

Modelos para dados de contagem: parte 1

Prof. Caio Azevedo

Exemplo 3: tempo de sobrevivências de bactérias

- Os dados correspondem ao número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto à uma temperatura de $300^{\circ}F$.
- Resposta: número (contagem) de bactérias sobreviventes.
- Variável explicativa: tempo de exposição.
- Nessas amostras de alimentos foram feitas 12 medições, a cada minuto, contabilizando a quantidade de bactérias vivas (do total original) sobreviventes.
- Novamente temos uma situação de medidas repetidas e, assim, as observações podem ter algum tipo de dependência.

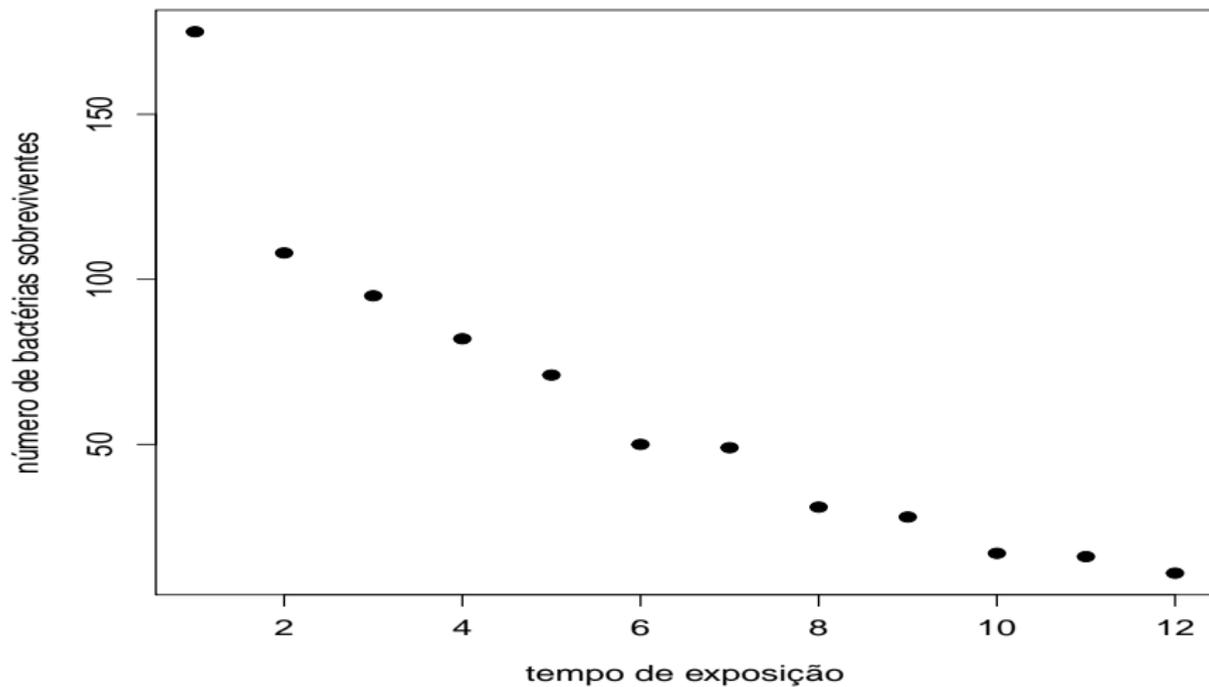
Dados oriundos do experimento

- Dados

número	175	108	95	82	71	50	49	31	28	17	16	11
tempo	1	2	3	4	5	6	7	8	9	10	11	12

- número: número de bactérias sobreviventes.
- tempo: tempo decorrido em minutos.
- Este exemplo já fora discutido anteriormente ([aqui](#))

Gráfico de dispersão



Observações

- A resposta é positiva e discreta. Assim, há possibilidade da resposta apresentar heterocedasticidade em função da variável explicativa.
- Considerar um MRNLH pode ser problemático uma vez que tal metodologia requer normalidade e homocedasticidade da resposta.

Observações

- Ademais, assumir normalidade para a resposta levará à um modelo que atribui probabilidades positivas de ocorrência a valores os quais não podem ser observados (negativos e não inteiros). Por outro lado, atribui probabilidade zero para eventos que são observados (valores pontuais).
- Além disso, a ligação identidade ($\mu_i = \mathbf{X}_i\beta$) imposta entre a média e a variável explicativa (fator) pode levar à valores negativos para as médias preditas.
- Também, a relação entre a resposta e a variável explicativa parecer ser não linear.

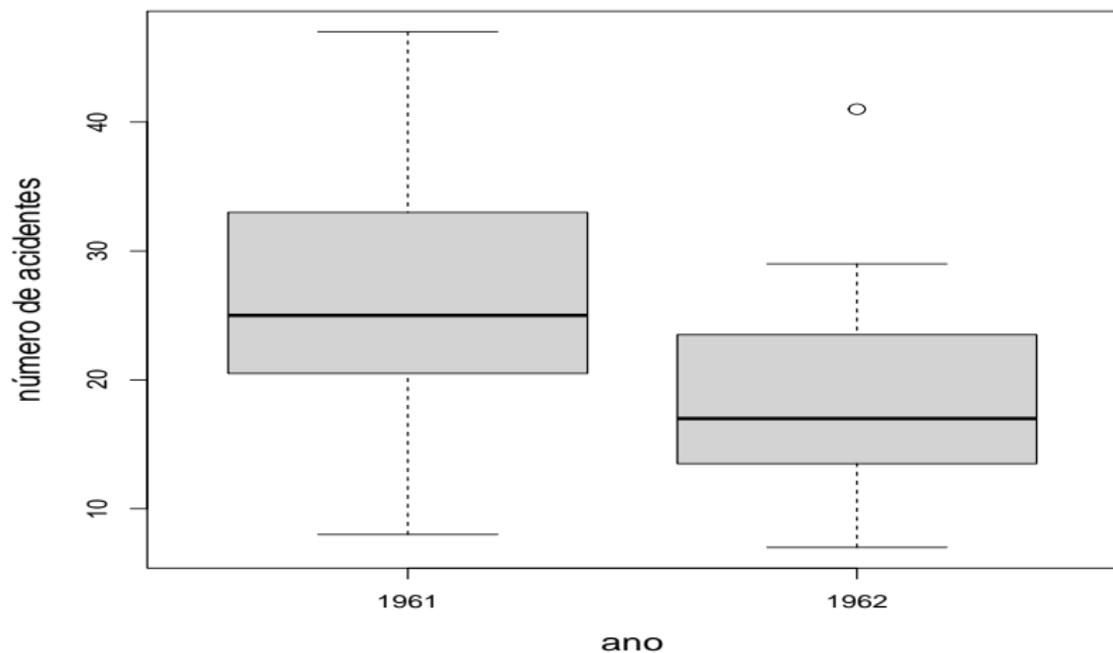
Exemplo 4: comparação do número de acidentes

- Descrição: número de acidentes (com algum tipo de trauma para as pessoas envolvidas) em 92 dias (correspondentes) em dois anos distintos (1961 e 1962), medidos em algumas regiões da Suécia.
- Considerou-se apenas 43 dias, correspondendo a dias de 1961 em que não havia limite de velocidade e de 1962 em que havia limites de velocidade (90 ou 100 km/h).

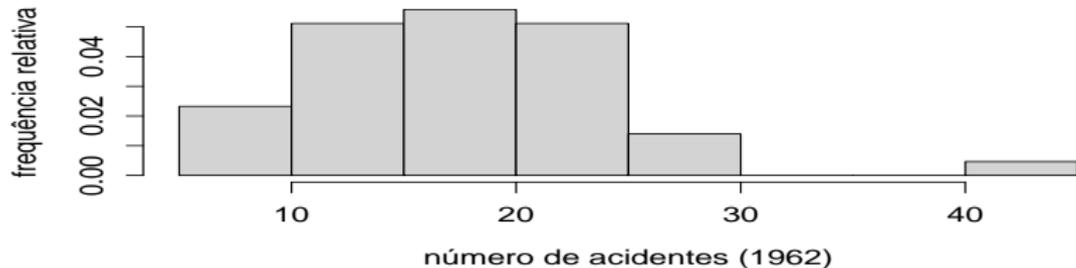
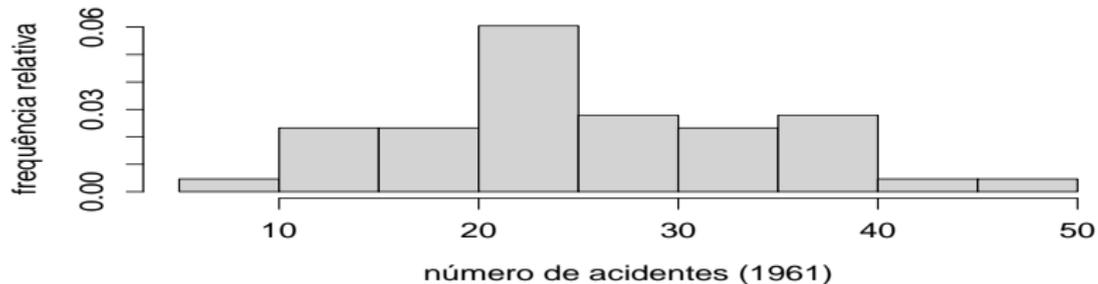
Exemplo 4: comparação do número de acidentes

- Resposta: número de acidentes.
- Variável explicativa: ano (presença/ausência de limite de velocidade).
- Questão de interesse: a imposição dos limites de velocidade levou à redução do número de acidentes?

