Modelos para dados de contagem: parte 1

Prof. Caio Azevedo

Exemplo 3: tempo de sobrevivências de bactérias

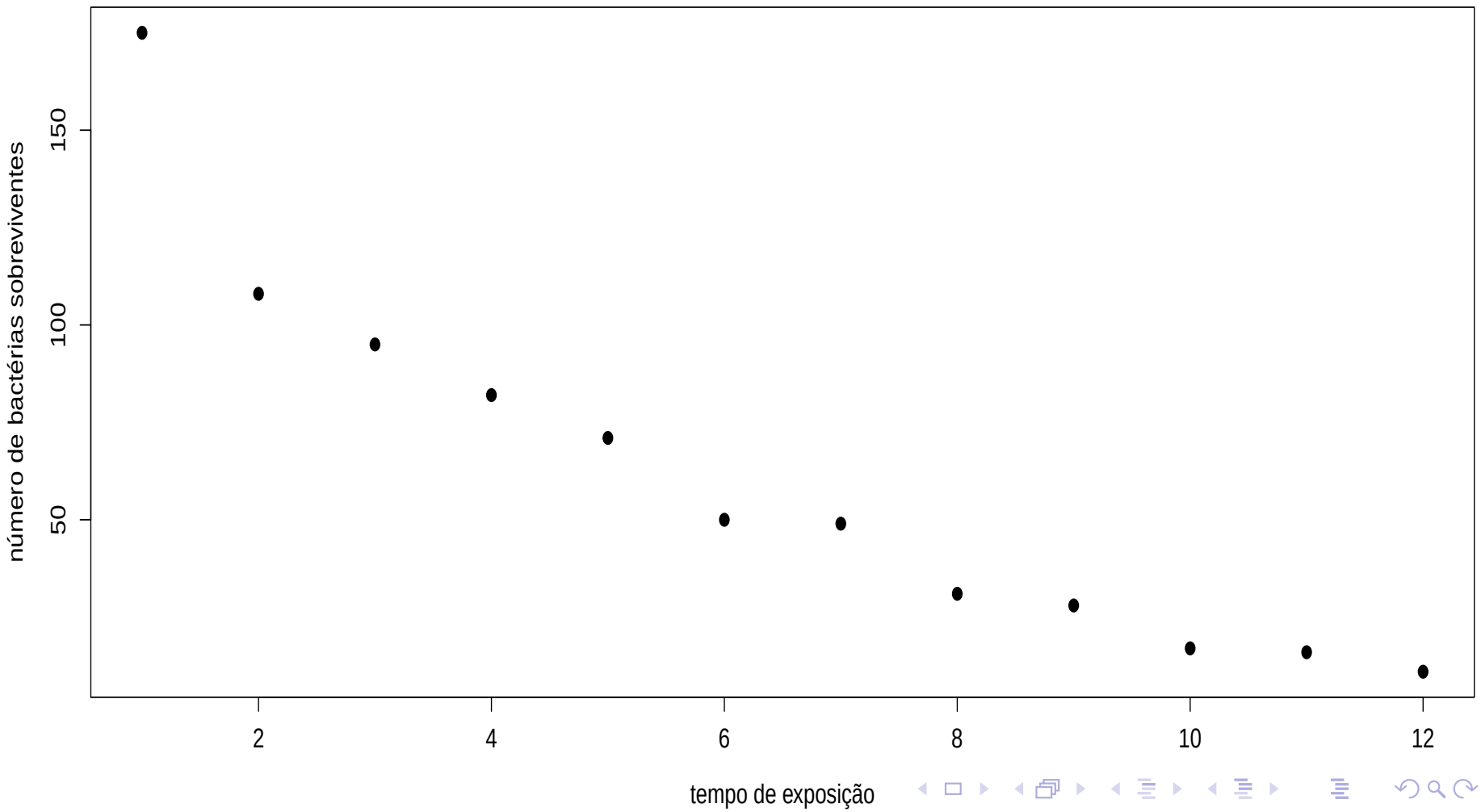
- Os dados correspondem ao número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto à uma temperatura de $300^o F$.
- Resposta: número (contagem) de bactérias sobreviventes.
- Variável explicativa: tempo de exposição.
- Nessas amostras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

Dados oriundos do experimento

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.

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 ($\mu_i = \boldsymbol{X}_i\boldsymbol{\beta}$) 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.

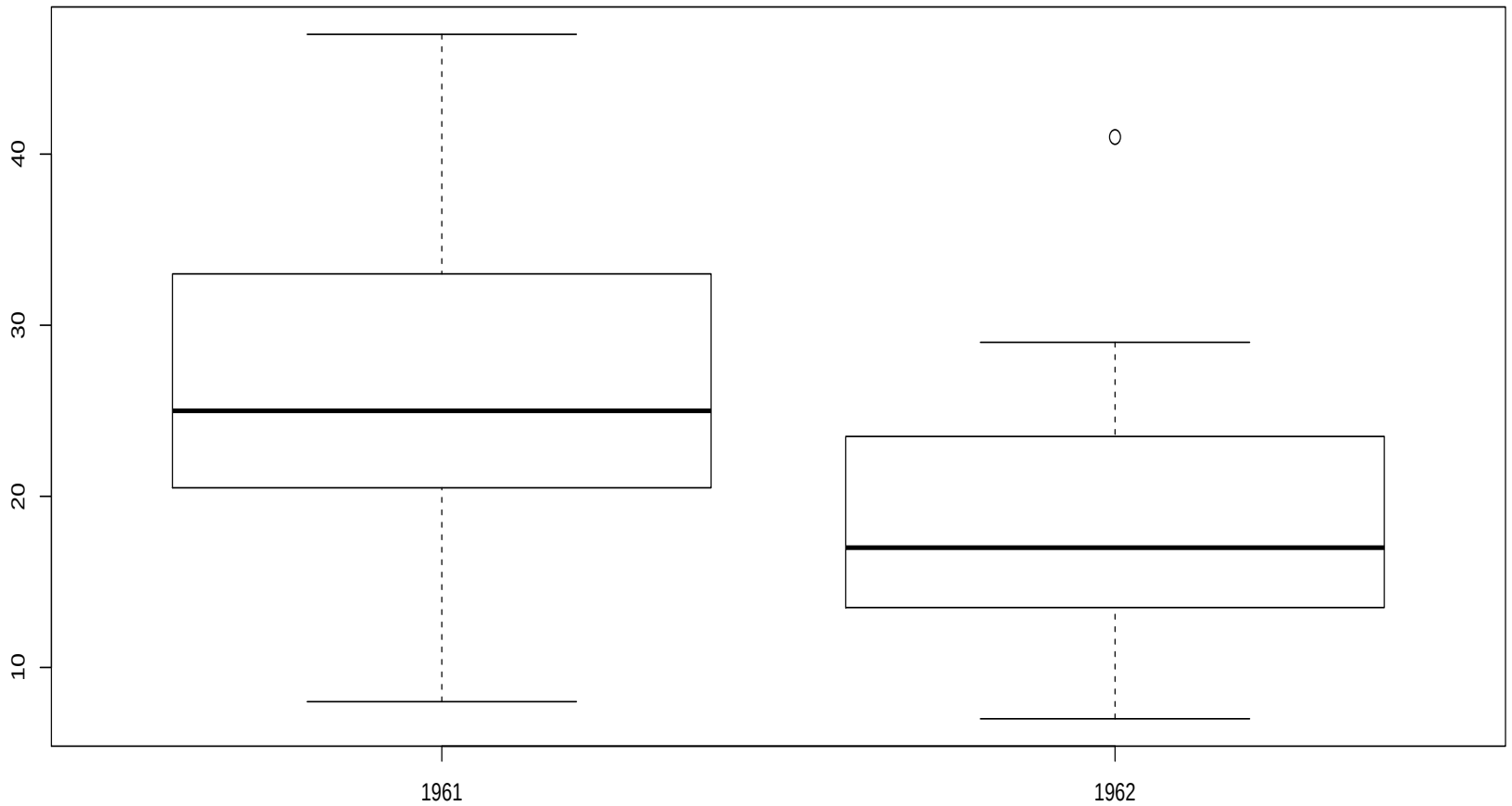
Exemplo 4: comparação do número de acidentes

- 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).

