Modelos de regressão para dados discretos (parte 2): dados de contagens

Prof. Caio Azevedo

Exemplo 14: tempo de sobrevivências de bactérias

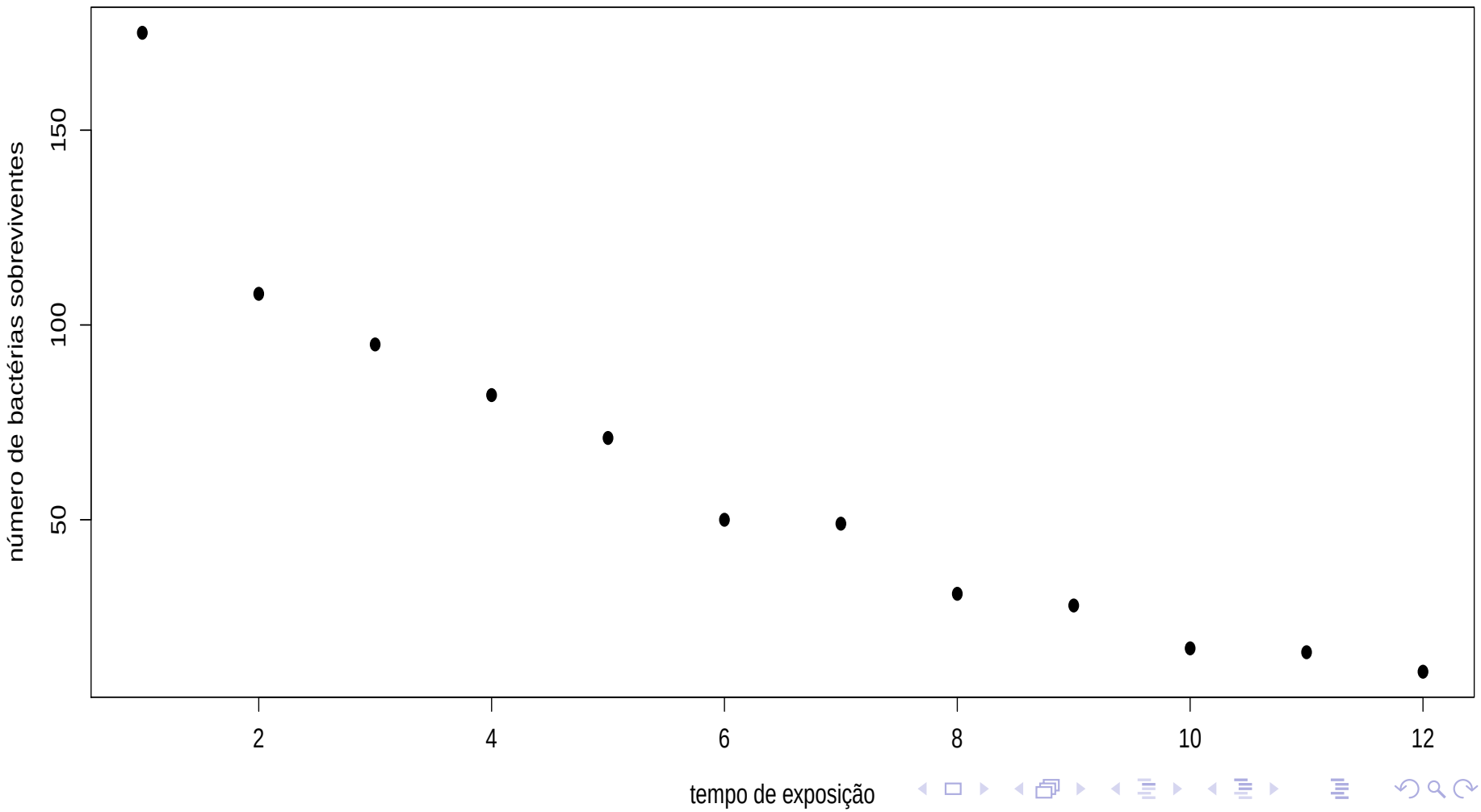
- Os dados correspondem ao número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto à uma temperatura de $300^o F (\approx 148,89^o C)$.
- Nessas amostras de alimentos foram feitas 12 medições, a cada minuto, contabilizando a quantidade 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



Características dos dados

- Variável resposta: número (Y_i); variável explicativa: tempo (x_i), i=1,2,...,12.
- A resposta corresponde à uma contagem. Assim

  $P(Y_i = k) > 0, \forall k \in \{0, 1, ...\}$ e

  $P(Y_i \in [r_1, r_2]) = 0, \forall r_1, r_2 < 0, r_1 \leq r_2$.
- Aparentemente, há uma relação não linear (curva de segundo grau, exponencial negativa etc) entre a resposta e a variável explicativa.

Modelo 0

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

- Y_i : número de bactérias sobreviventes ao tempo de exposição i.
- x_i : tempo de exposição i.
- $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. $\mathcal{E}(Y_i) = \mathcal{V}(Y_i) = \mu_i = e^{\beta_0 + \beta_1 x_i}$.
- ln(.): função de ligação (log ou logarítmica)

Interpretação dos parâmetros

- Lembrando que $\mu_i = e^{\beta_0+\beta_1 x_i}$, assim se $\mu_{i+1} = e^{\beta_0+\beta_1(x_i+1)}$ então $\mu_{i+1} = \mu_i e^{\beta_1}$.
- e^{β_0}: número médio de bactérias sobreviventes expostas durante 0 minutos à temperatura de 300^oF (em termos do problema, esta interpretação faz sentido?).
- e^{β_1}: incremento multiplicativo (positivo ou negativo) no número médio de bactérias sobreviventes para o aumento em 1 minuto no tempo de exposição à temperatura de 300^oF.
- Tem-se um modelo de regressão Poisson log-linear (função de ligação log).

Modelo 1

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

- Y_i : número de bactérias sobreviventes ao tempo de exposição i.
- x_i : tempo de exposição i e $\overline{x} = \frac{1}{12}\sum_{i=1}^{12} x_i$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. $\mathcal{E}(Y_i) = \mathcal{V}(Y_i) = \mu_i = e^{\beta_0 + \beta_1(x_i - \overline{x})}$.
- ln(.): função de ligação (log ou logarítmica)

Interpretação dos parâmetros

- Lembrando que $\mu_i = e^{\beta_0+\beta_1(x_i-\overline{x})}$, assim se $\mu_{i+1} = e^{\beta_0+\beta_1(x_i-\overline{x}+1)}$ então $\mu_{i+1} = \mu_i e^{\beta_1}$.
- e^{β_0}: número médio de bactérias sobreviventes expostas durante $\overline{x} = 6,5$ minutos à temperatura de $300^o F$.
- e^{β_1}: incremento multiplicativo (positivo ou negativo) no número médio de bactérias para o aumento em 1 minuto no tempo de exposição à temperatura de $300^o F$.

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.
- $-\dfrac{\beta_1}{2\beta_2} + \overline{x}$: minutos necessários para que o número de bactérias sobreviventes seja mínimo.

Modelo geral

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

$$\ln(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ji} \quad \rightarrow \quad \mu_i = e^{\sum_{j=1}^{p} \beta_j x_{ji}}, i = 1, 2, ..., n, j = 1, 2, ..., p$$

- Y_i : contagem de interesse associada a 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.
- ln(.) : função de ligação (log).
- 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 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 ln(.)) é apropriada.
  - O preditor linear é apropriado.

Inferência para o modelo

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

$$L(\boldsymbol{\beta}) = \frac{e^{-\sum_{i=1}^{n} \mu_i} \prod_{i=1}^{n} \mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} \propto e^{-\sum_{i=1}^{n} \mu_i} \prod_{i=1}^{n} \mu_i^{y_i}$$

- Logverossimilhança.

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \ln(\mu_i) + \text{const.}$$

Inferência para o modelo

- Como $\mu_i = e^{\eta_i}$, temos que a logverossimilhança traduz-se em

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} e^{\eta_i} + \sum_{i=1}^{n} y_i \eta_i = \sum_{i=1}^{n} (y_i \eta_i - e^{\eta_i}) \qquad (1)$$

- Vetor escore(**exercício**)

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} - e^{\eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \right) = \sum_{i=1}^{n} (y_i - e^{\eta_i}) \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} (y_i - e^{\eta_i}) \boldsymbol{X}_i = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned}$$

pois $\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_i$, $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ e $\mu_i = e^{\boldsymbol{\eta}_i}$.

Inferência para o modelo

- Além disso, $\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta}) \dfrac{\partial \eta_i}{\partial \boldsymbol{\beta}}$, em que $g_i(\boldsymbol{\beta}) = y_i - e^{\eta_i}$.

  Assim, $\mathcal{E}(g_i(\boldsymbol{\beta})) = \mathcal{E}(Y_i - e^{\eta_i}) = 0$ e $\dfrac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -e^{\eta_i} \dfrac{\partial \eta_i}{\partial \boldsymbol{\beta}}$.

- Por outro lado, a matriz Hessiana é dada por

$$\boldsymbol{H}(\boldsymbol{\beta}) = \frac{l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \left[g_i(\boldsymbol{\beta}) \frac{\partial \eta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} + \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \right]$$

- Note que $\dfrac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = -\dfrac{\partial e^{\eta_i}}{\partial \boldsymbol{\beta}'} = -e^{\eta_i} \dfrac{\partial \eta_i}{\partial \boldsymbol{\beta}'} = -e^{\eta_i} \boldsymbol{X}_i'$

Inferência para o modelo

- Logo, a informação de Fisher corresponde à

$$\boldsymbol{I}(\beta) = -\mathcal{E}(\boldsymbol{H}(\beta)) = \sum_{i=1}^{n} e^{\eta_i} \boldsymbol{X}_i \boldsymbol{X}_i' = \boldsymbol{X}' \boldsymbol{V} \boldsymbol{X}.$$

  em que $\boldsymbol{V} = \text{diag}(e^{\eta_1}, ..., e^{\eta_n})$.

- Repetir os desenvolvimentos considerando $\sqrt{\mu_i} = \boldsymbol{X}_i' \beta$ (função de ligação raiz quadrada).

Comentários

- Os resultados anteriores (modelos de regressão logística) continuam válidos, com pequenas modificações.
- $\mathcal{V}(Y_i) = \mu_i$.
- Desvio:

$$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{k} \left\{ [y_i \ln(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)] \, I_{\{1,2,\ldots\}}(y_i) + \widehat{\mu}_i \mathbb{1}_{\{0\}}(y_i) \right\}.$$

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

Comentários

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

$$\begin{aligned} T_{D_i} &= \pm \frac{\sqrt{2}}{\sqrt{1-\widehat{h}_{ii}}} \{y_i \ln(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\}^{1/2} I_{\{1,2,....\}}(y_i) \\ &\pm \frac{\sqrt{2\widehat{\mu}_i}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{0\}}(y_i). \end{aligned}$$

  em que $\pm$ assume o mesmo sinal de $y_i - \mu_i$ e $\widehat{h}_{ii}$ é o i-esimo elemento da diagonal principal da matriz $\boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{1/2}$.