Boxplots do número de acidentes por ano



Histogramas do número de acidentes por ano



Medidas Resumo

■ Medidas

Ano	Média	Var.	DP	CV(%)	Mín.	Med.	Máx.	CA	Curt.
1961	26,05	82,66	9,09	34,91	8,00	25,00	47,00	0,31	2,40
1962	18,05	44,71	6,69	37,05	7,00	17,00	41,00	0,90	4,43

- Aparentemente há uma superdispersão nos dados pois as variâncias amostrais (em cada grupo) são maiores do que as médias amostrais.

Distribuição de Poisson

- Função de probabilidade, se $Y \sim \text{Poisson}(\mu)$, $\mu > 0$, então

$$f_Y(y) = \frac{e^{-\mu} \mu^y}{y!} \mathbb{1}_{\{0,1,2,\dots\}}(y).$$

- Note que $\mathcal{E}(Y) = \mathcal{V}(Y) = \mu$.
- Há outras distribuições para contagem (sem limite superior) como a distribuição binomial negativa.
- Note que este tipo de contagem é diferente daquele representado pela distribuição binomial (a qual é limitada).

Modelo geral

$$Y_i \stackrel{ind.}{\sim} \text{Poisson}(\mu_i)$$
$$g(\mu_i) = \sum_{j=1}^p \beta_j x_{ji} \rightarrow \mu_i = g^{-1} \left(\sum_{j=1}^p \beta_j x_{ji} \right), i = 1, 2, \dots, n, j = 1, 2, \dots, p$$

- Y_i : contagem de interesse da i -ésima observação.
- x_{ji} : valor da variável explicativa j associada ao indivíduo i ;
- β_j : parâmetro associado ao impacto de cada covariável na média da supracitada contagem.
- $g(\cdot)$: função de ligação. Escolha usual $g(\cdot) = \ln(\cdot)$ (função de ligação canônica). Outras possibilidades: $\sqrt{(\cdot)}, 1/(\cdot)$
- Modelo com intercepto: $x_{1i} = 1, \forall i$.

Verificação da qualidade de ajuste do modelo

- No modelo em questão, temos, essencialmente, as seguintes suposições a serem avaliadas.
 - Apesar do modelo ser heterocedástico ($\mathcal{V}(Y_i) = \mu_i$), a variância observada pode ser menor do que a imposta pelo modelo (subdispersão) ou maior do (superdispersão).
 - As observações são independentes.
 - A função de ligação é apropriada.
 - A estrutura do preditor linear é adequada.

Estimação por MV: ligação canônica

- Lembremos que, no caso da distribuição de Poisson, $\phi = 1$.
- Lembremos que, para esse modelo, a função de ligação logaritmica, $\ln(\mu_i)$ é a função de ligação canônica.
- Função escore: $\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu})$, em que $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ e $\mu_i = e^{\eta_i}$.
- Informação de Fisher: $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$, $\mathbf{V} = \text{diag}(V_1, \dots, V_n)$, $V_i = \mu_i, i = 1, 2, \dots, n$.

Estimação por MV: ligação geral

- Função escore: $\mathbf{S}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \mathbf{X}' \mathbf{W}^{1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu})$, em que
 $\mu_i = g^{-1}(\eta_i)$, $\mathbf{W} = \text{diag}(\omega_1, \dots, \omega_n)$ e
 $\omega_i = \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 V_i^{-1} = \left(\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \right)^2 \mu_i^{-1}$, $i = 1, 2, \dots, n$.
- Informação de Fisher: $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}' \mathbf{W} \mathbf{X}$.
- Reveja os desenvolvimentos para os MLG [aqui](#) e [aqui](#).

Inferência por MV

- Dessa forma, devemos utilizar o processo iterativo (algoritmo Escore de Fisher), apresentado anteriormente [aqui](#), para obtermos estimativas para β .
- As formas do desvio e do RCD para o modelo Poisson já foram vistas anteriormente, respectivamente: slides 26 [aqui](#) e [aqui](#).
- Neste caso, a aproximação do desvio pela distribuição de qui-quadrado é apropriada quando $\mathcal{E}(Y_i) = \mu_i \rightarrow \infty, \forall i$.

Exemplo 3: Modelo 1

$$Y_i \stackrel{\text{ind.}}{\sim} \text{Poisson}(\mu_i)$$

$$\ln(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}) \rightarrow \mu_i = e^{\beta_0 + \beta_1(x_i - \bar{x})}, i = 1, 2, \dots, 12$$

- Y_i : número de bactérias sobreviventes no instante i .
- x_i : tempo de exposição no instante i , $\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = 6,5$.
- e^{β_0} : número esperado de bactérias sobreviventes no minuto 6,5.
- e^{β_1} : incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição aumenta em um minuto.

Exemplo 3: Modelo 2

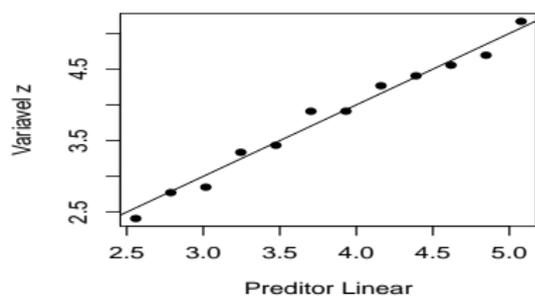
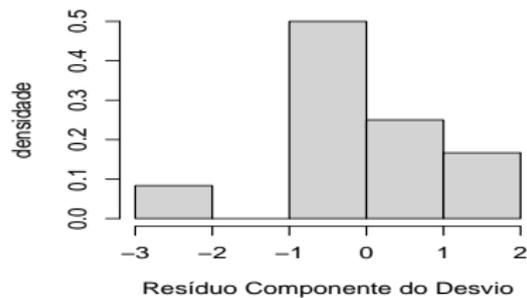
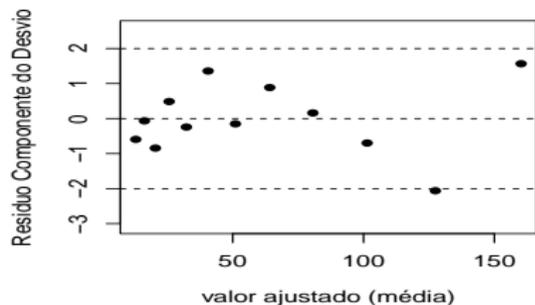
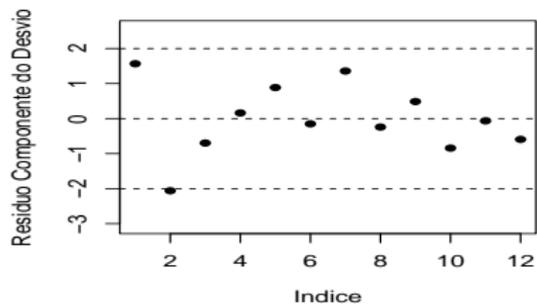
$$Y_i \stackrel{ind.}{\sim} \text{Poisson}(\mu_i)$$