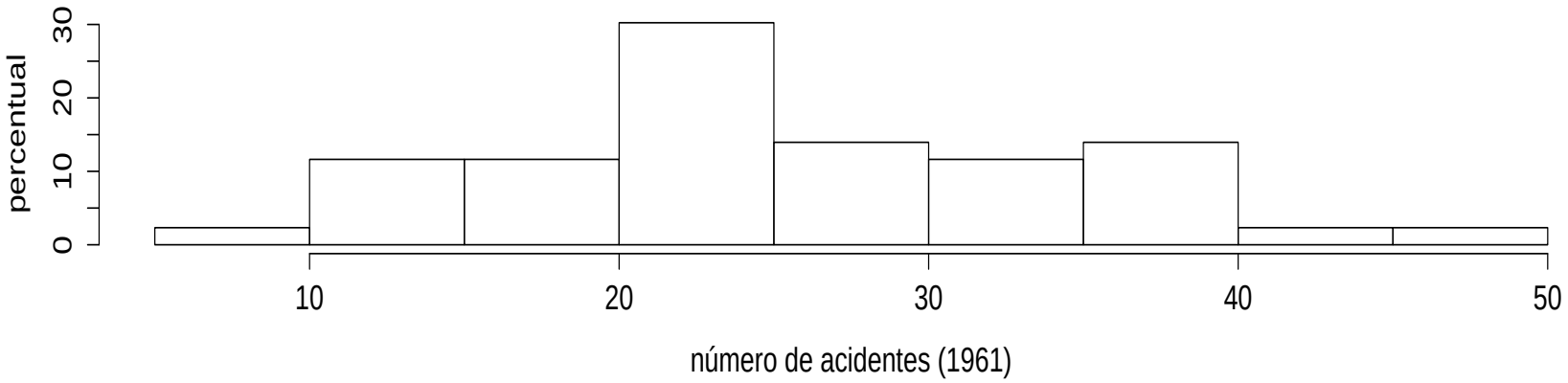
Exemplo 4: comparação do número de acidentes

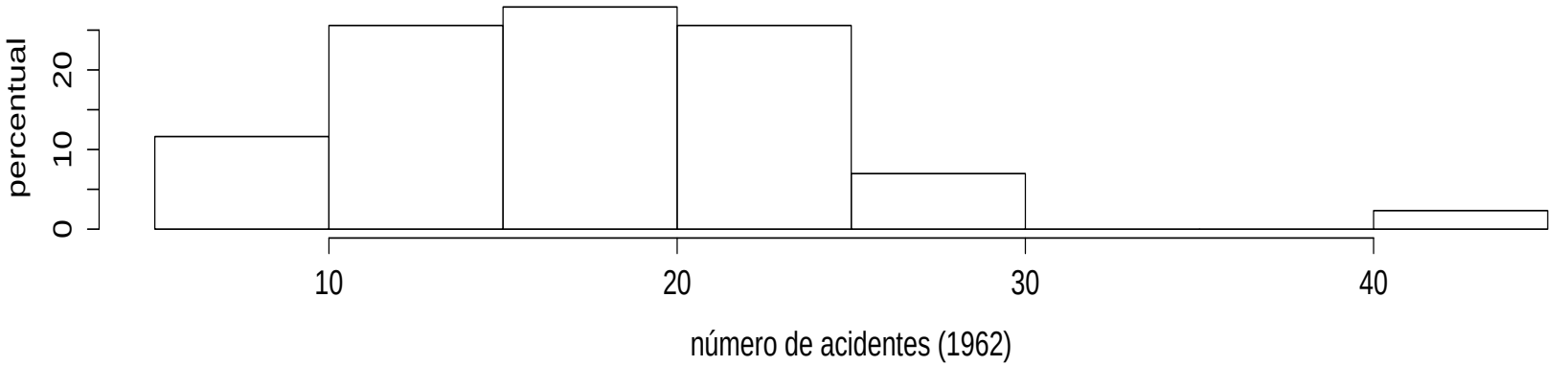
- Resposta: número de acidentes.
- Variável explicativa: ano (presença/ausência de limite de velocidade).
- Questão de interesse: a imposição dos limites de velocidade levou à redução do número de acidentes?

Boxplots do número de acidentes por ano



Histogramas do número de acidentes por ano





Medidas Resumo

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

Aparentemente há uma superdispersão nos dados (sob o modelo de Poisson) pois as variâncias amostrais (em cada grupo) são maiores do que as médias amostrais.

Distribuição de Poisson

- Função de probabilidade

  $Y \sim \text{Poisson}(\lambda)$, $\lambda > 0$, então

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

- 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

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

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

- 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.
- $g(.)$: função de ligação. Escolha usual $g(.) = \ln(.)$ (função de ligação canônica). Outras possibilidades: $\sqrt{(.)}, 1/(.)$
- 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

- Nesse caso, $\phi = 1$.
- Lembremos que, para esse modelo, a função de ligação logaritmica, $\ln(\mu_i)$ é a função de ligação canônica.
- Função escore: $\boldsymbol{S}(\boldsymbol{\beta}) = \dfrac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu})$, em que $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ e $\mu_i = e^{\eta_i}$.
- Informação de Fisher: $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$, $\boldsymbol{V} = \text{diag}(V_1, ..., V_n)$, $V_i = \mu_i, i = 1, 2, ..., n$.

Estimação por MV: ligação geral

- Função escore: $\boldsymbol{S}(\boldsymbol{\beta}) = \dfrac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$, em que $\mu_i = g^{-1}(\eta_i)$, $\boldsymbol{W} = \text{diag}(\omega_1, ..., \omega_n)$ e $\omega_i = \left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2 V_i^{-1} = \left(\dfrac{\partial g^{-1}(\eta_i)}{\partial \eta_i}\right)^2 \mu_i^{-1}$, $i = 1, 2, ..., n,$.
- Informação de Fisher: $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$.
- Reveja os desenvolvimentos para os MLG em `http://www.ime.unicamp.br/~cnaber/aula_Intro_MLG_MLG_1S_2016.pdf` e `http://www.ime.unicamp.br/~cnaber/aula_Intro_MLG_Parte2_MLG_1S_2016.pdf`.