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

Algoritmo escore de Fisher

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

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

até que algum critério de convergência seja satisfeito, como

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

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

Algoritmo escore de Fisher

- A equação (2) pode ser reescrita como

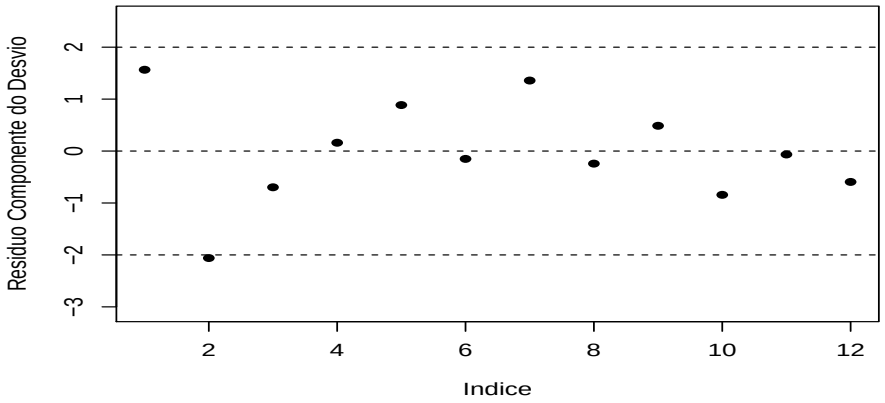
$$\beta^{(t+1)} = \left(\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)},$$

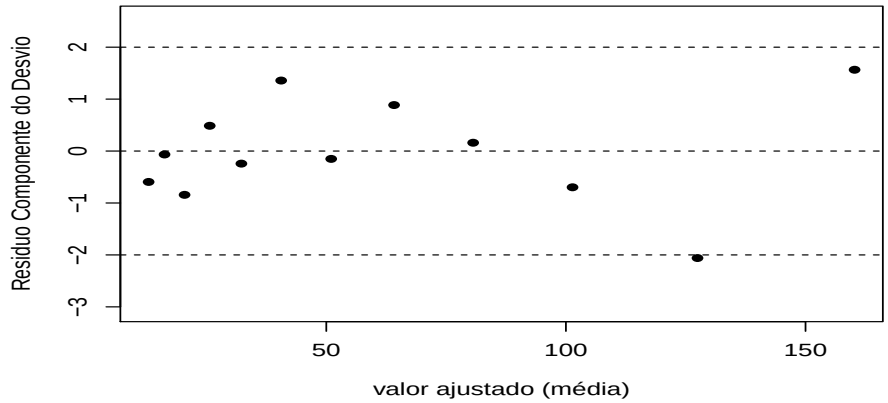
  em que $\boldsymbol{z} = \eta + \boldsymbol{W}^{-1/2}\boldsymbol{D}^{-1/2}(\boldsymbol{y} - \mu)$.

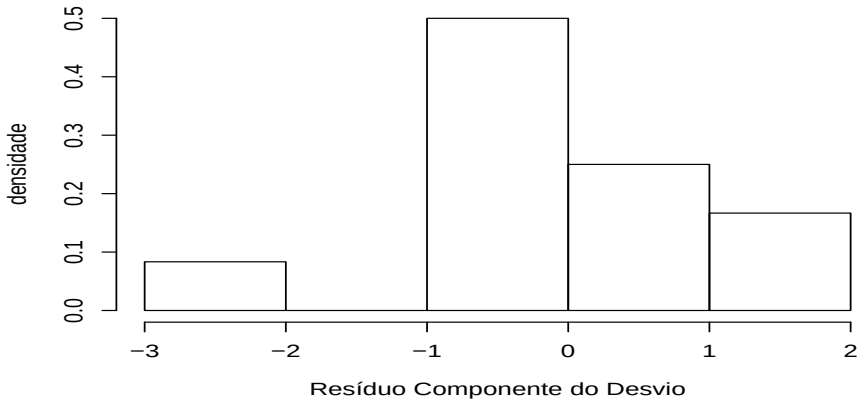
- Estimativas iniciais: é usual considerarmos $\eta^{(0)} = F^{-1}(\boldsymbol{y})$ (e.g., $F^{-1} = \ln(.)$).

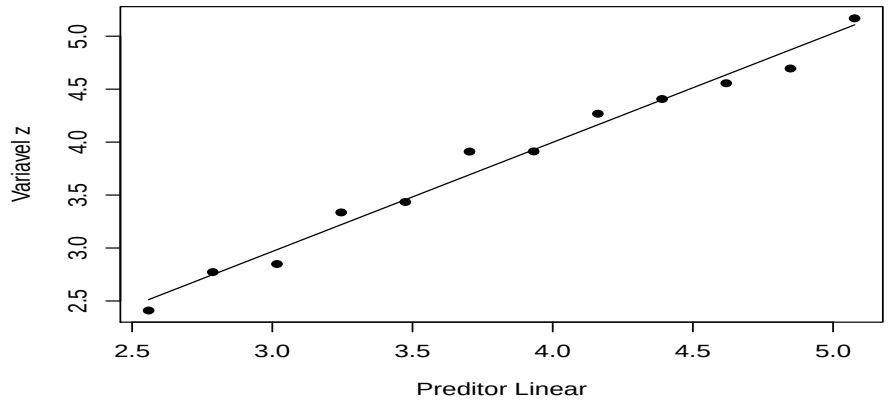
- Voltemos agora ao Exemplo 14.