$$\ln(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 \rightarrow$$

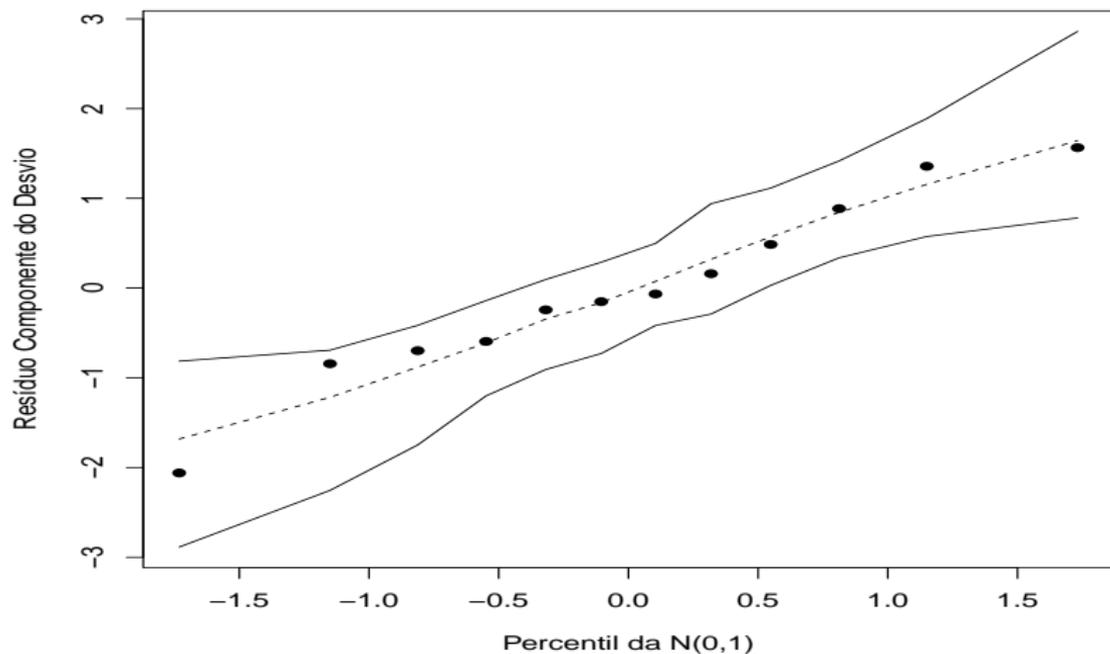
$$\mu_i = e^{\beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2}, i = 1, 2, \dots, 12$$

- Y_i : número de bactérias sobreviventes no instante i .
- x_i : tempo de exposição no instante i , $\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = 6,5$.
- e^{β_0} : número esperado de bactérias sobreviventes no minuto 6,5.
- $-\frac{\beta_1}{2\beta_2}$: minutos necessários para que o número de bactérias sobreviventes seja mínimo.