Inferência por MV

- Dessa forma, devemos utilizar o processo iterativo (algoritmo Escore de Fisher), apresentado anteriormente (slide 16) de `http://www.ime.unicamp.br/~cnaber/aula_Intro_MLG_Parte2_MLG_1S_2016.pdf`), para obtermos estimativas para $\boldsymbol{\beta}$.
- As formas do desvio e do RCD para o modelo binomial/Bernoulli já foram vistas anteriormente, respectivamente: slides 26 (`http://www.ime.unicamp.br/~cnaber/aula_Intro_MLG_Parte2_MLG_1S_2016.pdf`) e slide 7 (`http://www.ime.unicamp.br/~cnaber/aula_Ana_Res_MLG_MLG_1S_2016.pdf`).
- Neste caso, a aproximação do desvio pela distribuição de qui-quadrado é apropriada quando $\mu_i \to \infty, \forall i$.

Exemplo 3: Modelo 1

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

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

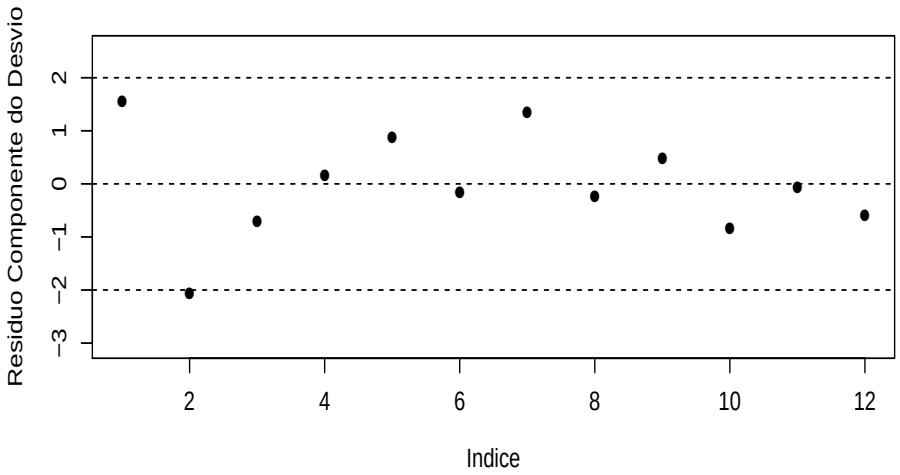
- Y_i : número de bactérias sobreviventes no instante i.
- x_i : tempo de exposição no instante i, $\overline{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = 6,5$.
- e^{β_0} : número esperado de bactérias sobreviventes no minuto 6,5.
- e^{β_1} : incremento (multiplicativo) no número esperado de bactérias sobreviventes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

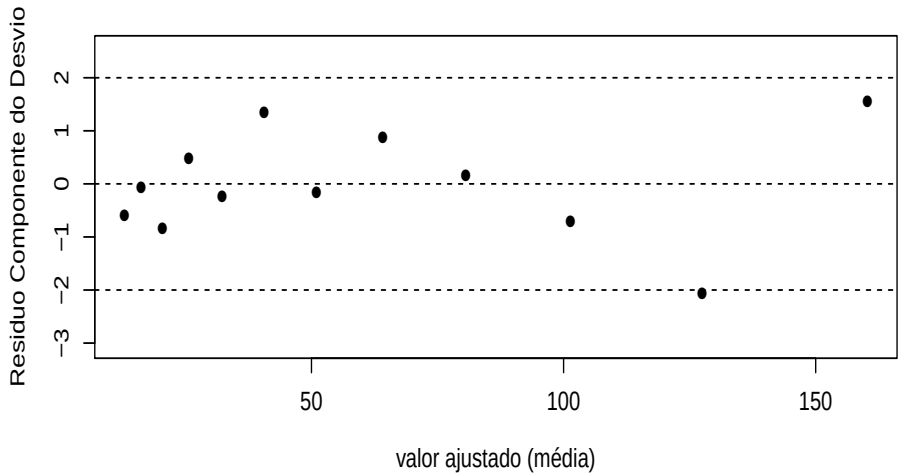
Exemplo 3: Modelo 2

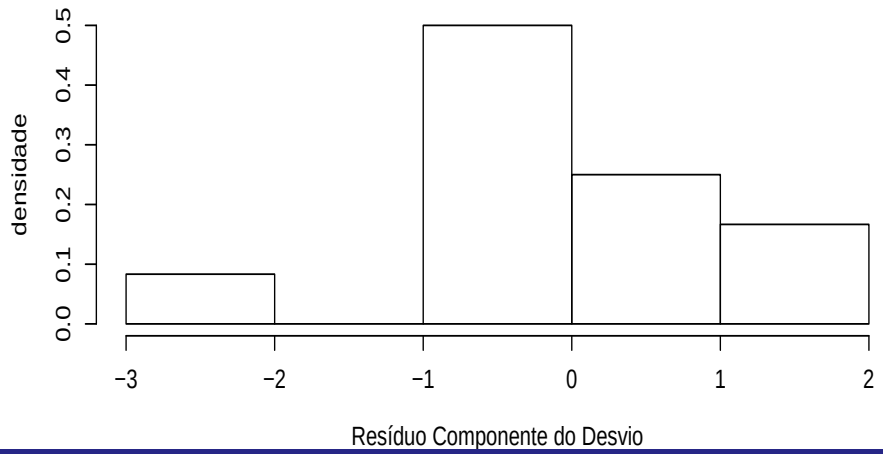
$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \text{Poisson}(\mu_i) \\
\ln(\mu_i) &= \beta_0 + \beta_1(x_i - \overline{x}) + \beta_2(x_i - \overline{x})^2 \rightarrow \\
\mu_i &= e^{\beta_0 + \beta_1(x_i - \overline{x}) + \beta_2(x_i - \overline{x})^2}, i = 1, 2, ..., 12
\end{aligned}
$$

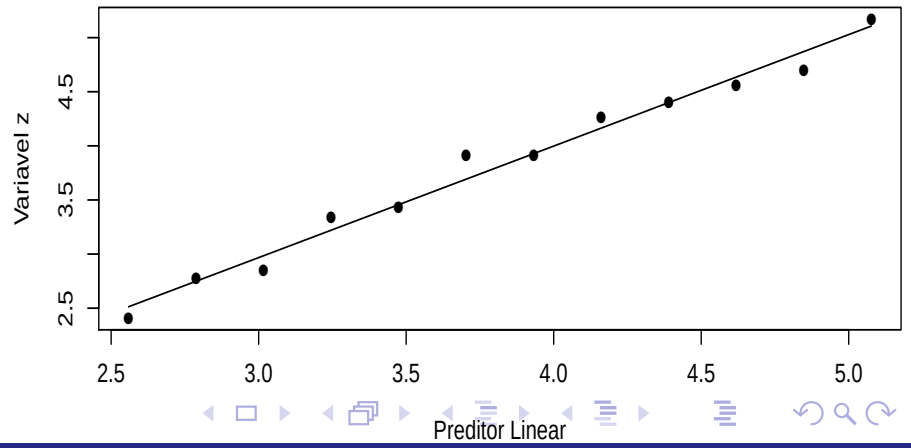
- Y_i : número de bactérias sobreviventes no instante i.
- x_i : tempo de exposição no instante i, $\overline{x} = \dfrac{1}{12}\sum_{i=1}^{12} x_i = 6,5$.
- e^{β_0} : número esperado de bactérias sobreviventes no minuto 6,5.
- $-\dfrac{\beta_1}{2\beta_2}$: minutos necessários para que o número de bactérias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