Gráficos de diagnóstico: modelo 1

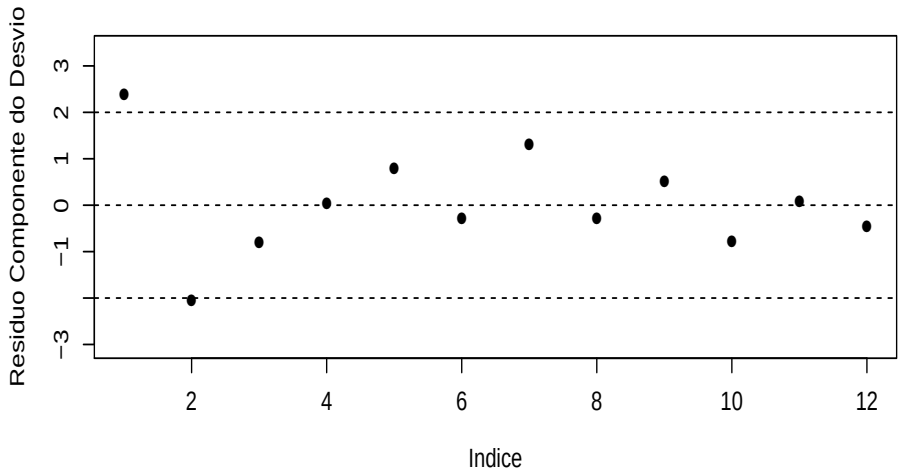


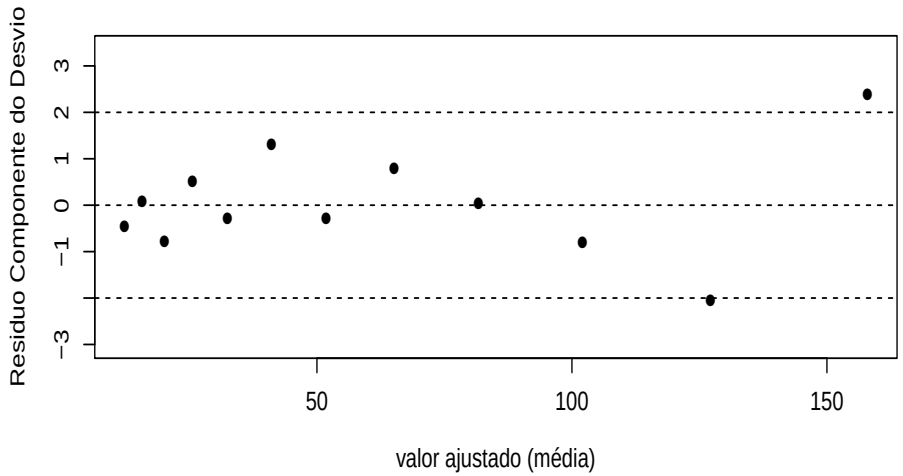


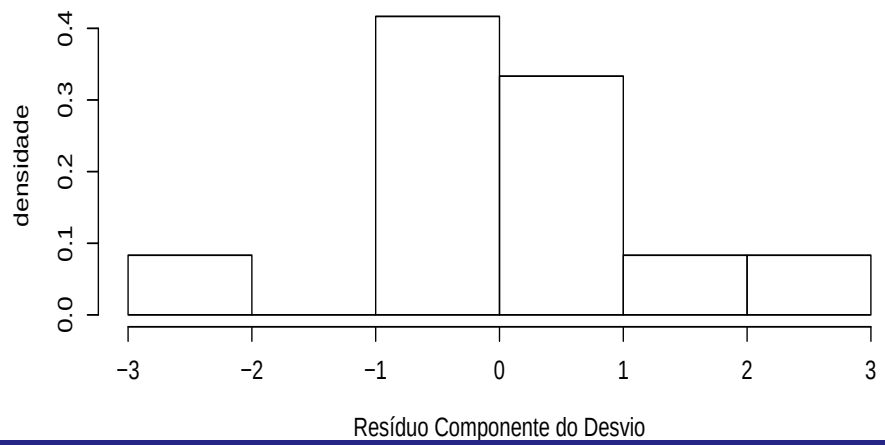


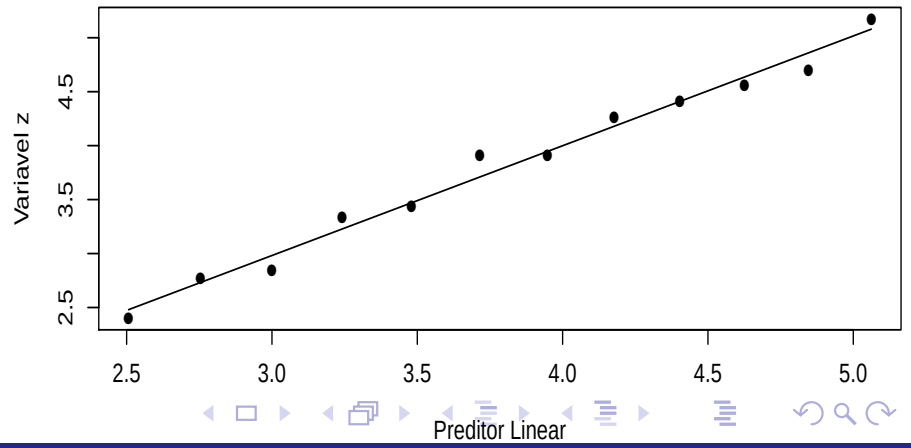


Gráficos de diagnóstico: modelo 2

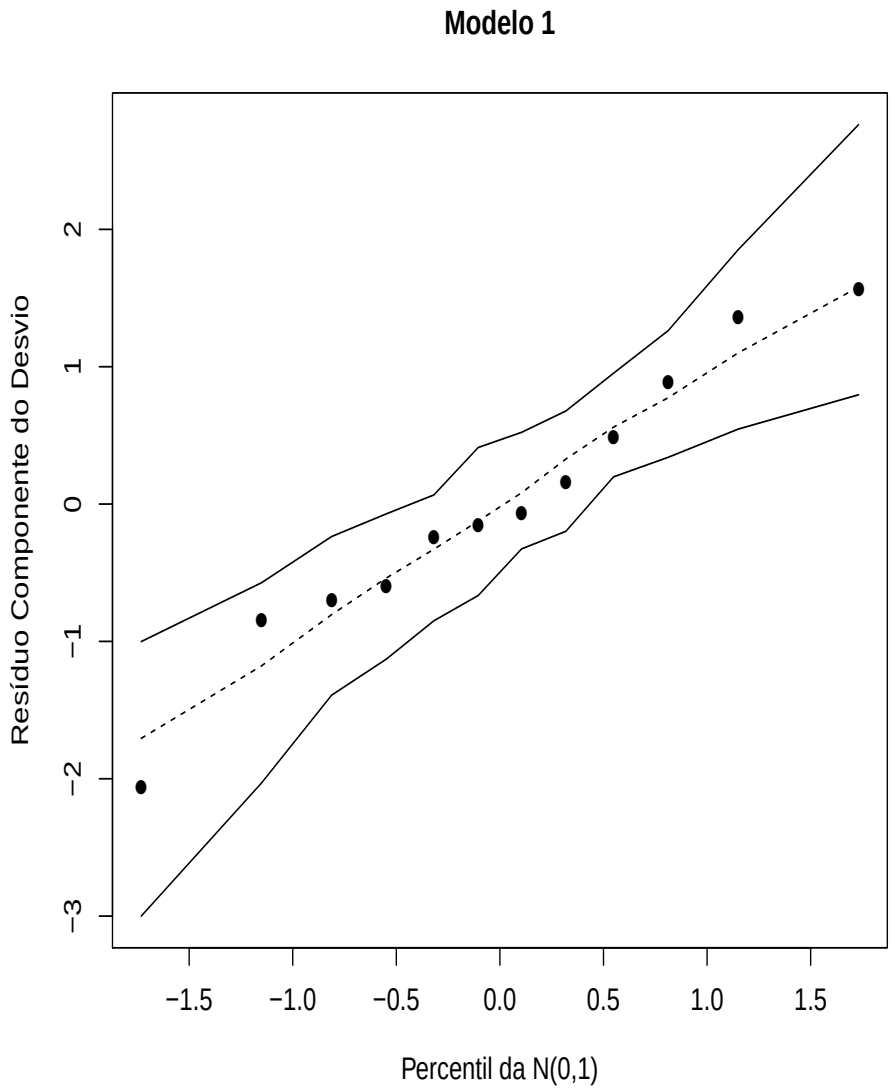


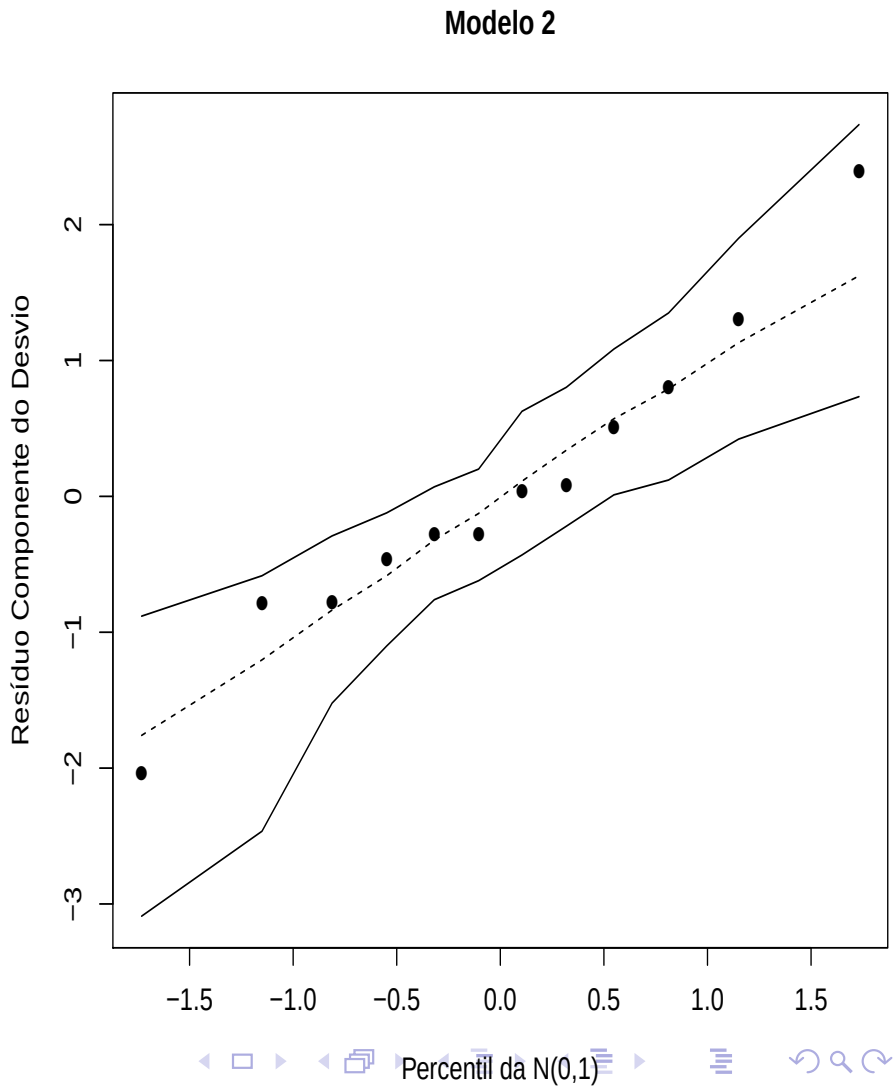




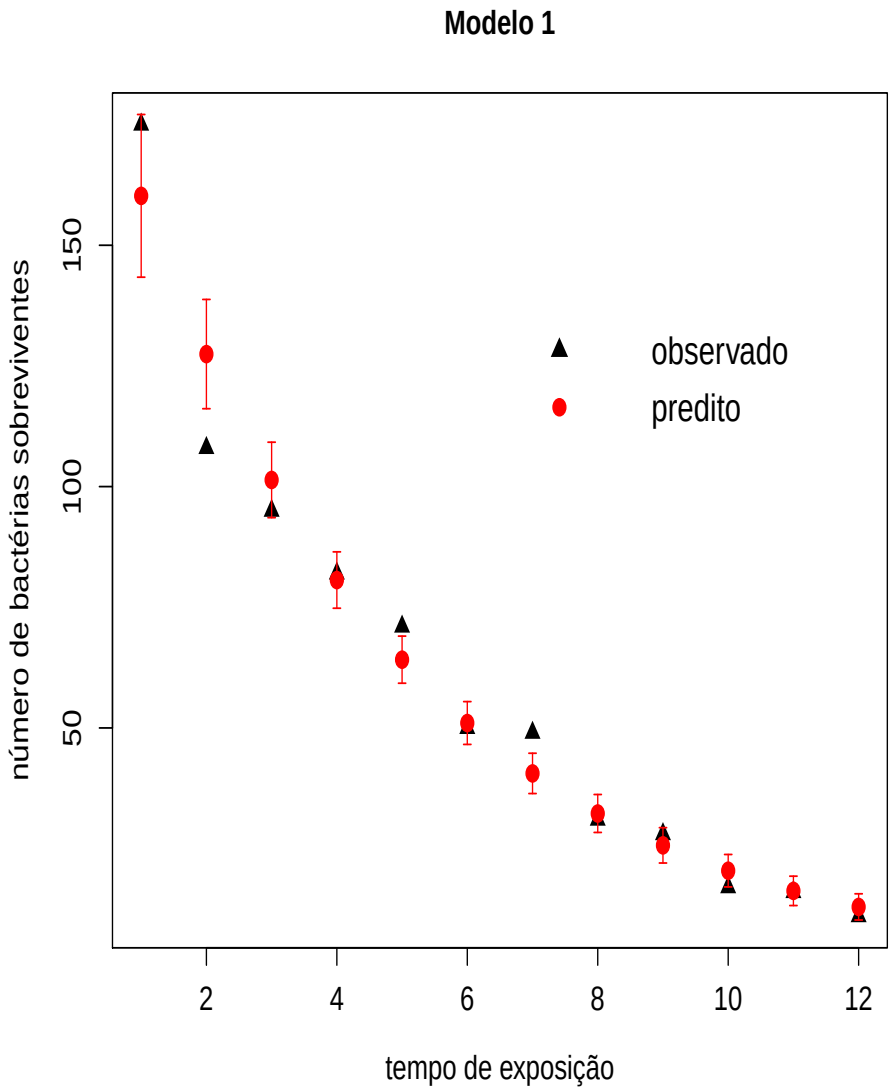


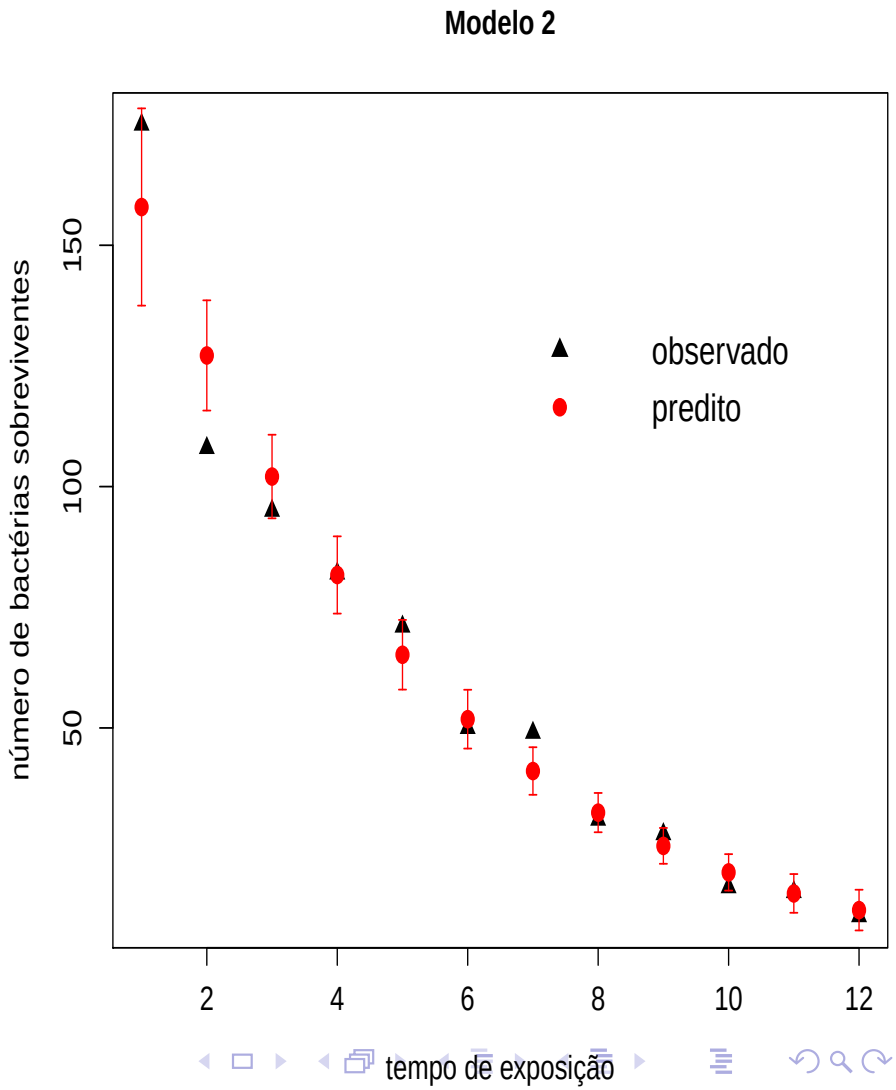
Gráficos de envelope: modelos 1 e 2





Valores observados e preditos : modelos 1 e 2





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]).

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.