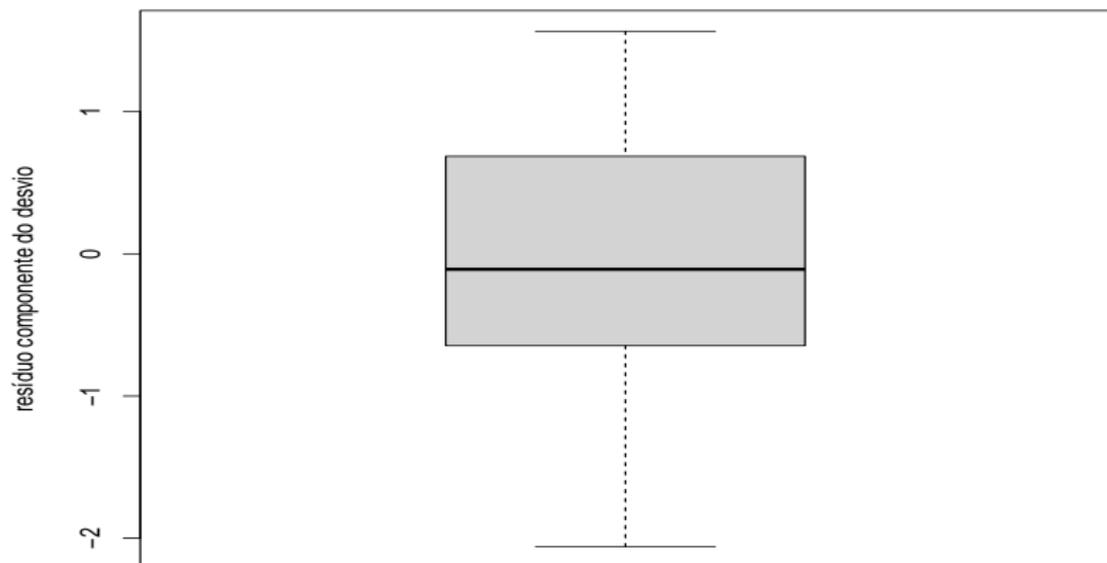
Gráficos de diagnóstico: Modelo 1



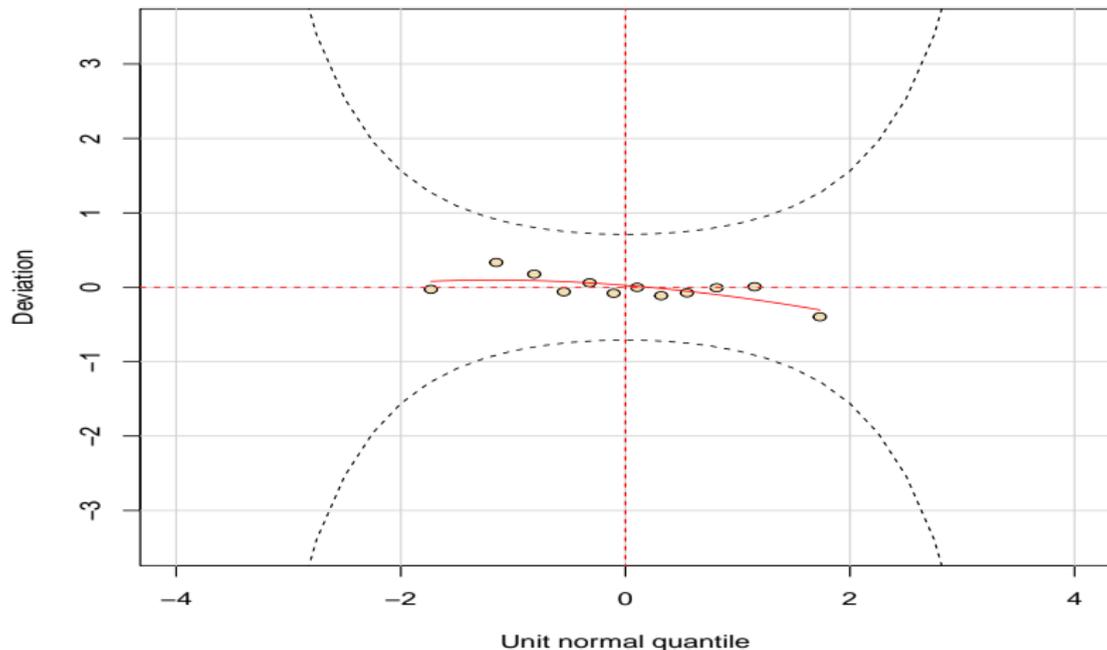
Envelope para os resíduos: Modelo 1



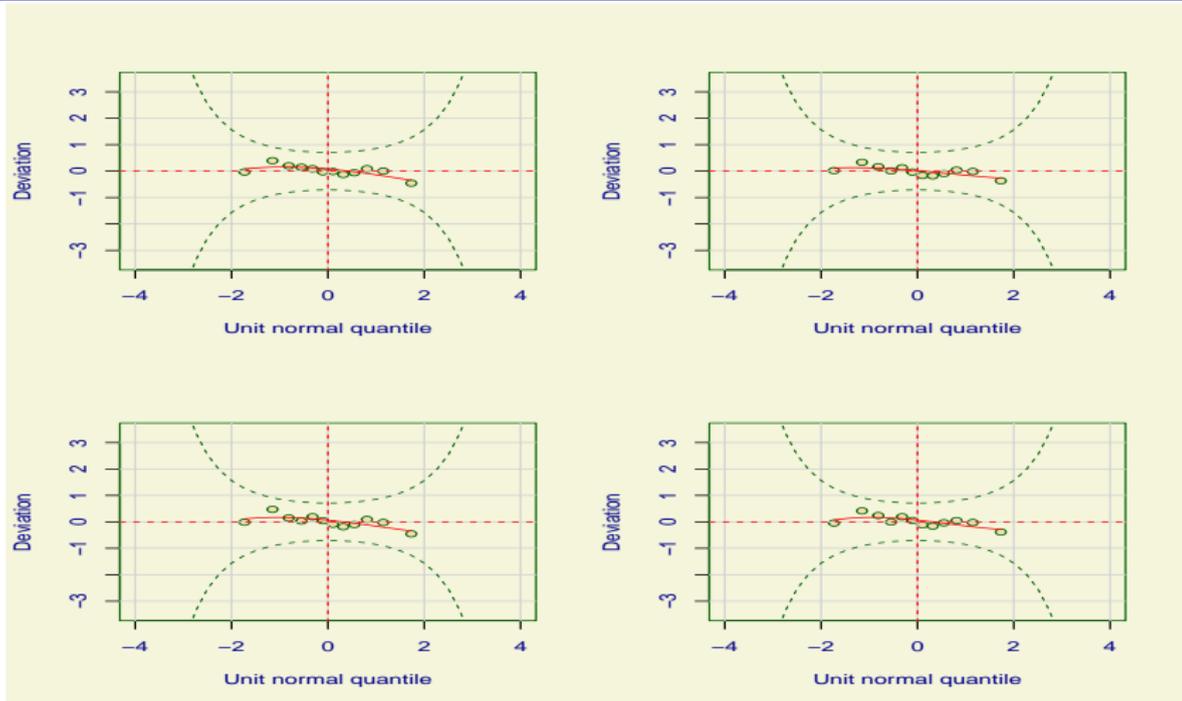
Box plot para os resíduos: Modelo 1



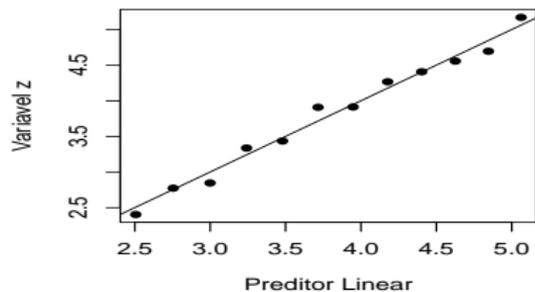
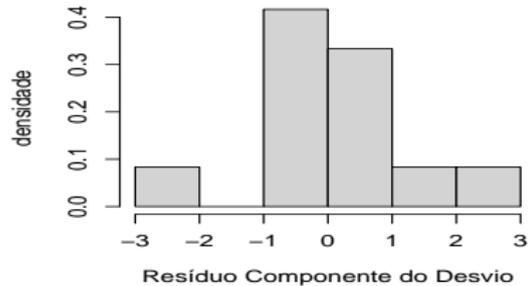
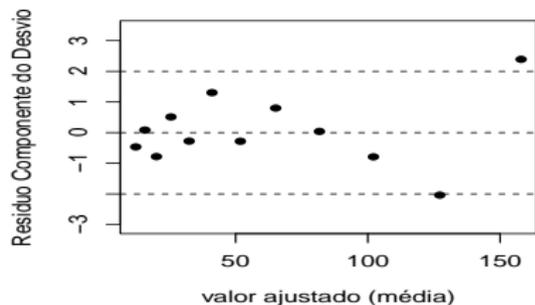
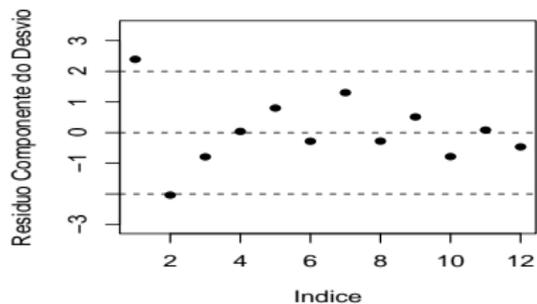
Worm plot para os resíduos: Modelo 1



