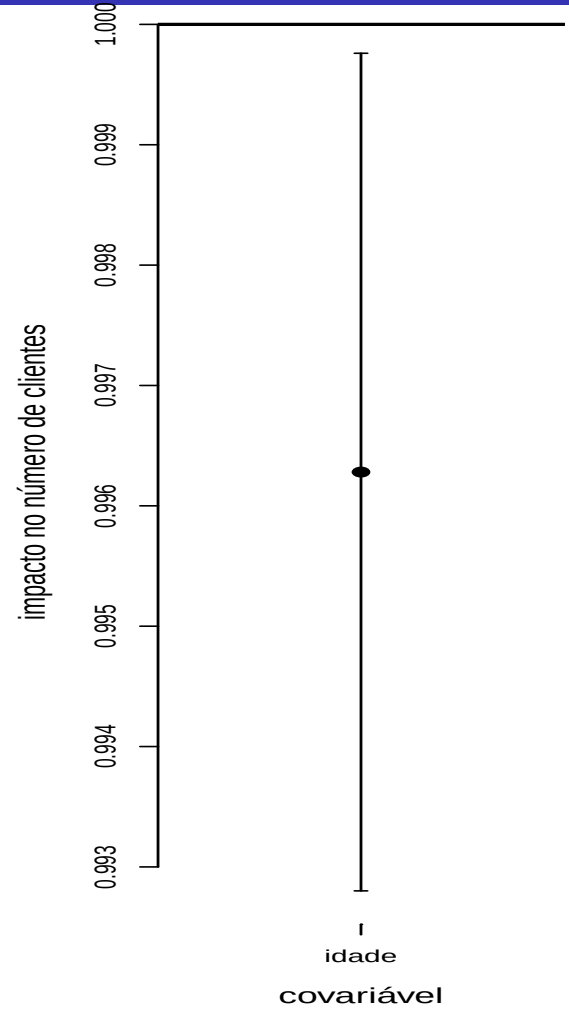
Estimativa do impacto que cada covariável (e^{β_j / s_j})



Estimativa do impacto das 3 primeiras covariáveis (e^{β_j / s_j})