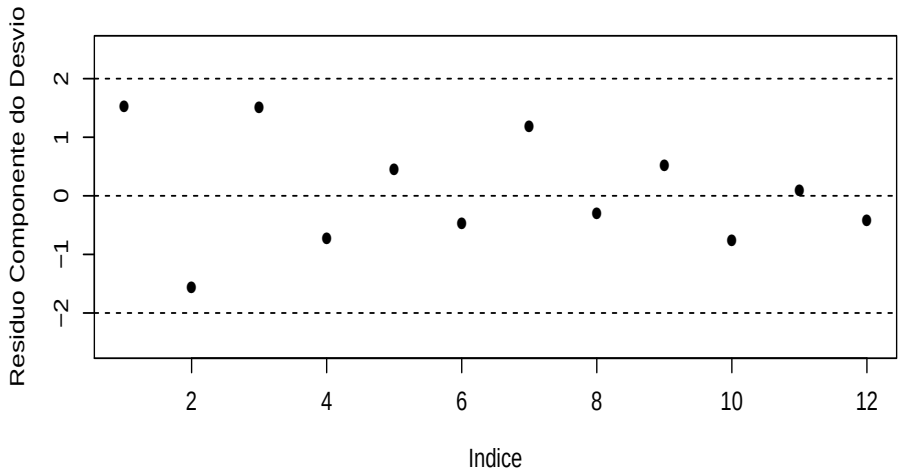
Exemplo 3: Modelo 3 (regressão segmentada)

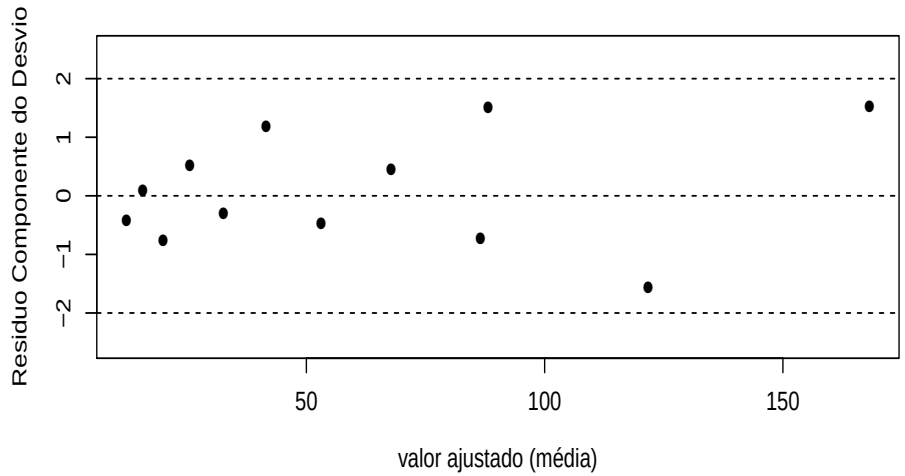
- $e^{\beta_{11}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição aumenta em um minuto, no primeiro intervalo $\{1, 2, 3\}$.
- $e^{\beta_{12}}$: incremento (multiplicativo) no número esperado de bactérias sobreviventes quando o tempo de exposição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

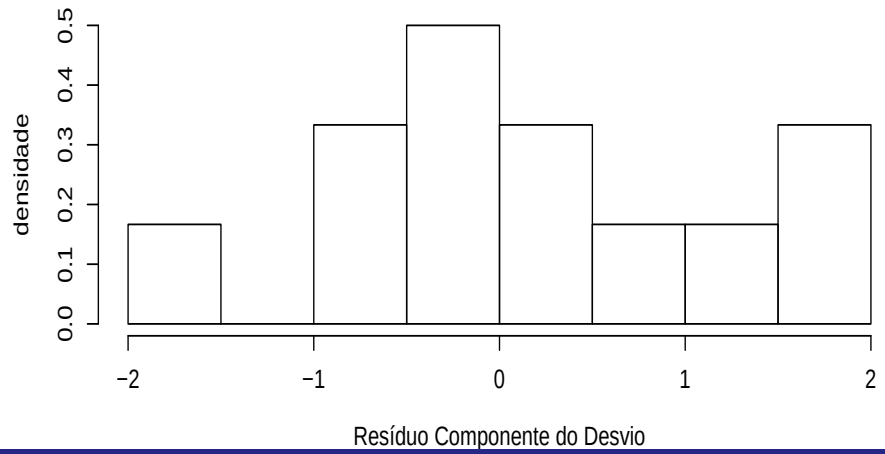
Modelo 3: matriz de planejamento

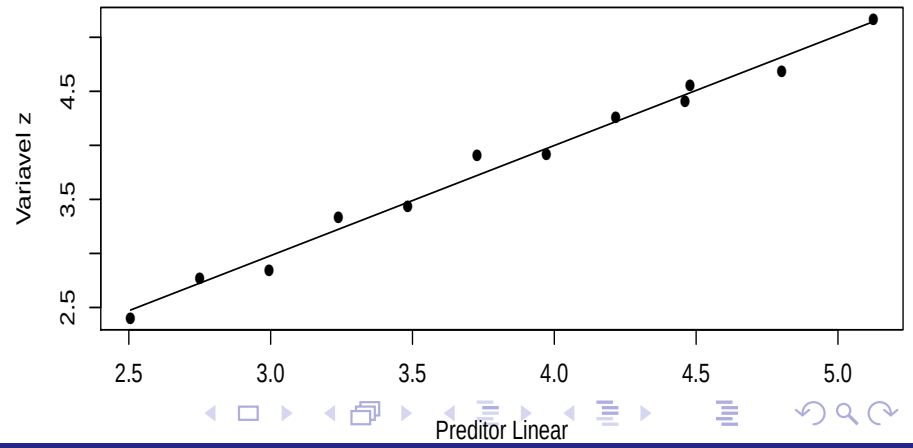
$$\boldsymbol{X} = \begin{bmatrix} 1 & -5,50 & 0 & 0 \\ 1 & -4,50 & 0 & 0 \\ 1 & -3,50 & 0 & 0 \\ 0 & 0 & 1 & -2,50 \\ 0 & 0 & 1 & -1,50 \\ 0 & 0 & 1 & -0,50 \\ 0 & 0 & 1 & 0,50 \\ 0 & 0 & 1 & 1,50 \\ 0 & 0 & 1 & 2,50 \\ 0 & 0 & 1 & 3,50 \\ 0 & 0 & 1 & 4,50 \\ 0 & 0 & 1 & 5,50 \end{bmatrix}$$