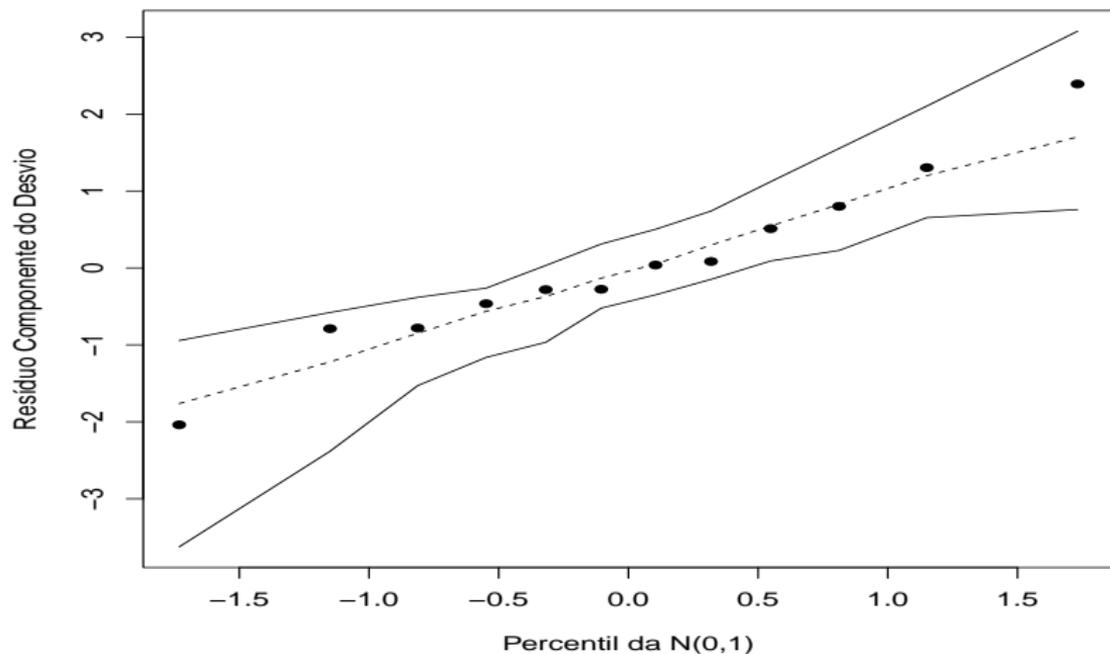
Worm plots para os resíduos: Modelo 1



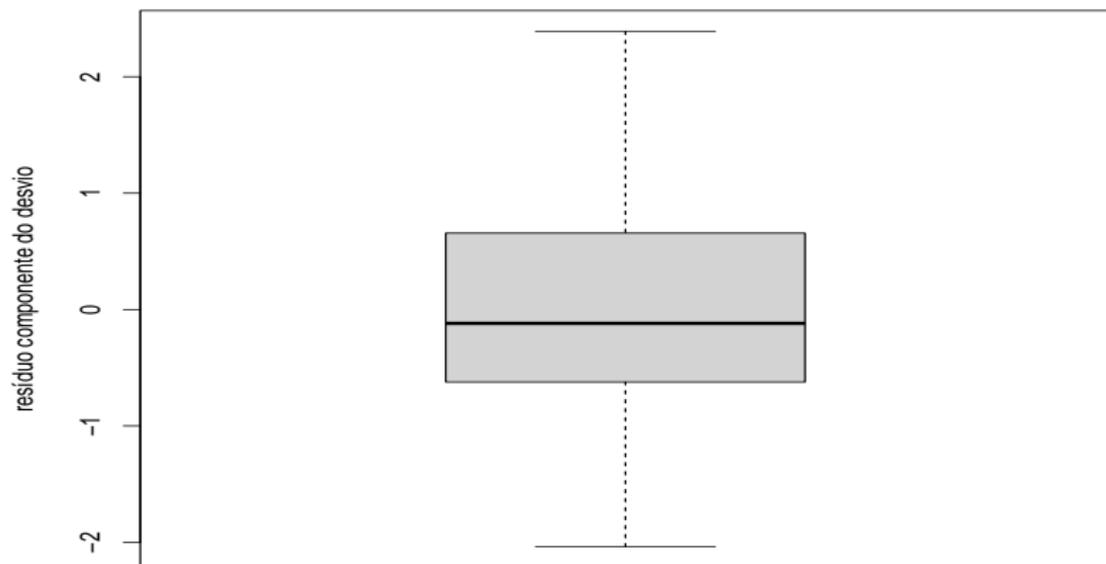
Gráficos de diagnóstico: Modelo 2



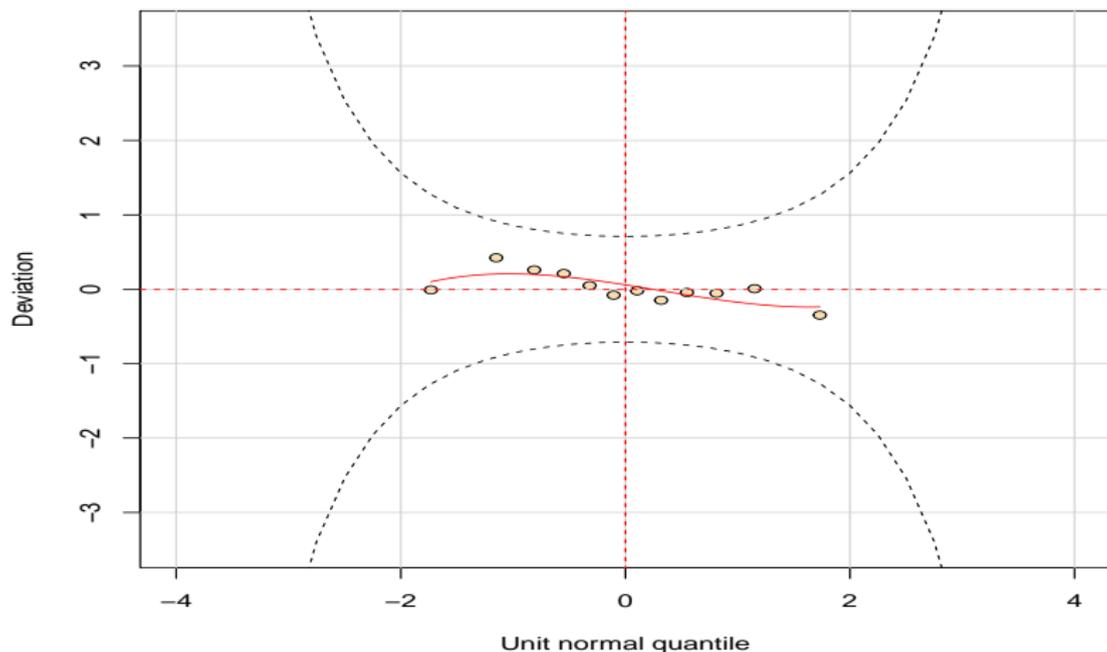
Envelope para os resíduos: Modelo 2



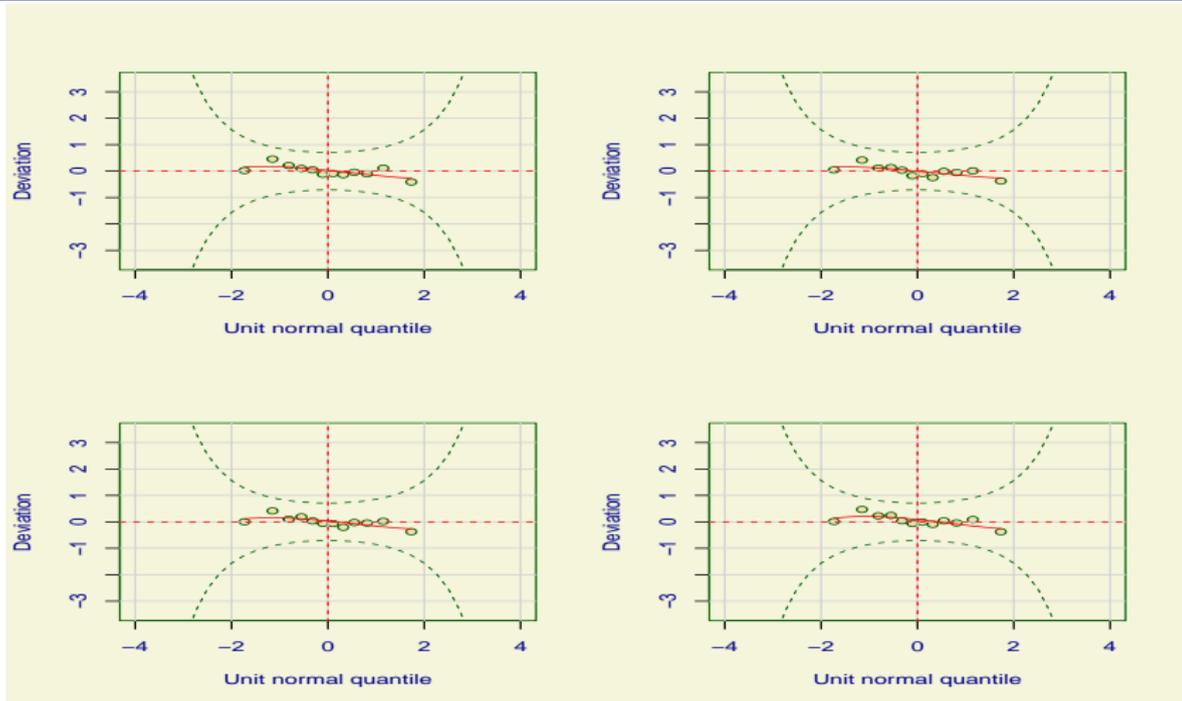
Box plot para os resíduos: Modelo 2



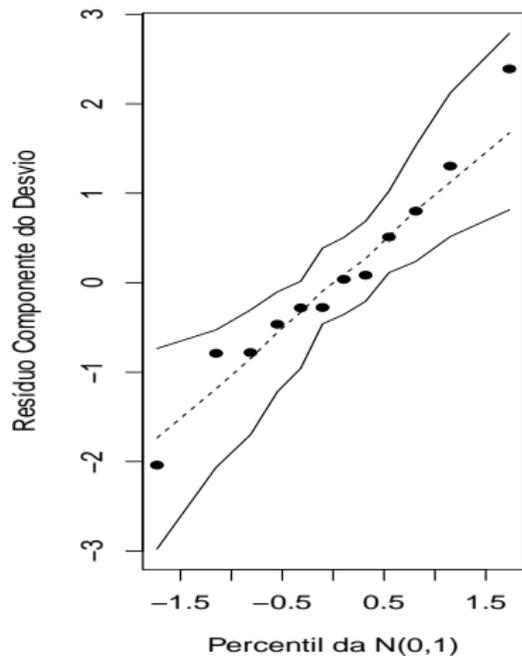
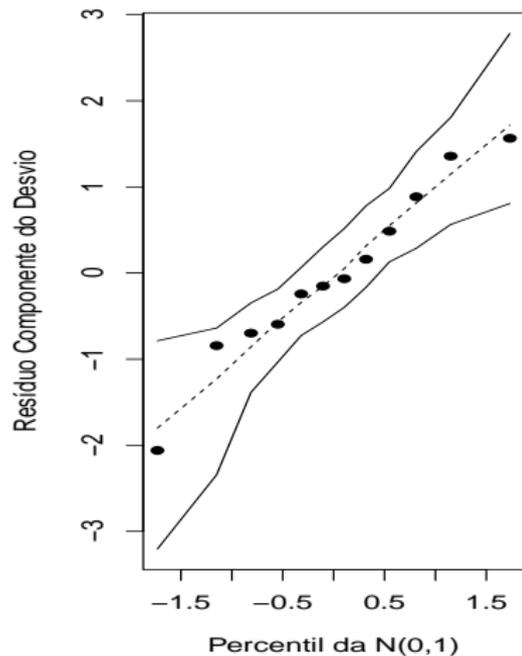
Worm plot para os resíduos: Modelo 2



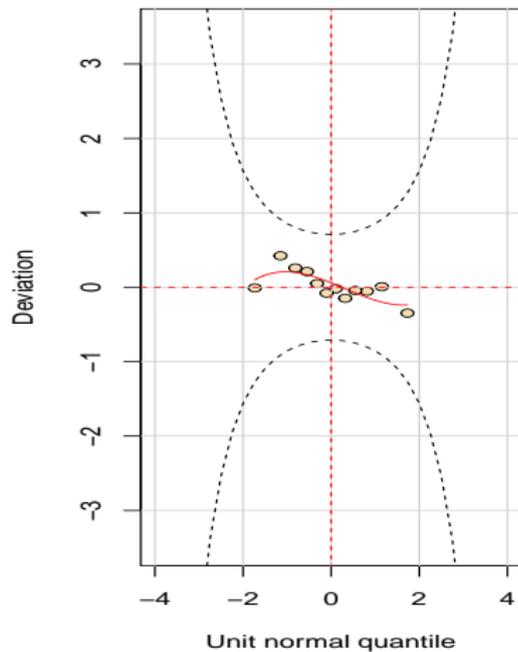
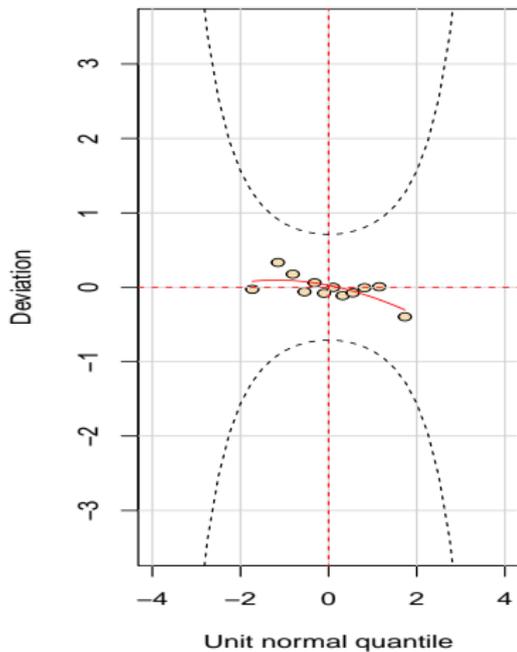
Worm plots para os resíduos: Modelo 2