Gráficos de diagnóstico: modelo 1

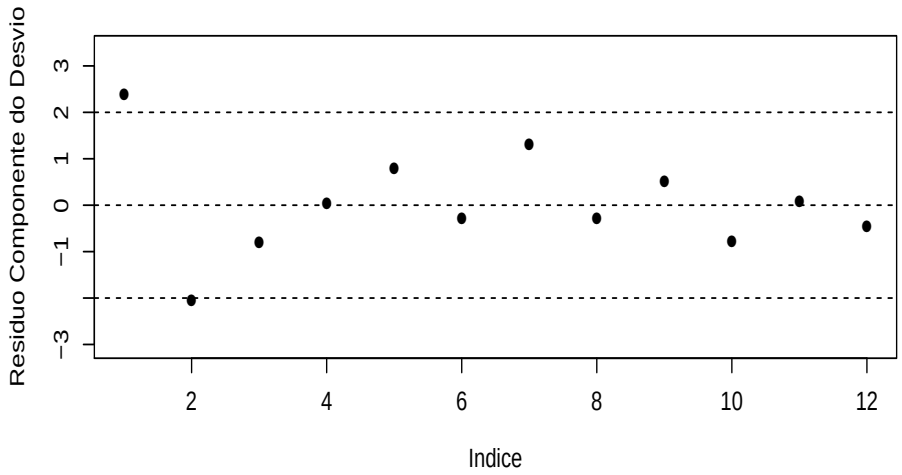


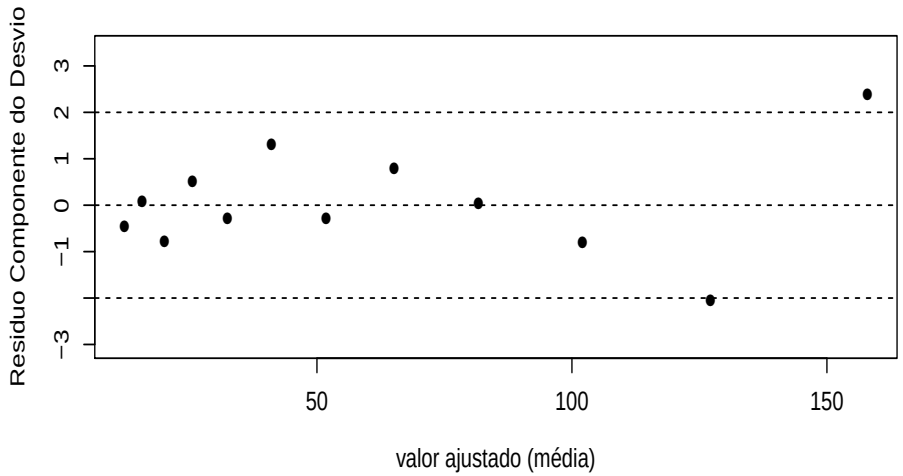


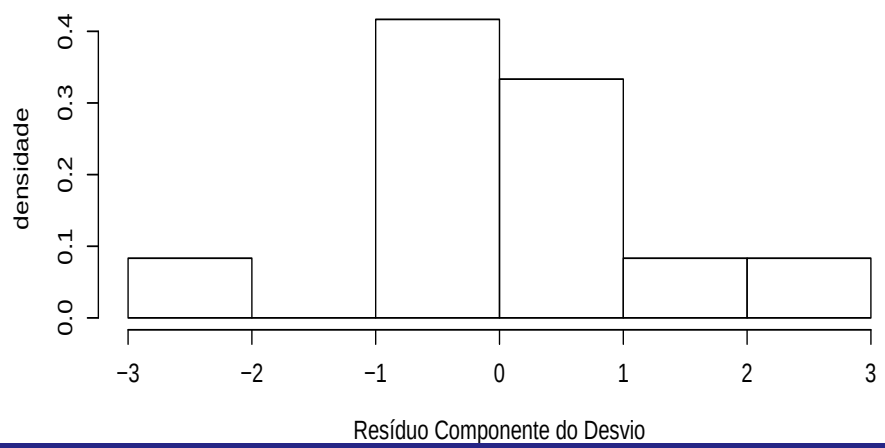


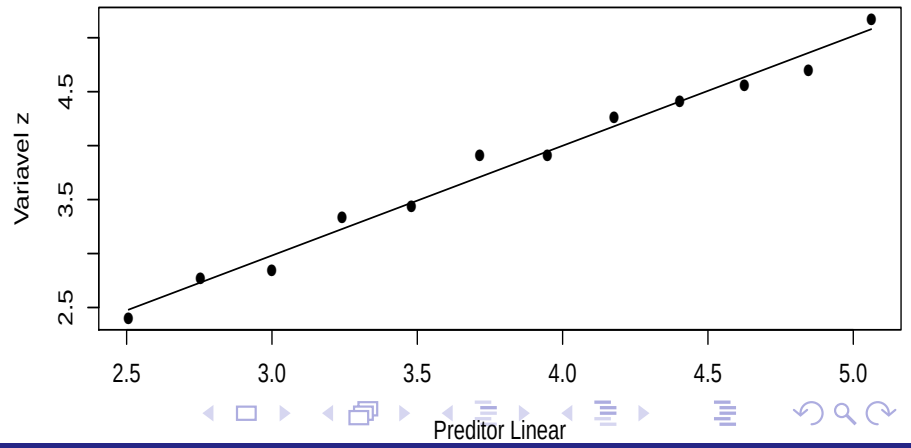


Gráficos de diagnóstico: modelo 2

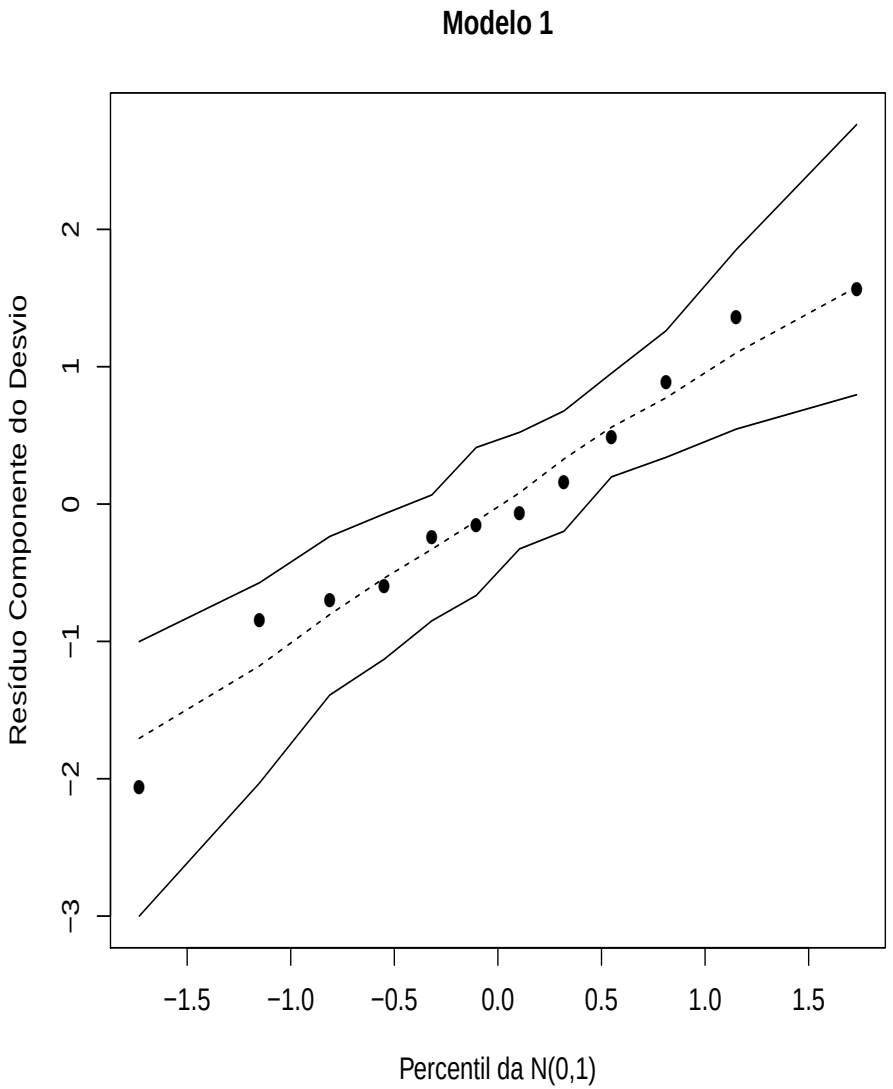


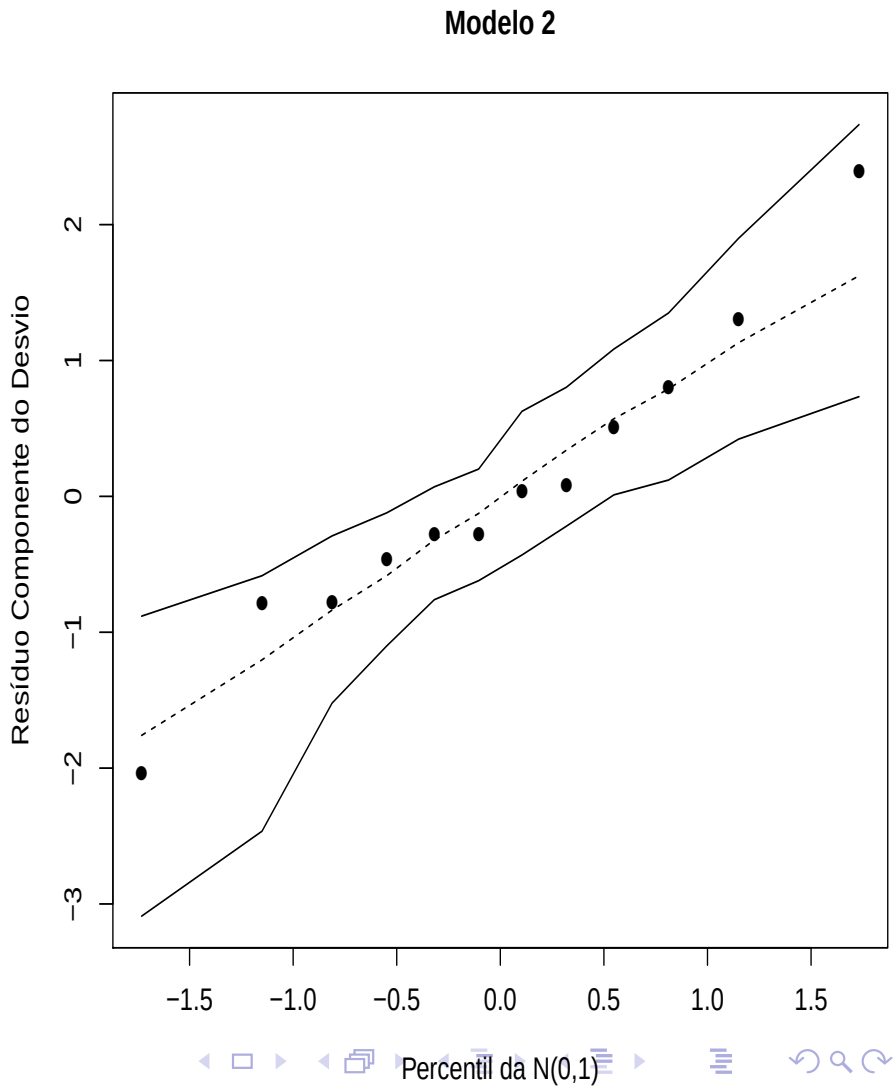




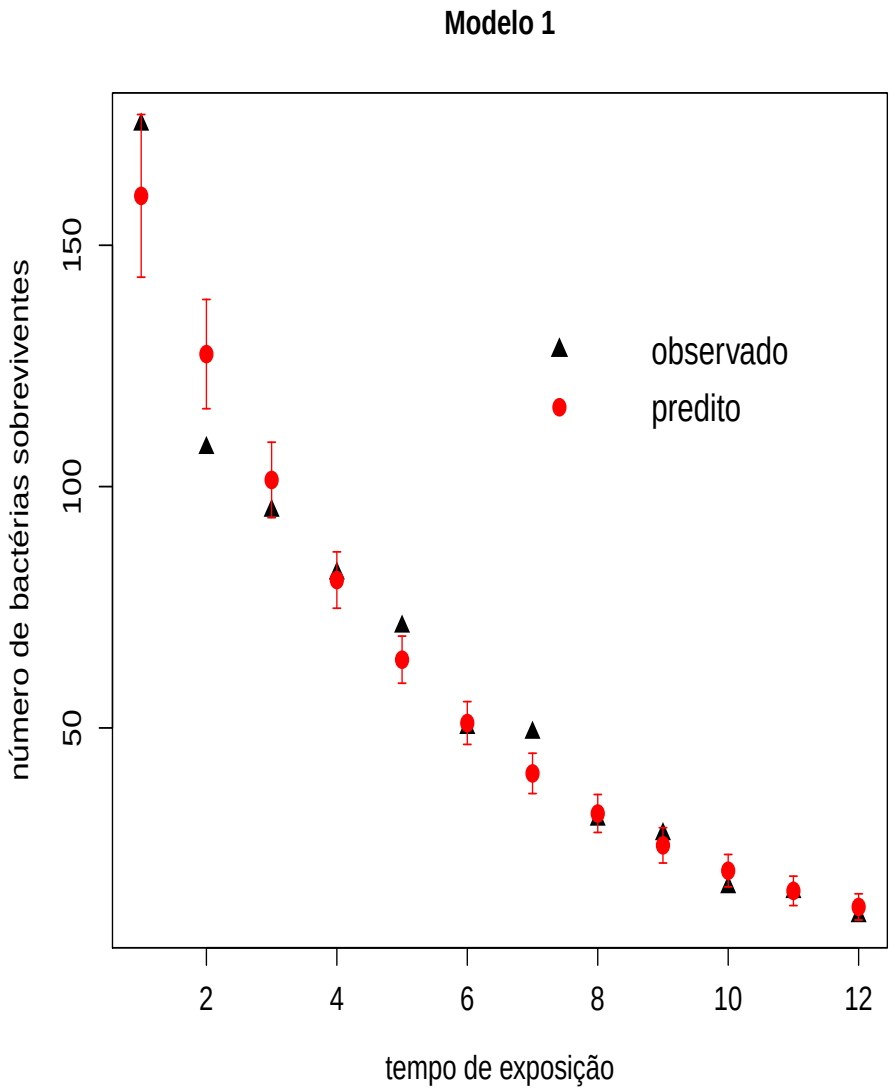


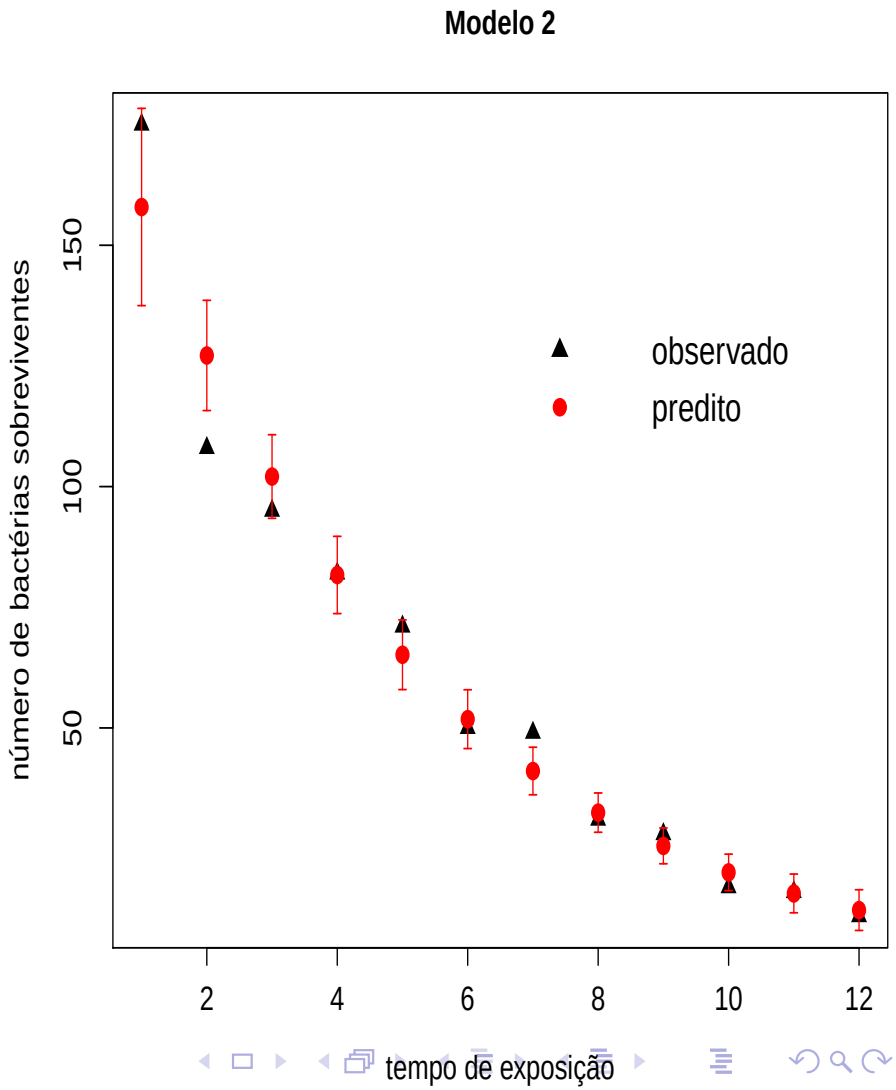
Gráficos de envelope: modelos 1 e 2





Valores observados e preditos : modelos 1 e 2





Estatísticas de comparação de modelos

Modelo	AIC	BIC	desvio	p-valor (desvio)
1	80,18	81,15	8,42	0,5877