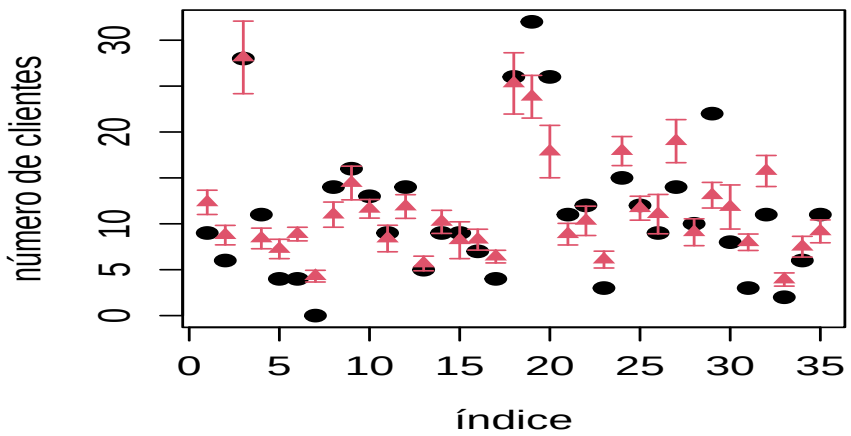


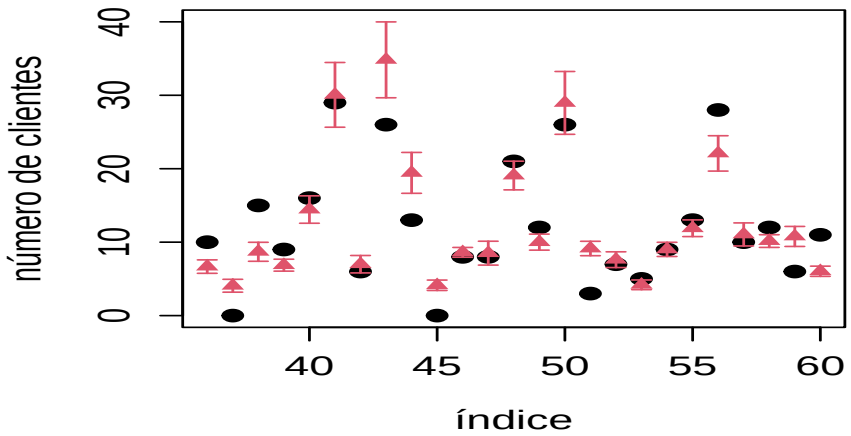
Mais sobre a escolha do modelo

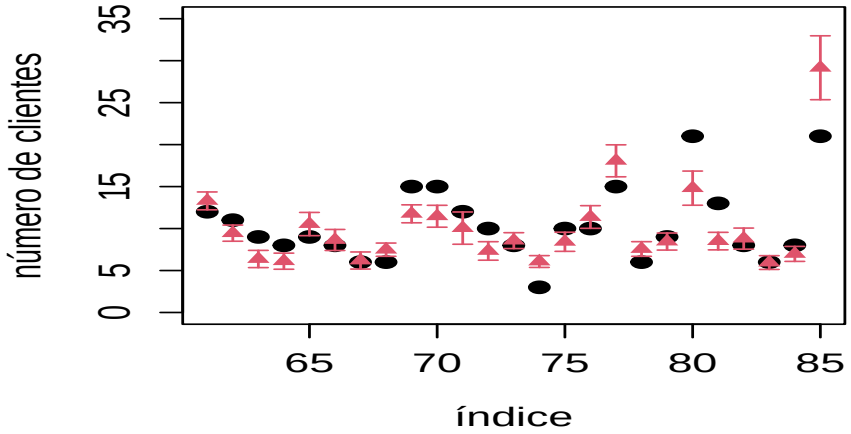
- A aplicação da metodologia stepwise, começando com o modelo só com o intercepto ou começando com o modelo completo, indicou, em ambos os casos, que todas as variáveis são significativas.
- A utilização do modelo com todas as covariáveis é preferível àquele sem a covariável idade.
- Critérios de Informação ($MAR = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widetilde{y}_i|$)

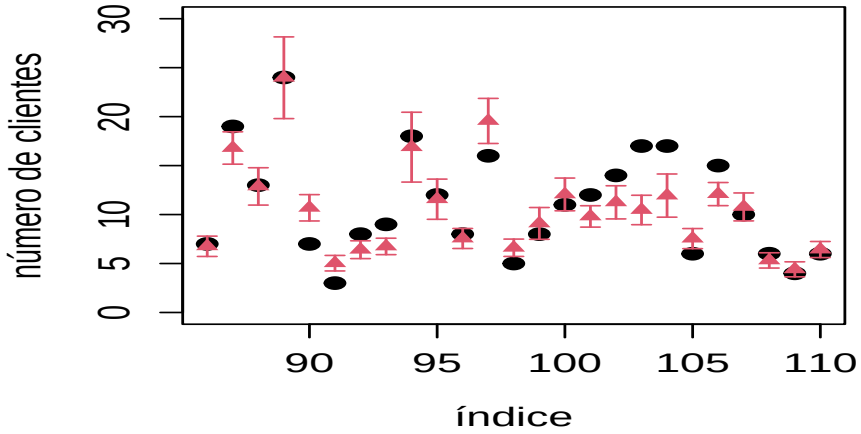
Modelo	AIC	BIC	AICc	SABIC	HQCIC	CAIC	MAR
Inicial	571,02	587,23	571,84	568,27	577,60	593,23	2,48
-Idade	573,40	586,91	573,98	571,11	578,88	591,91	2,52