Envelope para os resíduos: Modelos 1 e 2

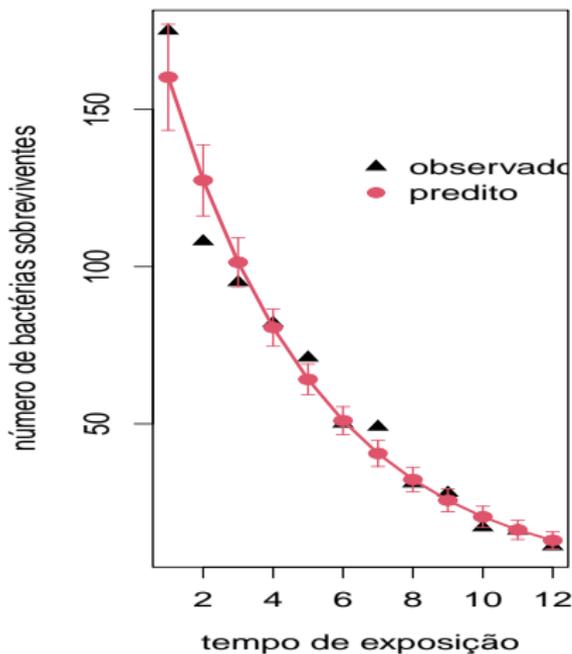


Worm plot para os resíduos: Modelos 1 e 2

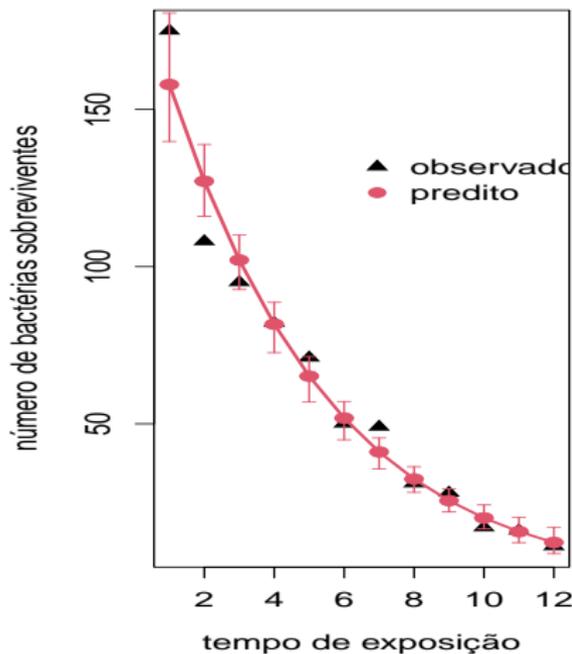


Valores observados e preditos : modelos 1 e 2

Modelo 1



Modelo 2



Critérios de Informação e de comparação

M	AIC	BIC	AICc	SABIC	HQCIC	CAIC	p	RAM	RPM
1	80,18	81,15	81,52	75,10	79,82	83,15	0,5877	67,44	-0,25
2	82,04	83,49	85,04	74,42	81,50	86,49	0,5067	67,73	-0,18

■ M: modelo.

■ p: p-valor do desvio.

■ RAM (Resíduo absoluto médio) = $\frac{1}{n} \sum_{i=1}^n |y_i - \tilde{y}_i|$.

■ RPM (Resíduo percentual médio) = $\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \tilde{y}_i}{y_i} \right)$.

Estimativas dos parâmetros dos modelos

Mod.	Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
1	β_0	3,82	0,05	[3,72 ; 3,91]	79,26	< 0,0001
	β_1	-0,23	0,01	[-0,25 ; -0,20]	-18,02	< 0,0001
2	β_0	3,83	0,06	[3,71 ; 3,95]	63,14	< 0,0001
	β_1	-0,23	0,02	[-0,26 ; -0,20]	-14,80	< 0,0001
	β_2	-0,0016	0,0041	[-0,0097 ; 0,0065]	-0,3818	0,7026

Comentários

- Ambos os modelos se ajustaram bem aos dados, tanto em relação aos resíduos, bem como em relação à predição.
- O modelo 1 é preferível, particularmente, no que diz respeito aos critérios de informação, predição e significância dos parâmetros.
- Via modelo 1, a estimativa da taxa de decaimento do número de bactérias (e^{β_1}) (com intervalos de confiança de 95% entre parênteses) é : 0,80 ([0,78;0,82]).

Exemplo 3: Modelo 3 (regressão segmentada)

$$Y_i \stackrel{ind.}{\sim} \text{Poisson}(\mu_i)$$

$$\ln(\mu_i) = (\beta_{01} + \beta_{11}(x_i - \bar{x}))\mathbb{1}_{\{1,2,3\}}(x_i)$$

$$+ (\beta_{02} + \beta_{12}(x_i - \bar{x}))\mathbb{1}_{\{4,5,\dots,12\}}(x_i) \rightarrow$$

$$\mu_i = e^{(\beta_{01} + \beta_{11}(x_i - \bar{x}))\mathbb{1}_{\{1,2,3\}}(x_i) + (\beta_{02} + \beta_{12}(x_i - \bar{x}))\mathbb{1}_{\{4,5,\dots,12\}}(x_i)}, i = 1, 2, \dots, 12$$

- Y_i : número de bactérias sobreviventes no instante i .
- x_i : tempo de exposição no instante i , $\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = 6,5$.
- $e^{\beta_{02}}$: número esperado de bactérias sobreviventes no minuto 6,5.

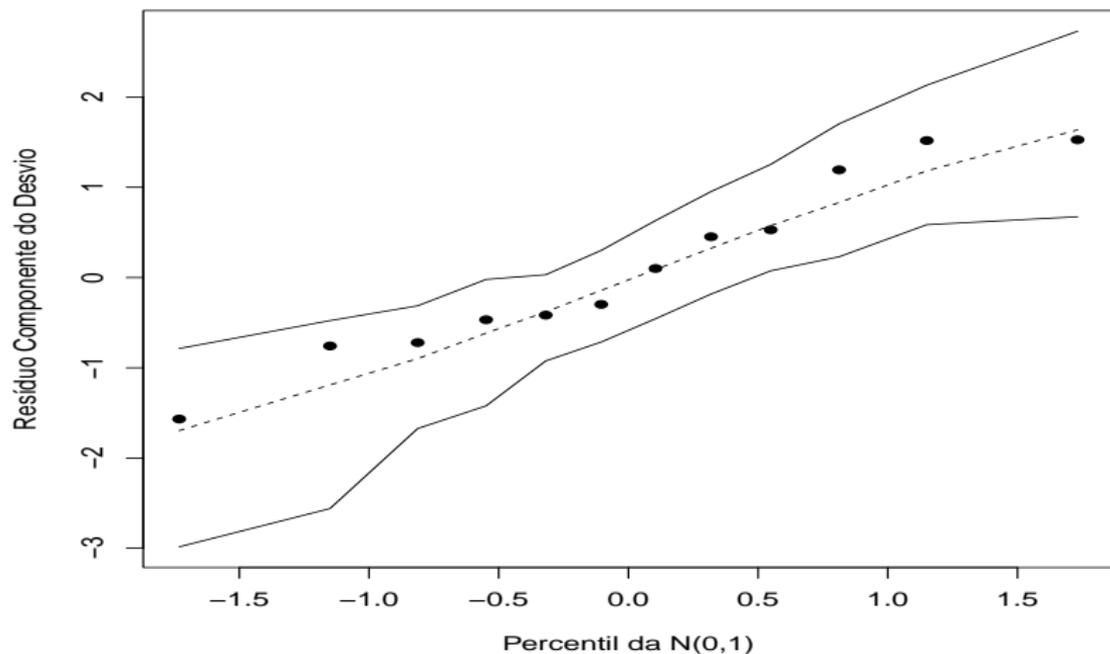
Exemplo 3: Modelo 3 (regressão segmentada)

- $e^{\beta_{11}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição aumenta em um minuto, no primeiro intervalo $\{1, 2, 3\}$.
- $e^{\beta_{12}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição aumenta em um minuto, no segundo intervalo $\{4, 5, \dots, 12\}$.