Gráficos de diagnóstico: modelo 3

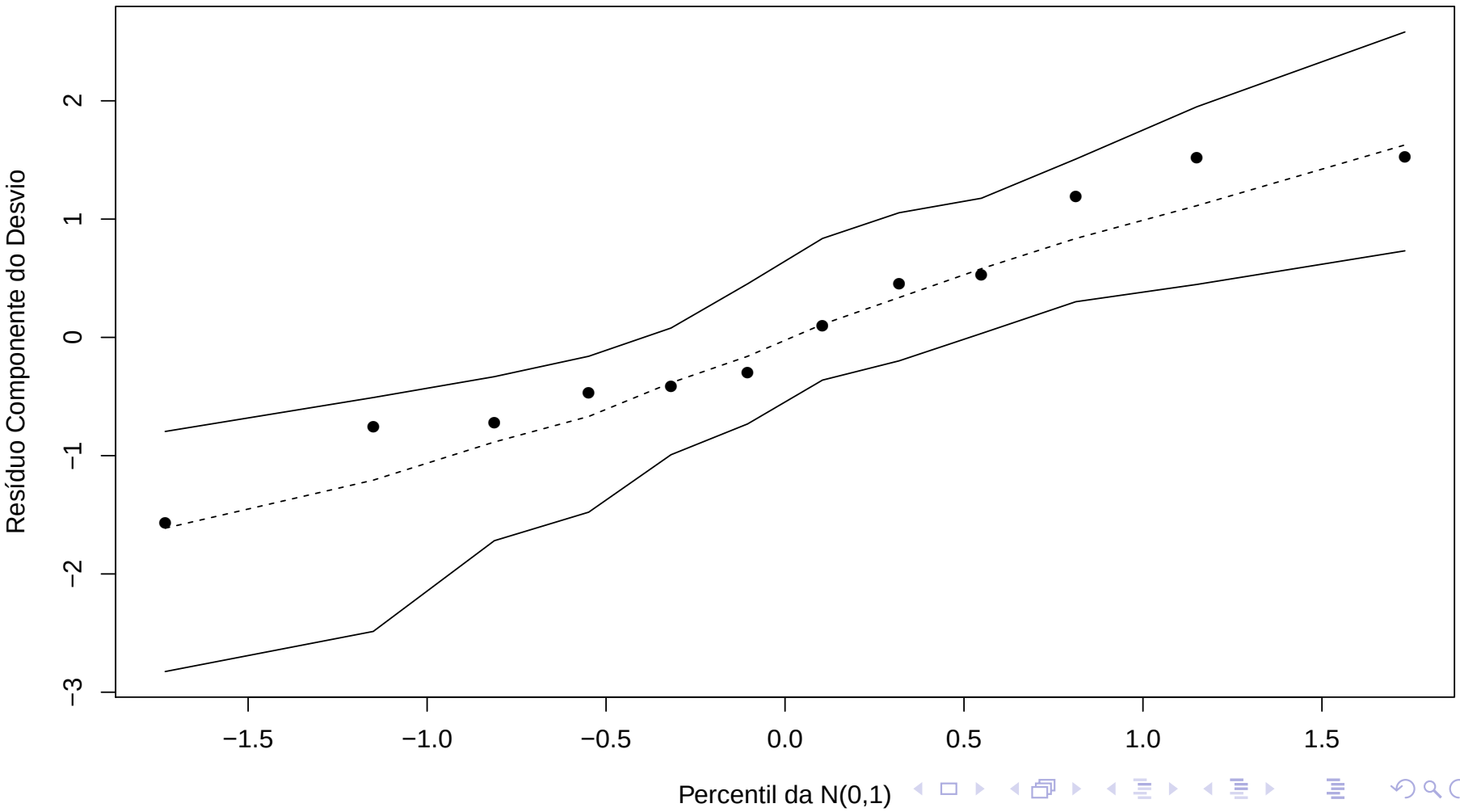








Gráficos de envelope: modelo 3



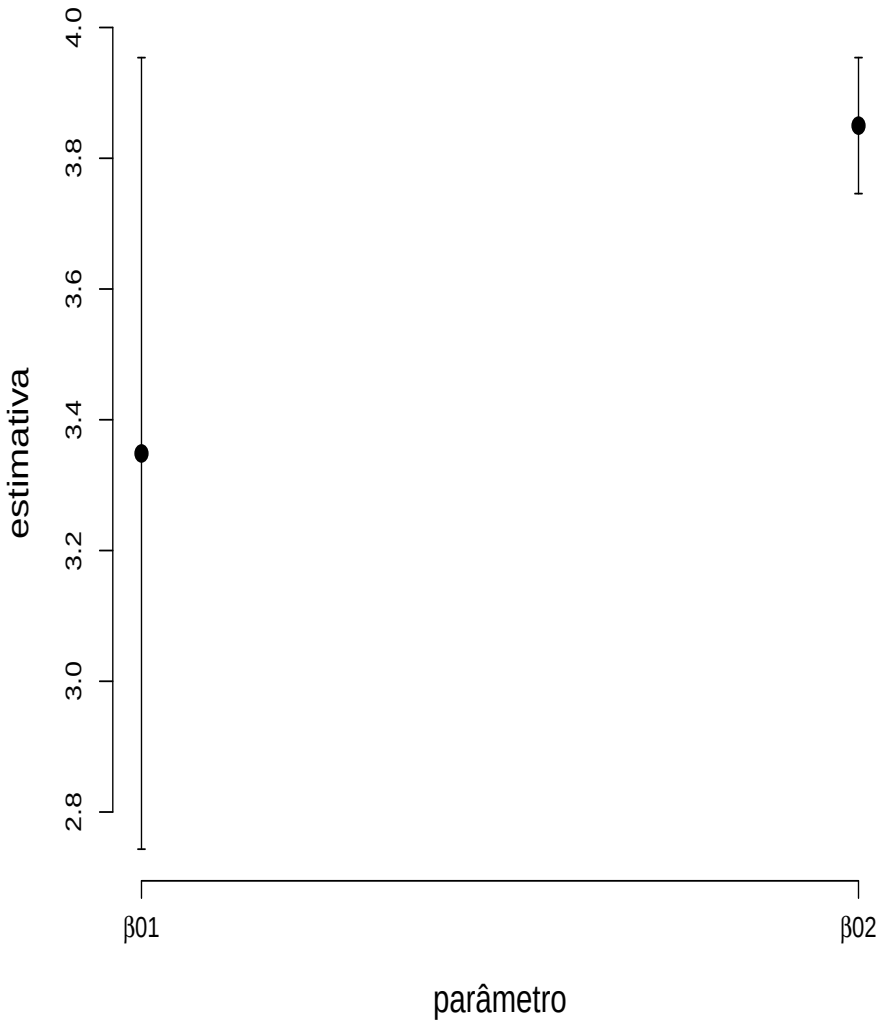
Estatísticas de comparação de modelos

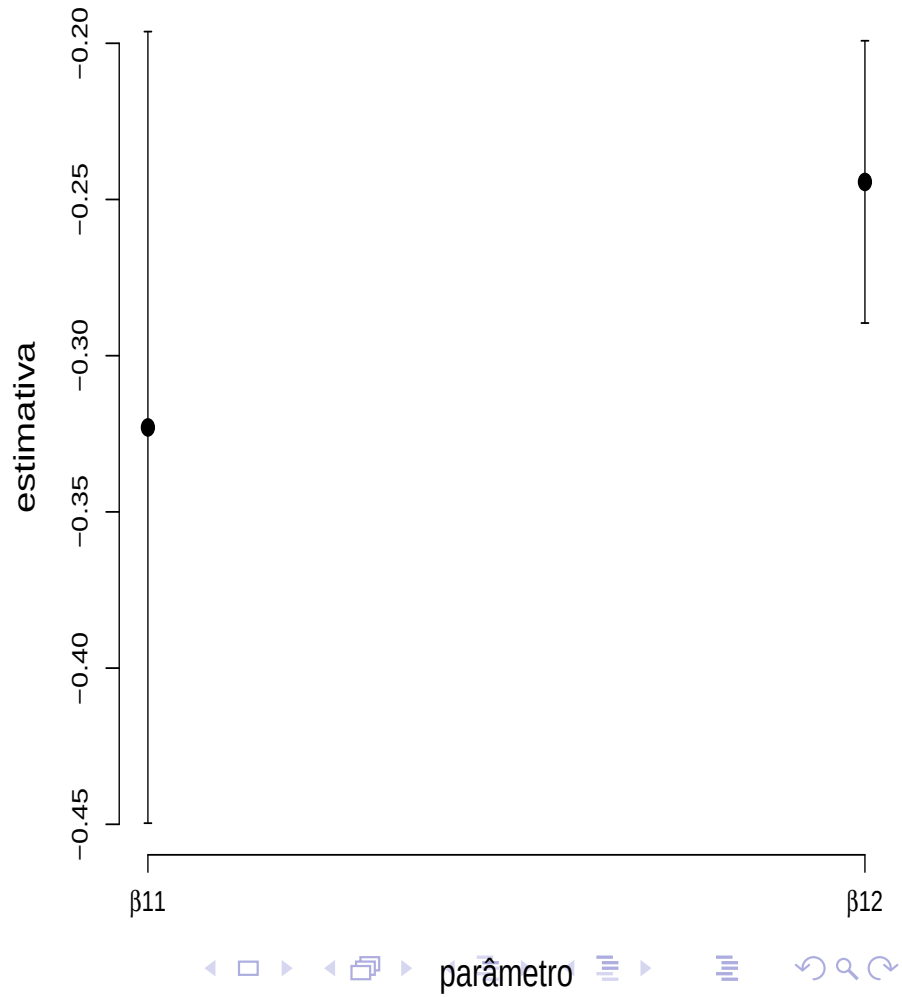
Modelo	AIC	BIC	desvio	p-valor (desvio)
1	80,18	81,15	8,42	0,5877
2	82,04	83,49	8,27	0,5067
3	80,92	82,86	5,16	0,7406

Estimativas dos parâmetros do modelo 3

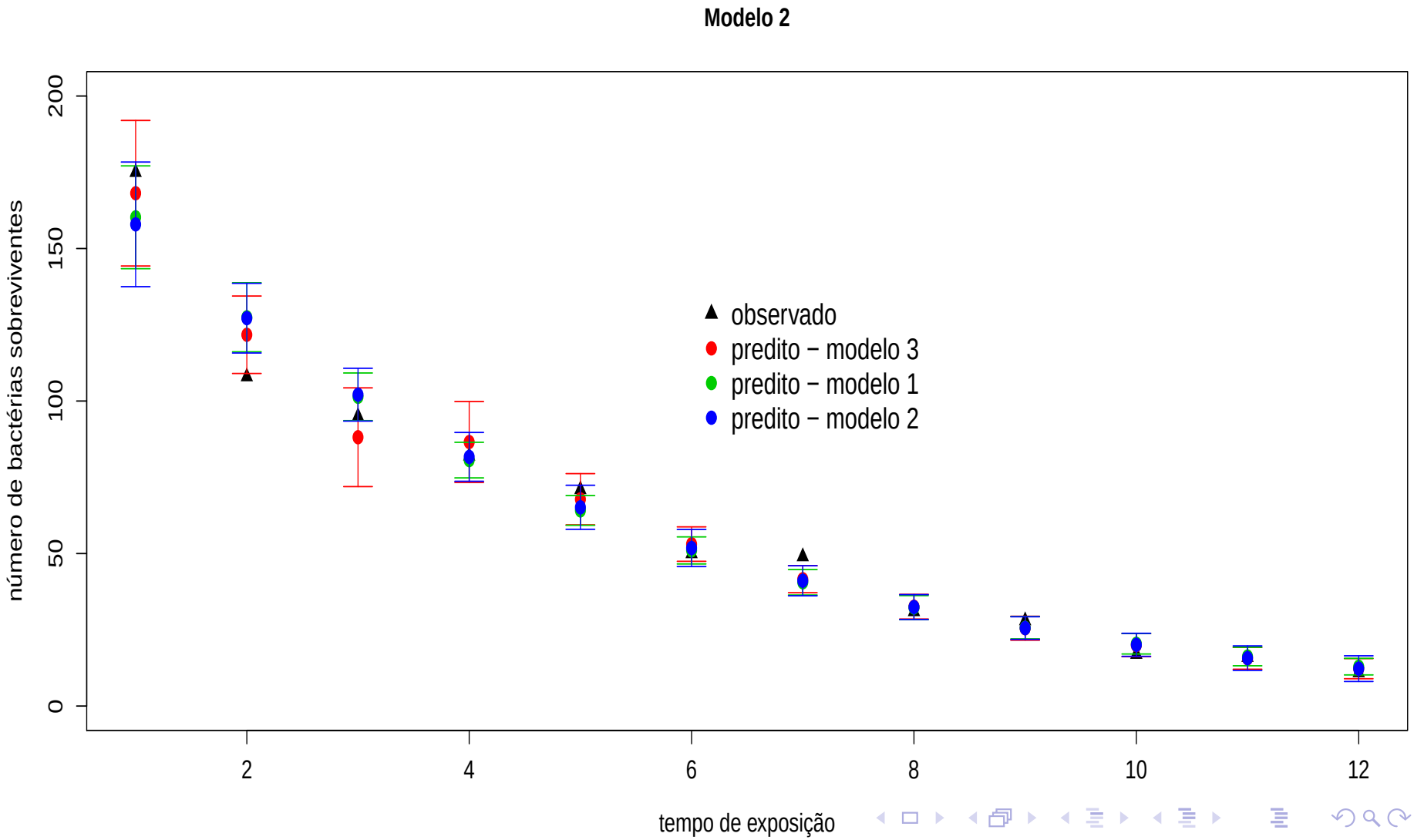
Par.	Est.	EP	IC(95%)	Estat. Z_t	p-valor
β_{01}	3,35	0,31	[2,74 ; 3,95]	10,84	$<$ 0,0001
β_{11}	-0,32	0,06	[-0,45 ; -0,20]	-5,00	$<$ 0,0001
β_{02}	3,85	0,05	[3,75 ; 3,95]	72,54	$<$ 0,0001
β_{12}	-0,24	0,02	[-0,29 ; -0,20]	-10,60	$<$ 0,0001

Estimativas pontuais e intervalares: modelo 3





Valores observados e preditos : modelos 1, 2 e 3



Comentários

- Alternativa: manter a estrutura geral do modelo substituindo-se o preditor linear (η_i) por um preditor não linear (modelos não lineares generalizados) e/ou considerar efeitos aleatórios. Outra alternativa: estrutura de regressão não paramétrica.

Comentários

- De acordo com o modelo, há uma diminuição (percentual) significativa no número de bactérias da ordem de $exp(-0,229) \approx 0,795 \equiv 79,5\% \approx 80,0\%$, para o aumento em 1 minuto no tempo de exposição. Espera-se que tal diminuição oscile entre $IC(e^{-\beta_1}, 95\%) = [0,775; 0,815] \equiv [77,5\%; 81,5\%]$ (lembrando que esse intervalo de confiança é assintótico).

Exemplo 2: 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).
- Questão de interesse: a imposição dos limites de velocidade levou à redução do número de acidentes?