Valores preditos x observados

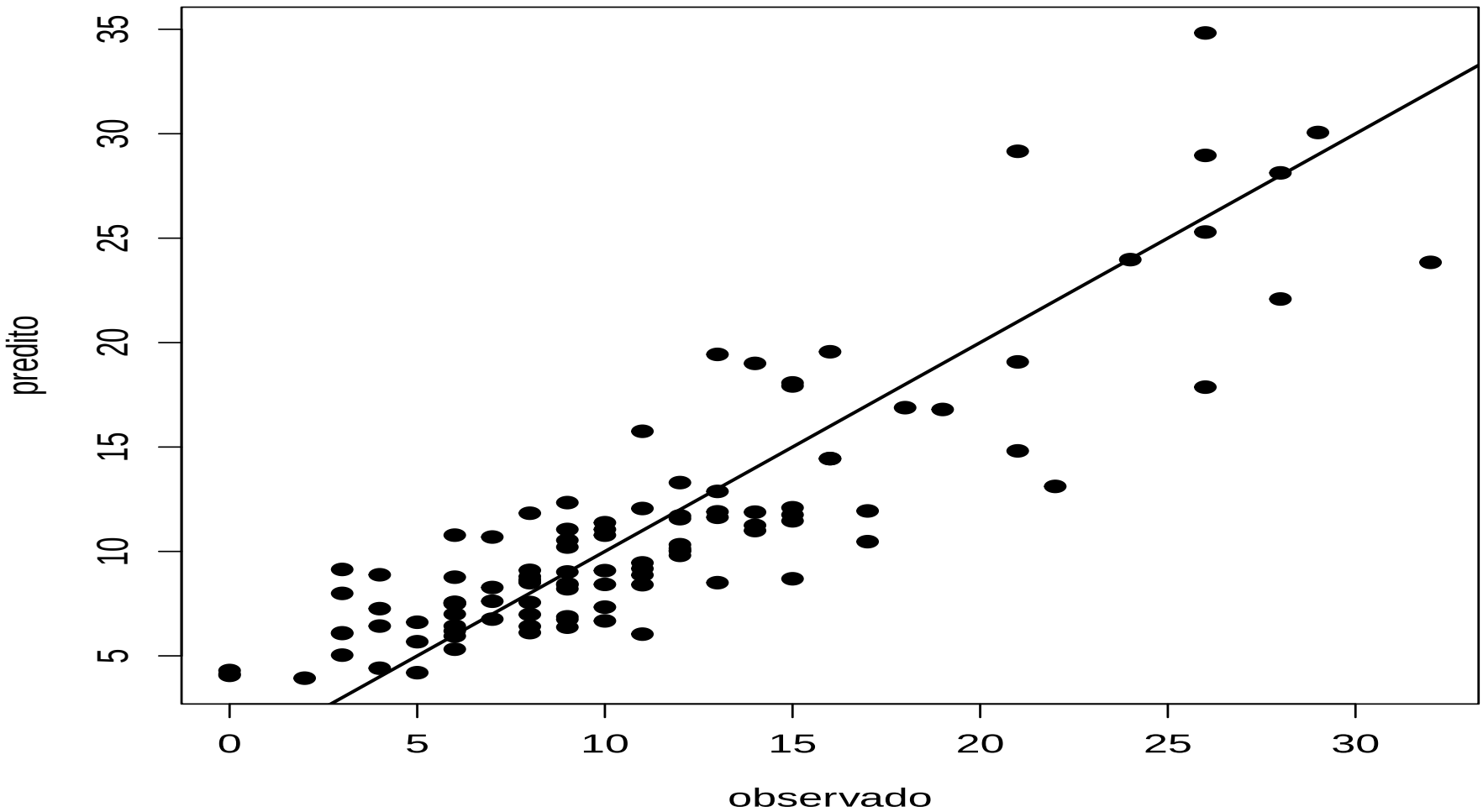




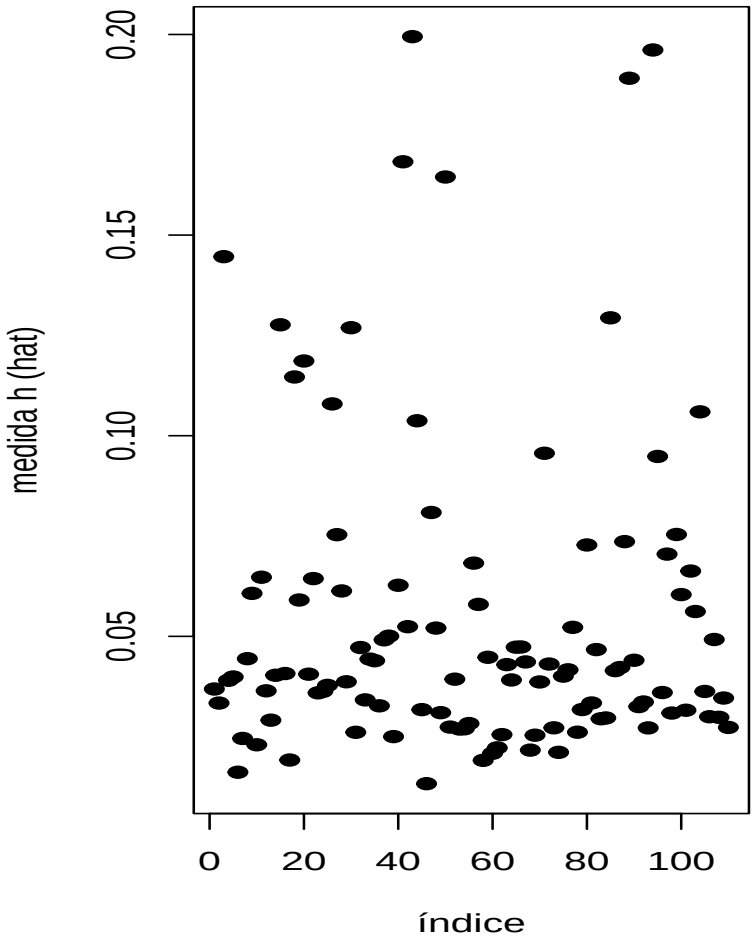


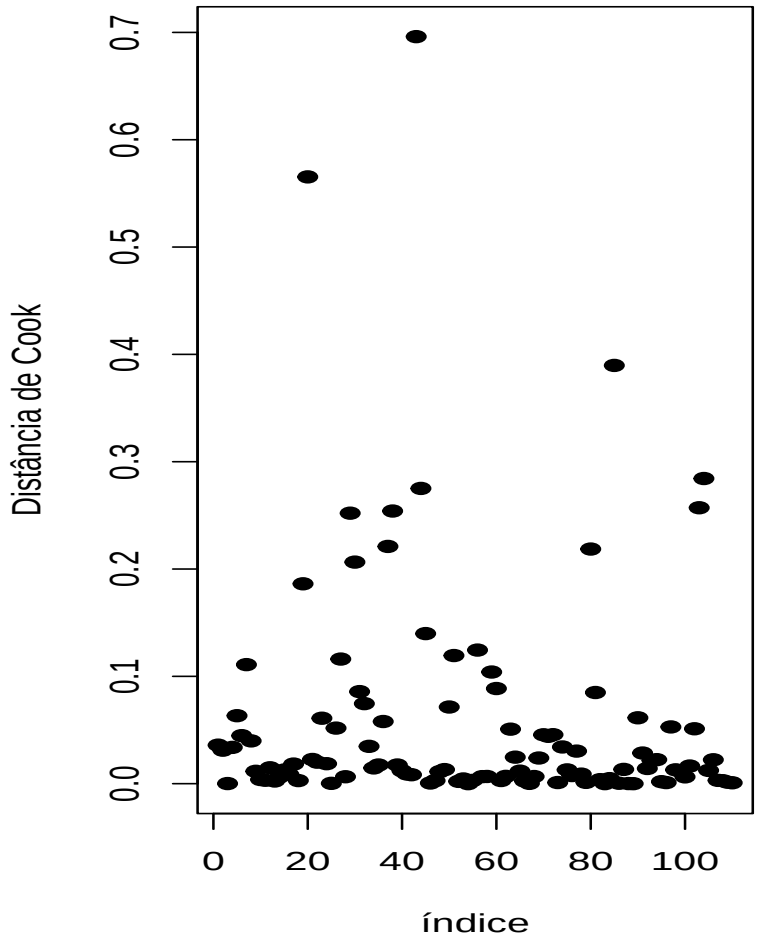


Valores preditos x observados



Pontos alavanca e Distância de Cook





Análise de sensibilidade

parâmetro (β_0)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	2,30	0,03	[2,24;2,36]	72,92	$< 0{,}0001$
-# 20	2,29	0,03	[2,23;2,36]	72,21	$< 0{,}0001$
-# 40	2,30	0,03	[2,24;2,36]	72,56	$< 0{,}0001$
-# 20,40	2,29	0,03	[2,23;2,35]	71,86	$< 0{,}0001$

Análise de sensibilidade

parâmetro (β_1)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	0,16	0,04	[0,09;0,23]	4,26	$< 0{,}0001$
-# 20	0,16	0,04	[0,09;0,24]	4,36	$< 0{,}0001$
-# 40	0,16	0,04	[0,09;0,24]	4,27	$< 0{,}0001$
-# 20,40	0,17	0,04	[0,09;0,24]	4,36	$< 0{,}0001$

Análise de sensibilidade

parâmetro (β_2)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	-0,22	0,04	[-0,29;-0,14]	-5,53	$< 0{,}0001$
-# 20	-0,23	0,04	[-0,31;-0,15]	-5,78	$< 0{,}0001$