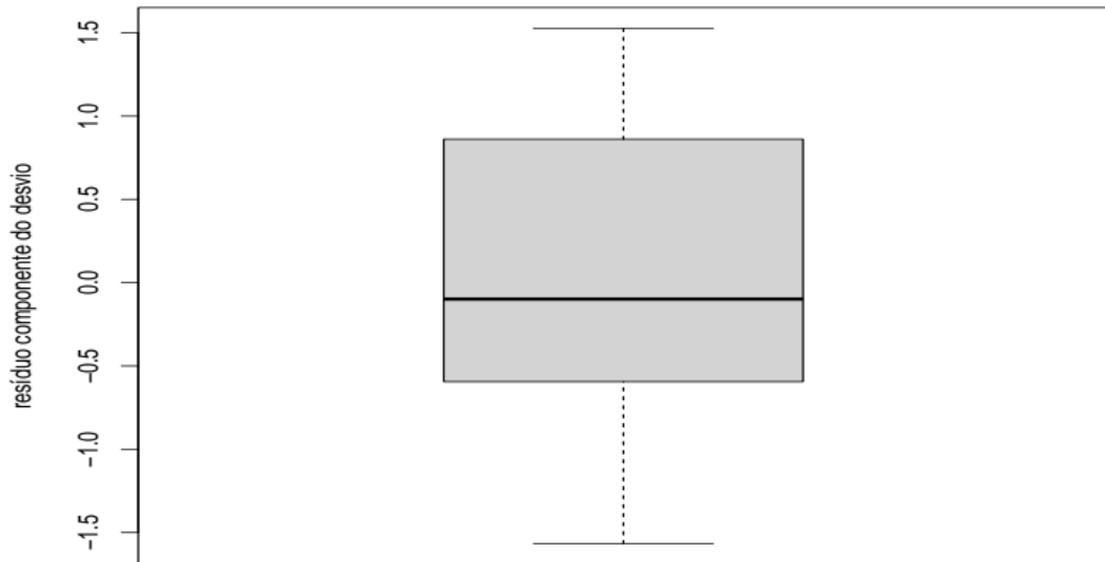
Modelo 3: matriz de planejamento

$$X = \begin{bmatrix} 1 & -5,50 & 0 & 0 \\ 1 & -4,50 & 0 & 0 \\ 1 & -3,50 & 0 & 0 \\ 0 & 0 & 1 & -2,50 \\ 0 & 0 & 1 & -1,50 \\ 0 & 0 & 1 & -0,50 \\ 0 & 0 & 1 & 0,50 \\ 0 & 0 & 1 & 1,50 \\ 0 & 0 & 1 & 2,50 \\ 0 & 0 & 1 & 3,50 \\ 0 & 0 & 1 & 4,50 \\ 0 & 0 & 1 & 5,50 \end{bmatrix}$$

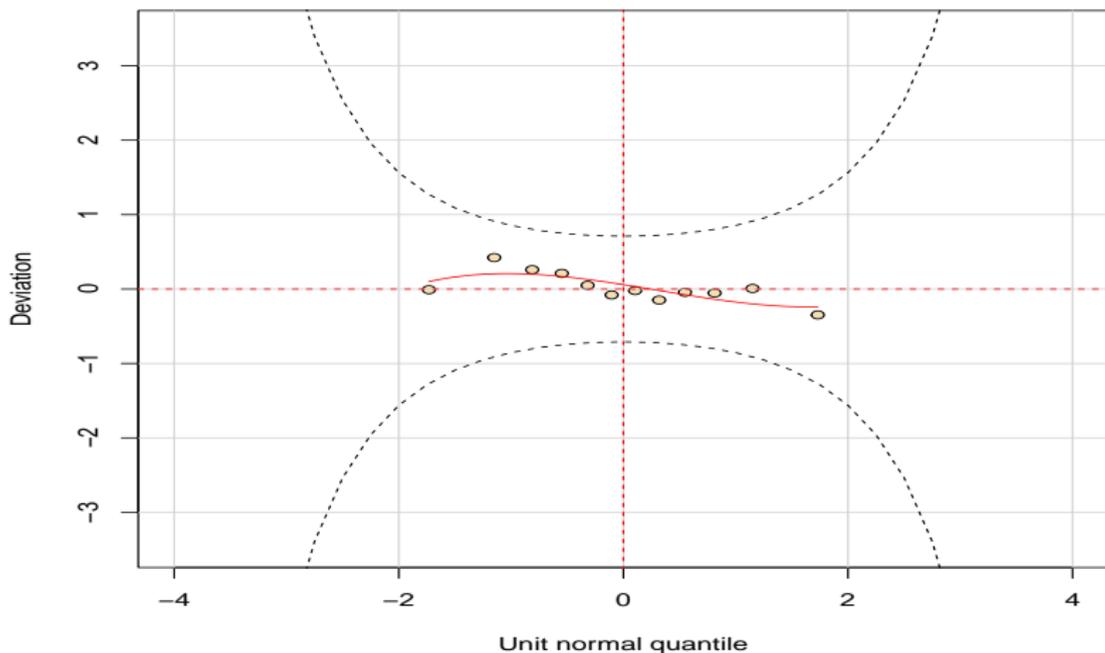
Envelope para os resíduos: Modelo 3



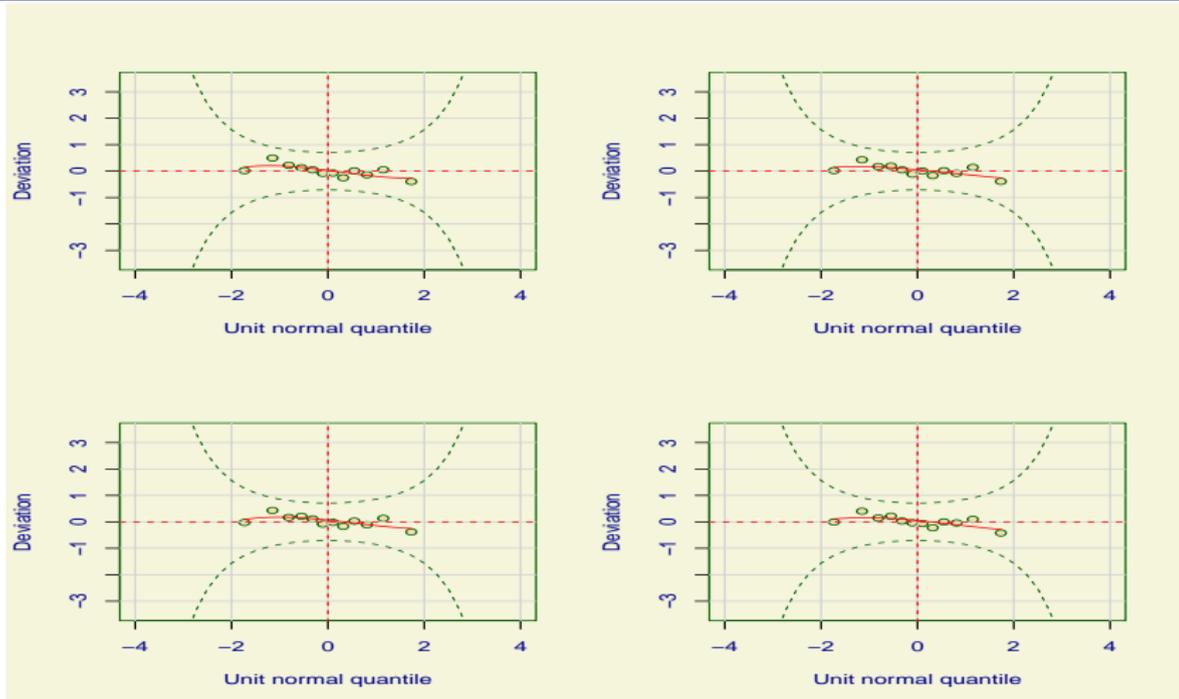
Box plot para os resíduos: Modelo 3



Worm plot para os resíduos: Modelo 3



Worm plots para os resíduos: Modelo 3



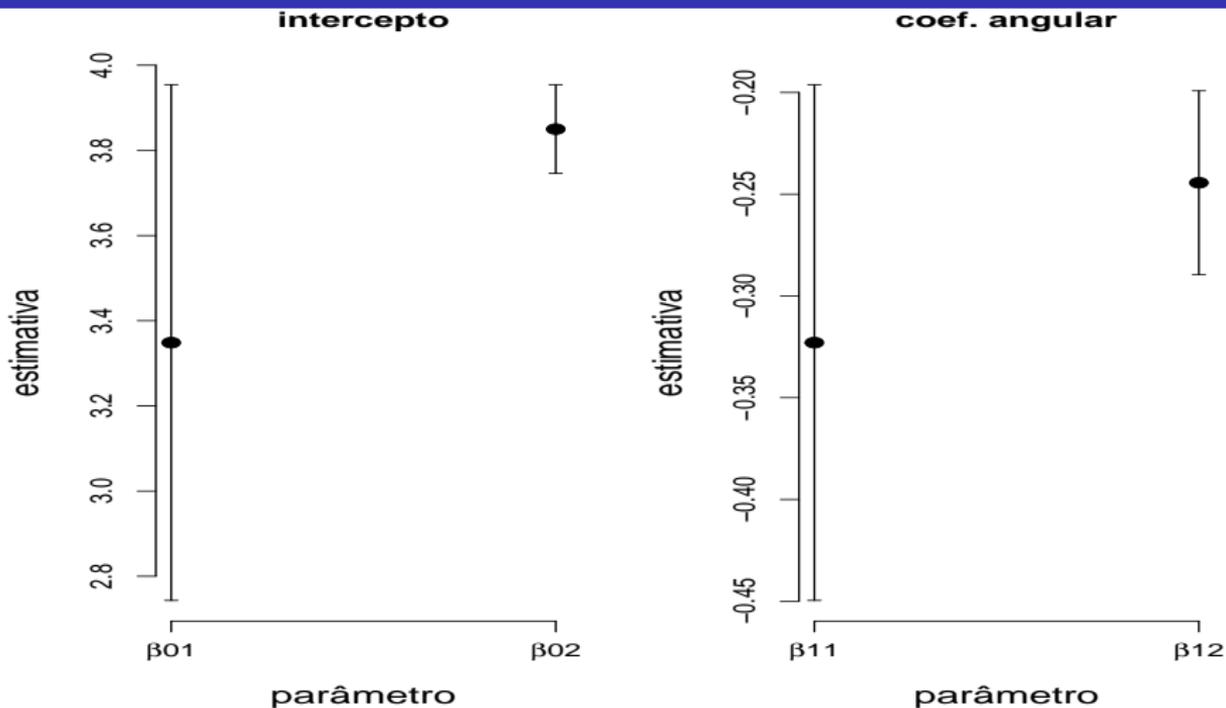
Critérios de informação e medidas de comparação