Modelo

- Considere (i =1, ano de 1961, i=2, ano de 1962). Lembrando que: 1961 (sem limite de velocidade) e 1962 (com limite de velocidade), temos

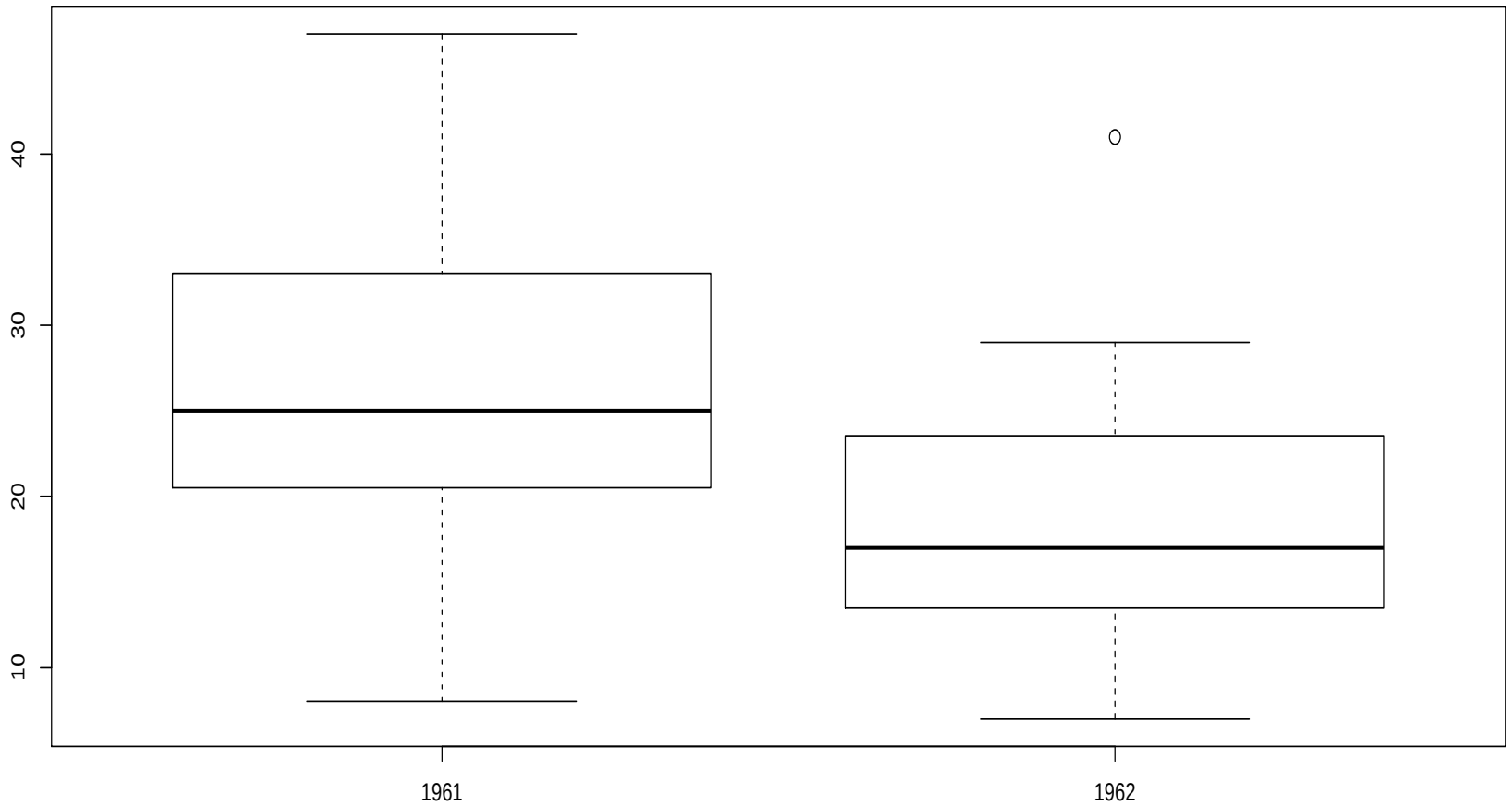
$$\begin{aligned} Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_i), i = 1, 2; j = 1, ..., 43 \\ \ln \mu_i &= \mu + \alpha_i, \alpha_1 = 0 \end{aligned}$$

em que $\boldsymbol{\beta} = (\mu, \alpha_2)'$. Assim, tem-se que $\mathcal{E}(Y_{ij}) = e^{\mu+\alpha_i}$. Além disso, e^{α_2} é o incremento multiplicativo (positivo ou negativo) da média do ano de 1962 em relação à média do ano de 1961 ($\mu_2 = \mu_1 e^{\alpha_2}$).

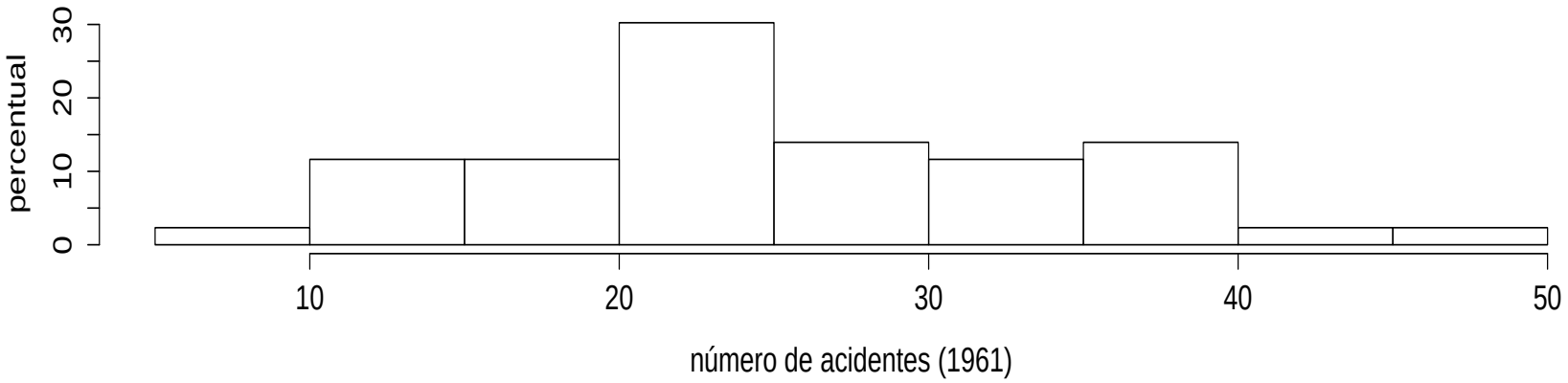
Medidas Resumo

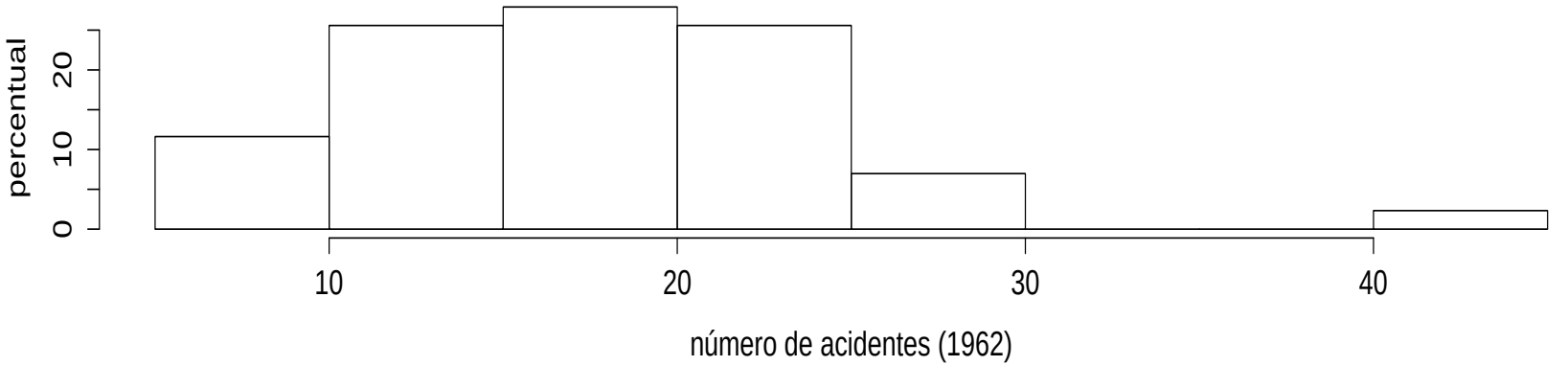
Ano	Média	Var.	DP	CV(%)	Mín.	Med.	Máx.
1961	26,05	82,66	9,09	34,91	8,00	25,00	47,00
1962	18,05	44,71	6,69	37,05	7,00	17,00	41,00

Boxplots do número de acidentes por ano

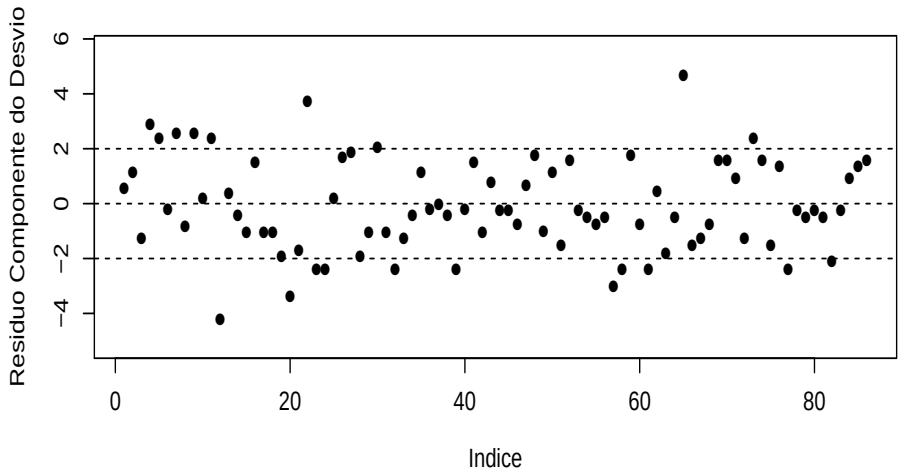


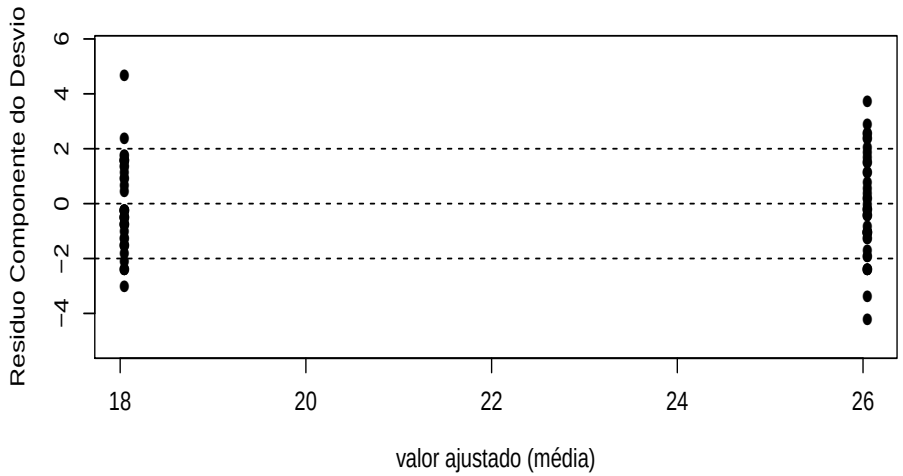
Histogramas do número de acidentes por ano