2	82,04	83,49	8,27	0,5067
4	107,63	109,08	10,00	0,4405
5	96,71	98,65	9,00	0,4373
6	78,39	79,84	12,02	0,2834
7	77,86	79,80	12,02	0,2122

Os modelos 4 e 5 correspondem a modelos normais lineares enquanto que os modelos 6 e 7 correspondem a MLG com distribuição gama e função de ligação log (respectivamente com os preditores lineares dos modelos 1 e 2).

Estimativas dos parâmetros dos modelos

Modelo	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

O modelo 1 é preferível. A estimativa da taxa de decaimento do número de bactérias (e^{β_1}) (com intervalos de confiança de 95% entre parênteses) é : 0,79([0,77;0,81]).

Exemplo 3: Modelo 3 (regressão segmentada)

$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \text{Poisson}(\mu_i) \\
\ln(\mu_i) &= (\beta_{01} + \beta_{11}(x_i - \overline{x}))\mathbb{1}_{\{1,2,3\}}(x_i) + (\beta_{02} + \beta_{12}(x_i - \overline{x}))\mathbb{1}_{\{4,5,\ldots,12\}}(x_i) \\
\mu_i &= e^{(\beta_{01}+\beta_{11}(x_i-\overline{x}))\mathbb{1}_{\{1,2,3\}}(x_i)+(\beta_{02}+\beta_{12}(x_i-\overline{x}))\mathbb{1}_{\{4,5,\ldots,12\}}(x_i)}, i = 1, 2, ..., 12
\end{aligned}
$$

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

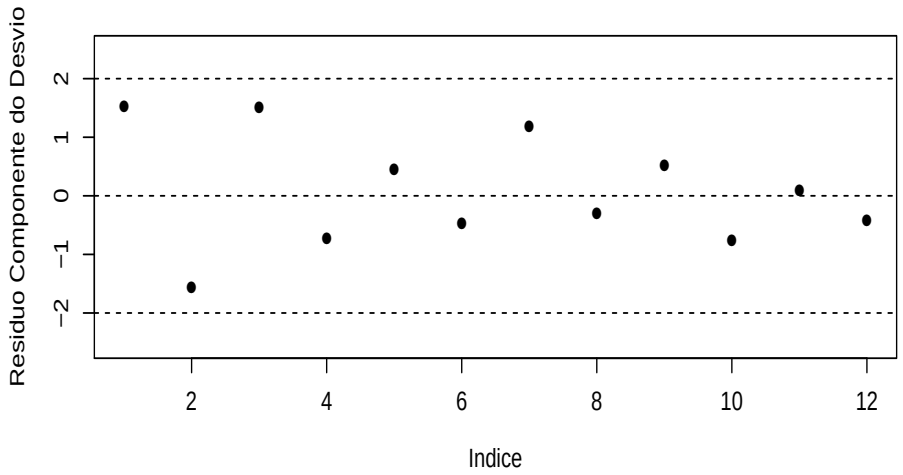
Exemplo 3: Modelo 3 (regressão segmentada)