M	AIC	BIC	AICc	SABIC	HQCIC	CAIC	p	RAM	RPM
1	80,18	81,15	81,52	75,10	79,82	83,15	0,5877	67,44	-0,25
2	82,04	83,49	85,04	74,42	81,50	86,49	0,5067	67,73	-0,18
3	80,92	82,86	86,63	70,76	80,20	86,86	0,7406	54,39	-0,17

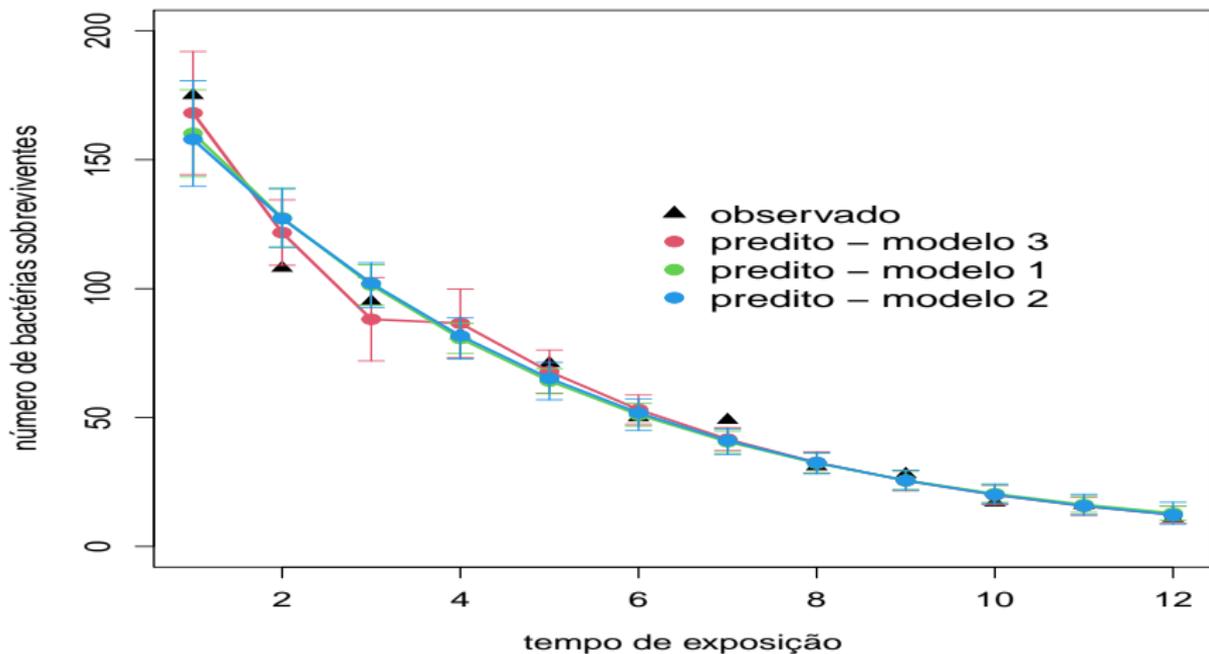
Estimativas dos parâmetros do modelo 3

Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
β_{01}	3,35	0,31	[2,74 ; 3,95]	10,84	< 0,0001
β_{11}	-0,32	0,06	[-0,45 ; -0,20]	-5,00	< 0,0001
β_{02}	3,85	0,05	[3,75 ; 3,95]	72,54	< 0,0001
β_{12}	-0,24	0,02	[-0,29 ; -0,20]	-10,60	< 0,0001

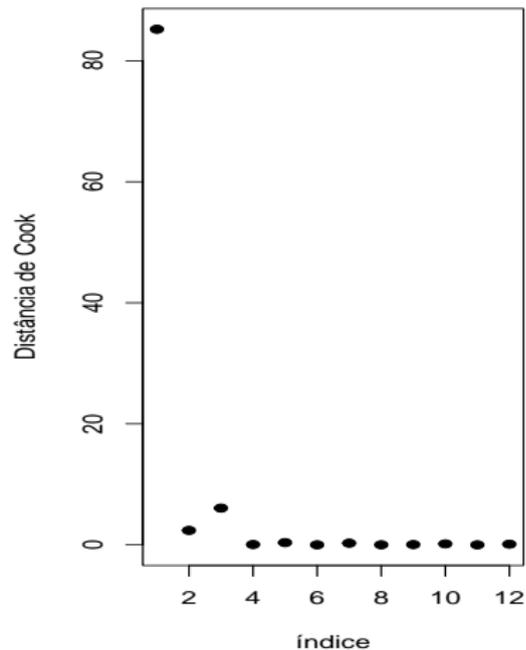
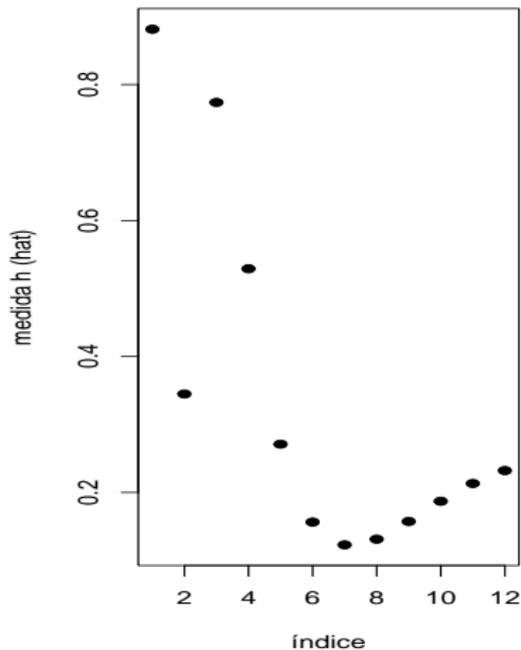
Estimativas pontuais e intervalares: modelo 3



Valores observados e preditos : modelos 1, 2 e 3



Pontos alavanca e Distância de Cook



Análise de sensibilidade

Parâmetro (β_0)

Observações	Est.	EP	IC(95%)	z value	p-valor
todas	3,82	0,05	[3,72;3,91]	79,26	< 0,0001
#-1	3,81	0,05	[3,72;3,91]	79,54	< 0,0001
#-2	3,83	0,05	[3,74;3,93]	78,76	< 0,0001
#-3	3,82	0,05	[3,73;3,92]	77,83	< 0,0001
#-4	3,82	0,05	[3,72;3,91]	76,50	< 0,0001

Análise de sensibilidade

Parâmetro (β_1)

Observações	Est.	EP	IC(95%)	z value	p-valor
todas	-0,23	0,01	[-0,25;-0,20]	-18,02	< 0,0001
#-1	-0,22	0,02	[-0,25;-0,19]	-14,11	< 0,0001
#-2	-0,24	0,01	[-0,26;-0,21]	-17,81	< 0,0001
#-3	-0,23	0,01	[-0,26;-0,21]	-17,99	< 0,0001
#-4	-0,23	0,01	[-0,25;-0,20]	-18,00	< 0,0001

Comentários

- Os modelos 1 e 3 foram os melhores, com ajustes muito próximos e ambos se mostraram ajustados aos dados. Por simplicidade, escolheremos o modelo 1.
- Há uma redução esperada de 20% em relação ao número de bactérias para cada minuto exposto.
- Alternativa para melhorar: utilizar um preditor não linear.
- Nenhum dos pontos candidatos a alavanca/influentes levaram a mudanças significativas em termos inferenciais.