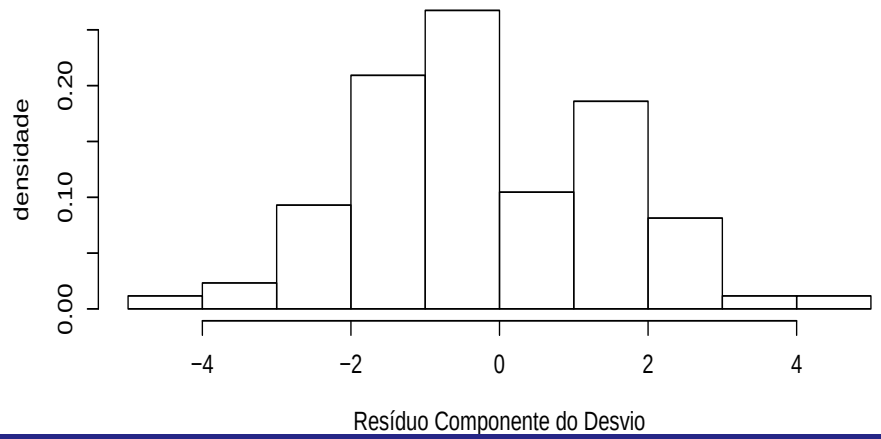


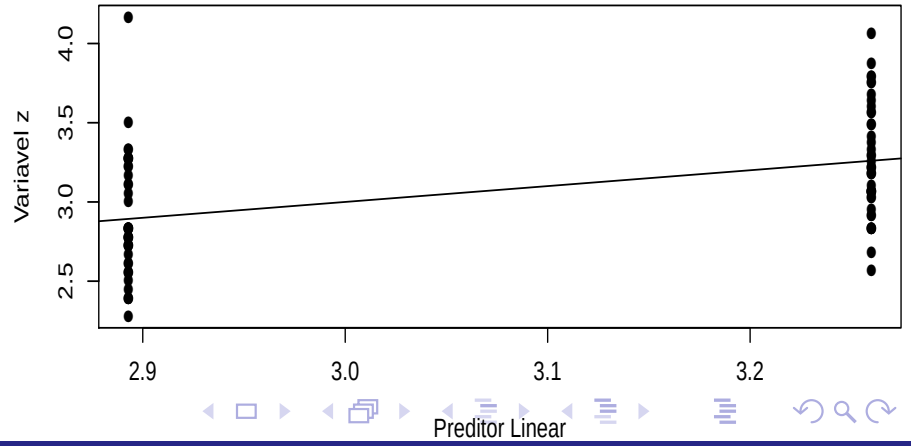


Gráficos de diagnóstico

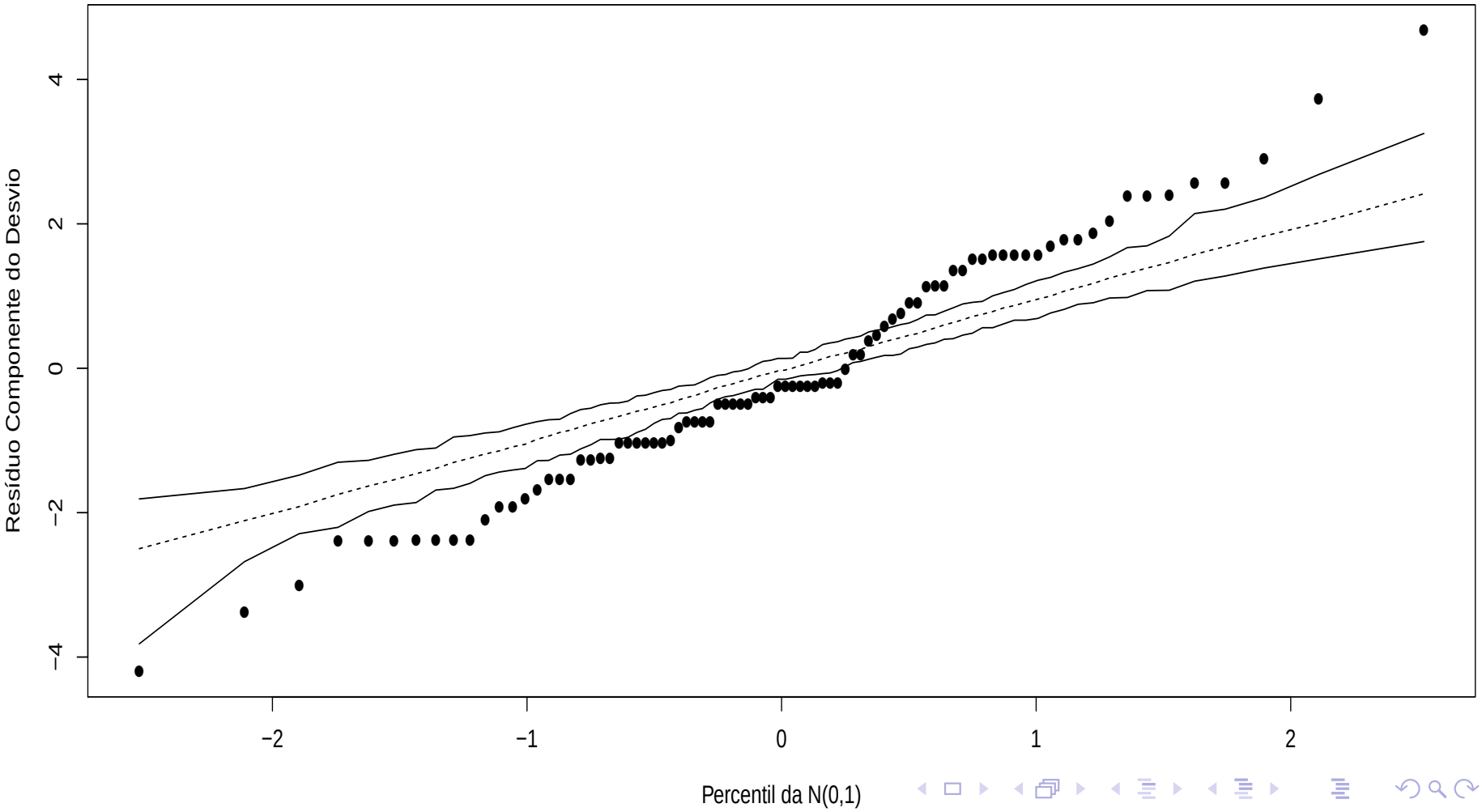




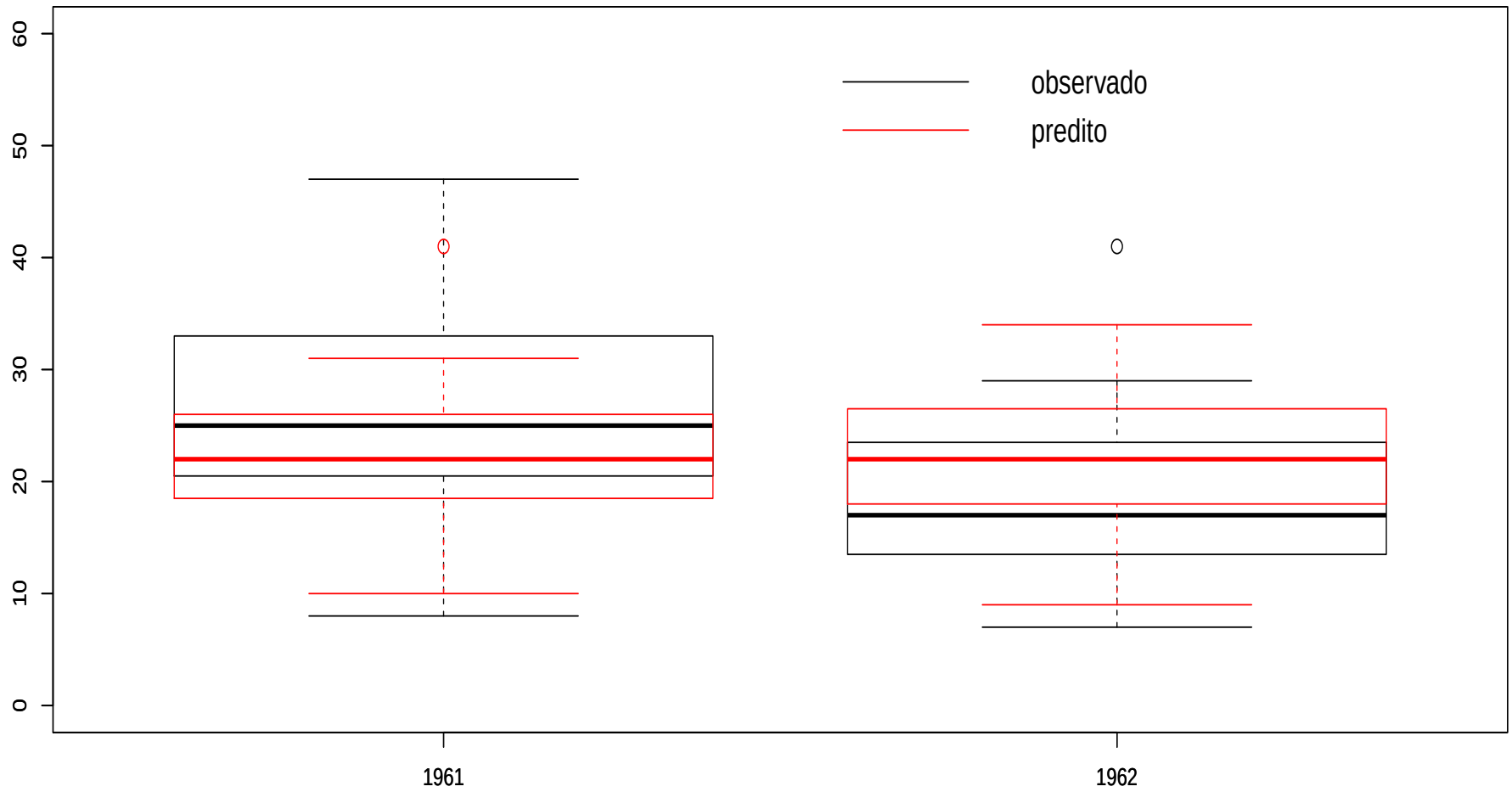




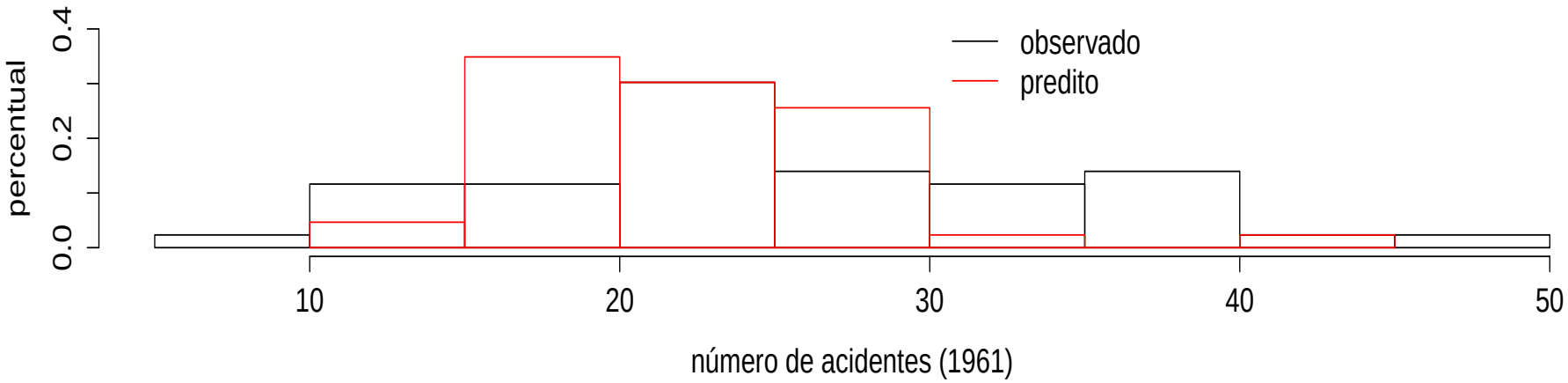
Gráficos de envelope

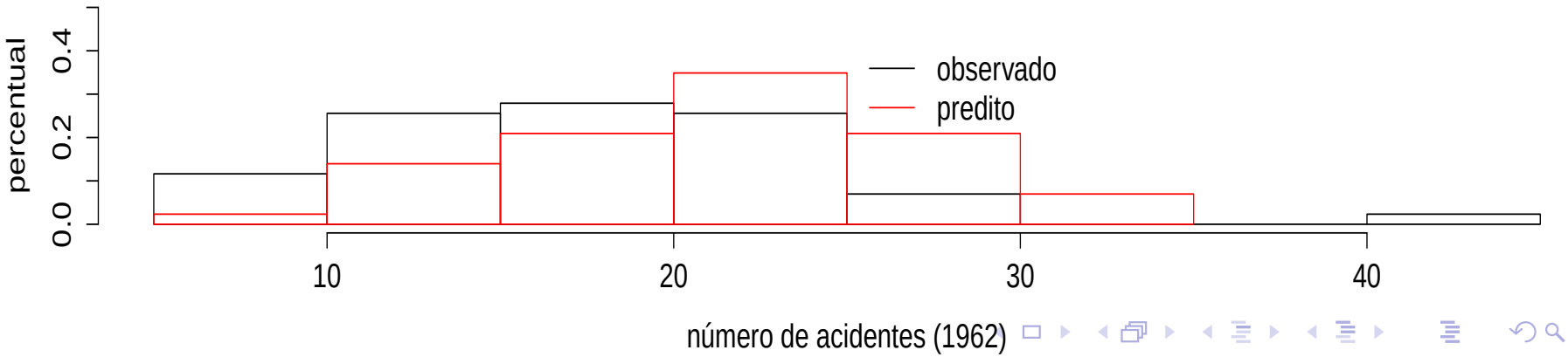


Distribuições preditas e observadas (boxplot)



Distribuições preditas e observadas (histograma)





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

$D(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}) = 235,17$ (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.