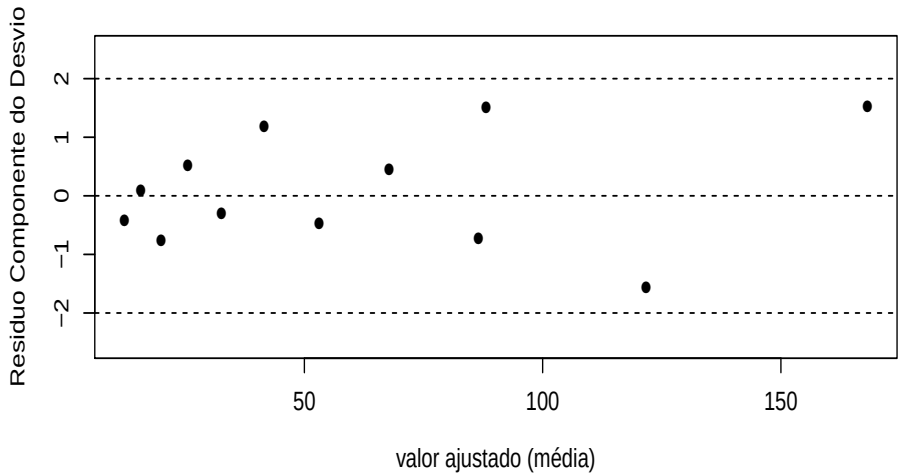
- $e^{\beta_{11}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição aumenta em um minuto, no primeiro intervalo $\{1, 2, 3\}$.
- $e^{\beta_{12}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição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

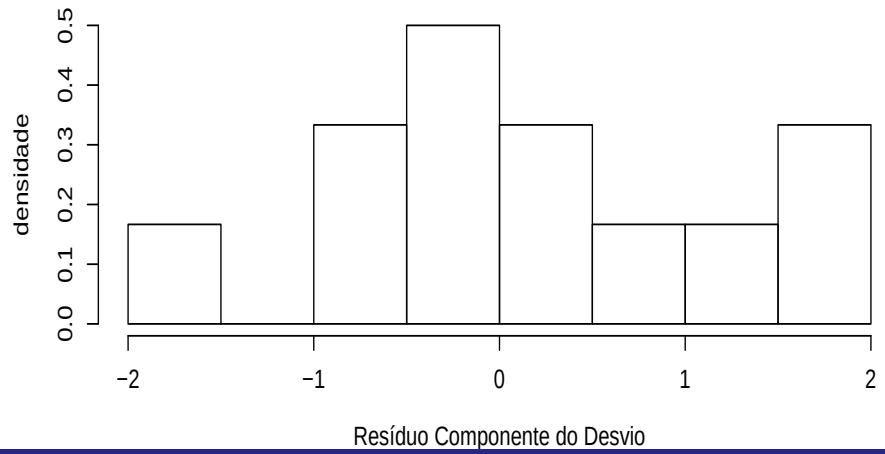
Modelo 3: matriz de planejamento

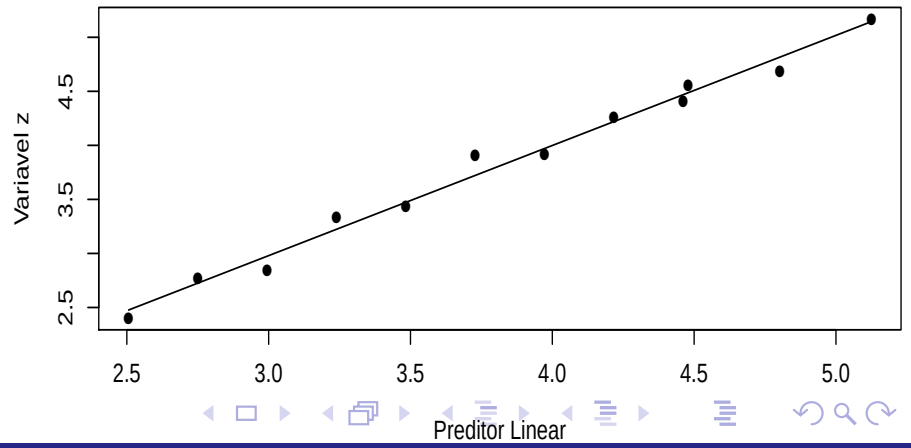
$$\boldsymbol{X} = \begin{bmatrix} 1 & -5,50 & 0 & 0 \\ 1 & -4,50 & 0 & 0 \\ 1 & -3,50 & 0 & 0 \\ 0 & 0 & 1 & -2,50 \\ 0 & 0 & 1 & -1,50 \\ 0 & 0 & 1 & -0,50 \\ 0 & 0 & 1 & 0,50 \\ 0 & 0 & 1 & 1,50 \\ 0 & 0 & 1 & 2,50 \\ 0 & 0 & 1 & 3,50 \\ 0 & 0 & 1 & 4,50 \\ 0 & 0 & 1 & 5,50 \end{bmatrix}$$