-# 40	-0,22	0,04	[-0,29;-0,14]	-5,52	$< 0{,}0001$
-# 20,40	-0,23	0,04	[-0,31;-0,15]	-5,76	$< 0{,}0001$

Análise de sensibilidade

parâmetro (β_3)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	-0,06	0,03	[-0,12;-0,00]	-2,09	0,0365
-# 20	-0,08	0,03	[-0,14;-0,02]	-2,47	0,0136
-# 40	-0,06	0,03	[-0,12;-0,00]	-2,09	0,0362
-# 20,40	-0,08	0,03	[-0,14;-0,02]	-2,47	0,0135

Análise de sensibilidade

parâmetro (β_4)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	0,25	0,04	[0,18;0,33]	6,53	$< 0{,}0001$
-# 20	0,26	0,04	[0,18;0,34]	6,67	$< 0{,}0001$
-# 40	0,25	0,04	[0,18;0,33]	6,53	$< 0{,}0001$
-# 20,40	0,26	0,04	[0,18;0,34]	6,66	$< 0{,}0001$

Análise de sensibilidade

parâmetro (β_5)

Observações	Est.	EP	IC(95%)	Estat. Z_t	p-valor
todas	-0,30	0,04	[-0,37;-0,22]	-7,95	$< 0,0001$
-# 20	-0,29	0,04	[-0,36;-0,22]	-7,77	$< 0,0001$
-# 40	-0,29	0,04	[-0,37;-0,22]	-7,87	$< 0,0001$
-# 20,40	-0,29	0,04	[-0,36;-0,21]	-7,68	$< 0,0001$