Gráficos de diagnóstico: modelo 3

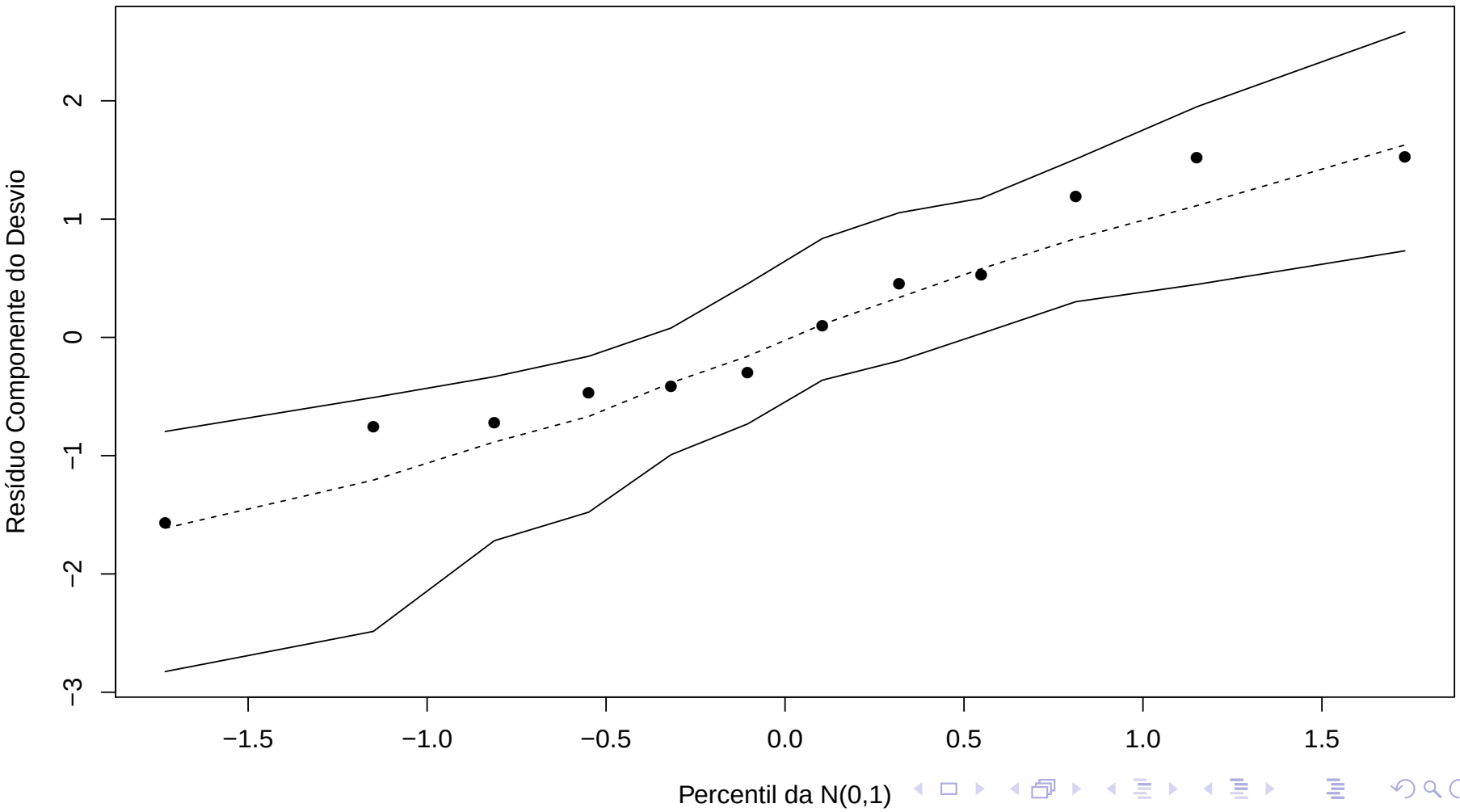








Gráficos de envelope: modelo 3



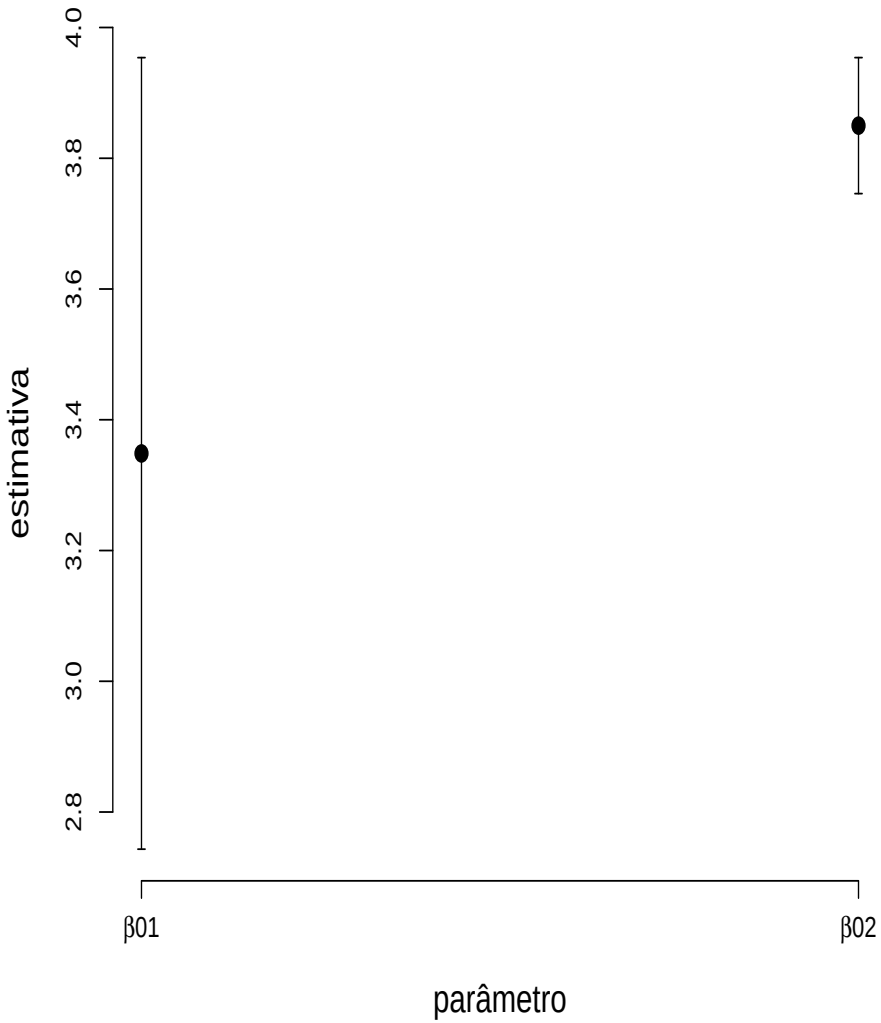
Estatísticas de comparação de modelos

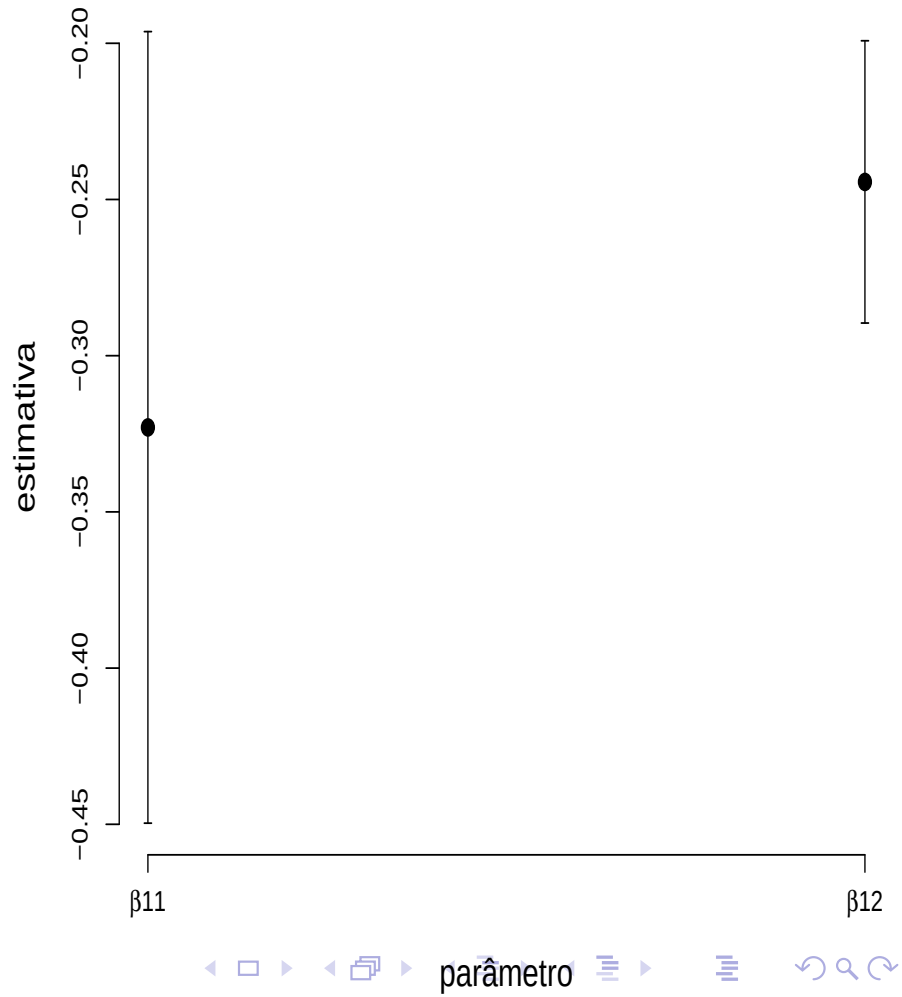
Modelo	AIC	BIC	desvio	p-valor (desvio)
1	80,18	81,15	8,42	0,5877
2	82,04	83,49	8,27	0,5067
3	80,92	82,86	5,16	0,7406

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