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.

Comentários

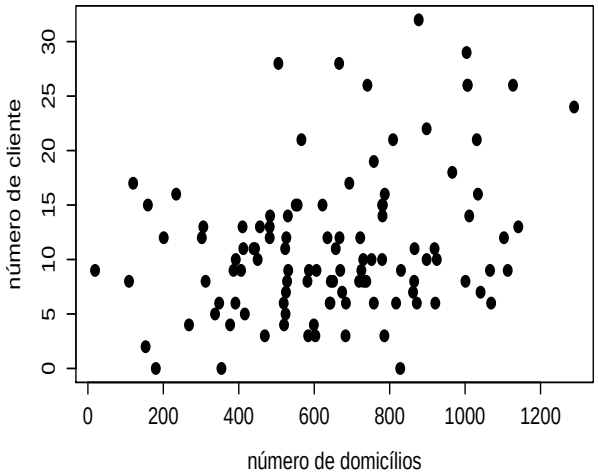
- A análise de diagnóstico indicou que o modelo não se ajustou bem aos dados, portanto ele não pode ser utilizado para analisá-los.
- Isso ocorreu, possivelmente, devido à um problema de superdispersão.
- Alternativas de análise: modelo de regressão de Poisson de efeitos aleatórios (binomial-negativo), modelos livre de distribuição (sem supor alguma distribuição específica para a variável resposta) que contemple a superdispersão, outros modelos de regressão hetorocedásticos (modelos de superdispersão).

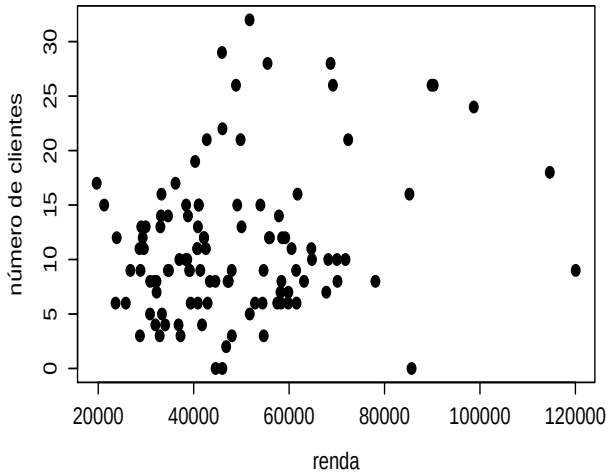
Exemplo 15: perfil dos clientes de uma loja

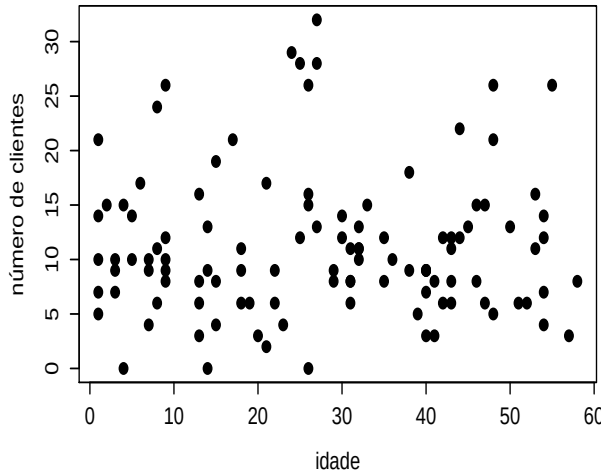
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável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

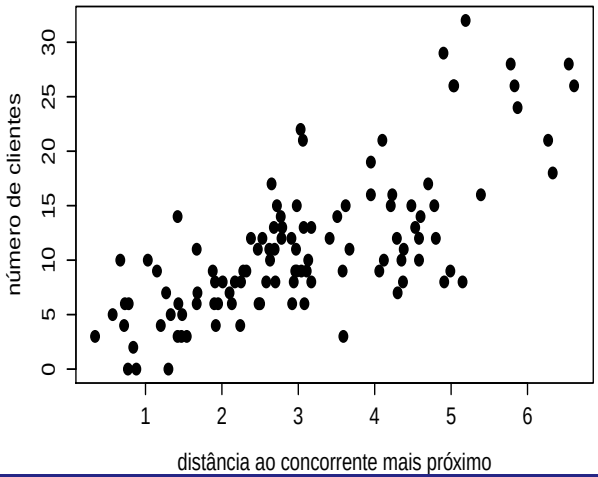
Mapa de Fortaleza (ilustração)

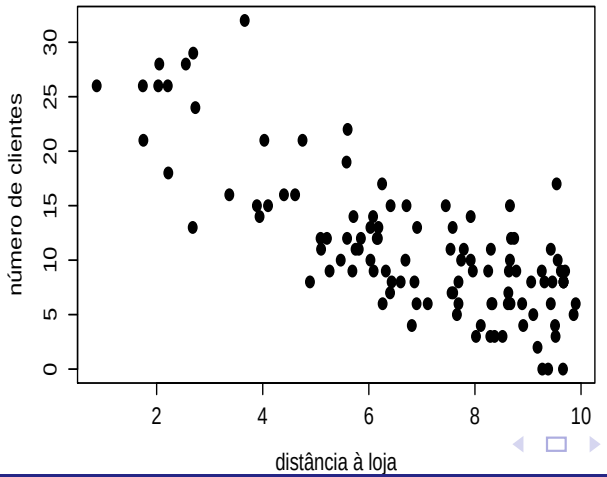
Gráficos de dispersão











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

Modelo (completo)

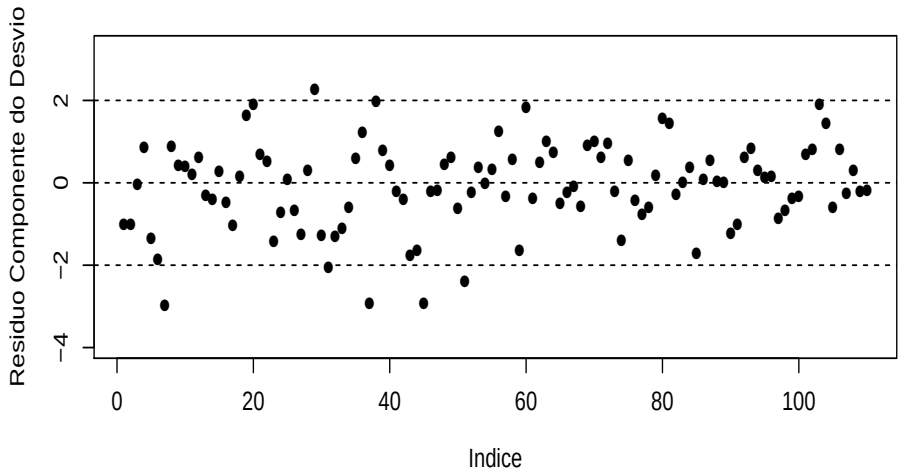
$$
\begin{aligned}
Y_i &\overset{ind.}{\sim} \text{Poisson}(\mu_i) \\
\ln(\mu_i) &= \beta_0 + \beta_1\left(\frac{x_{1i}-\overline{x}_1}{s_1}\right) + \beta_2\left(\frac{x_{2i}-\overline{x}_2}{s_2}\right) + \beta_3\left(\frac{x_{3i}-\overline{x}_3}{s_3}\right) + \\
&+ \beta_4\left(\frac{x_{4i}-\overline{x}_4}{s_4}\right)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

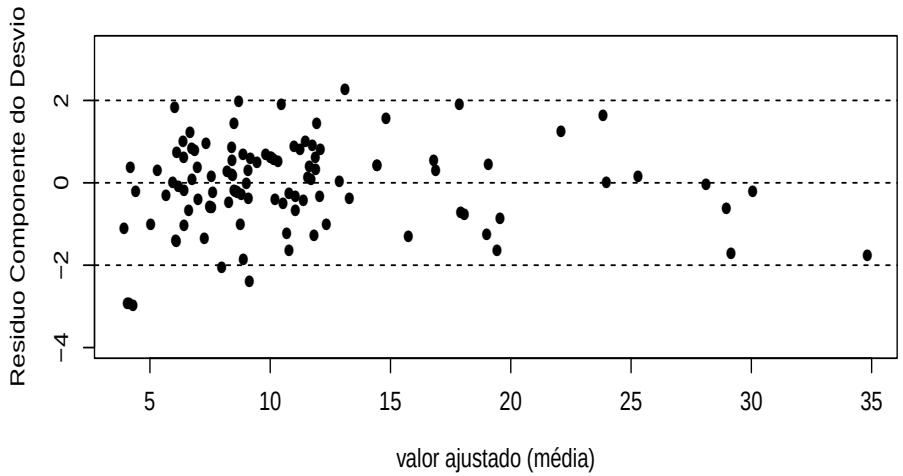
- x_{ji} : valor da variável explicativa j, $j = 1, 2, ..., 5$, associada à área i, $\overline{x}_j = \frac{1}{110}\sum_{i=1}^{110} x_{ji}$, e $s_j = \frac{\sum_{i=1}^{n}(x_{ji}-\overline{x}_j)^2}{109}$ $j = 1, 2, ..., 5$.

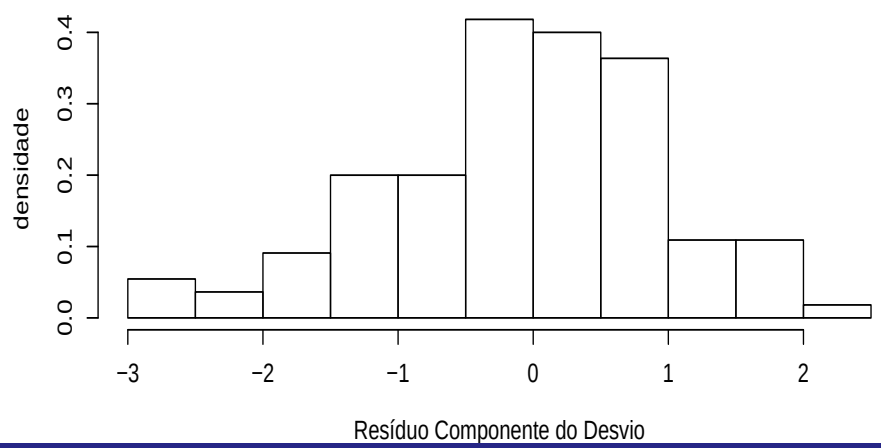
Modelo (completo)

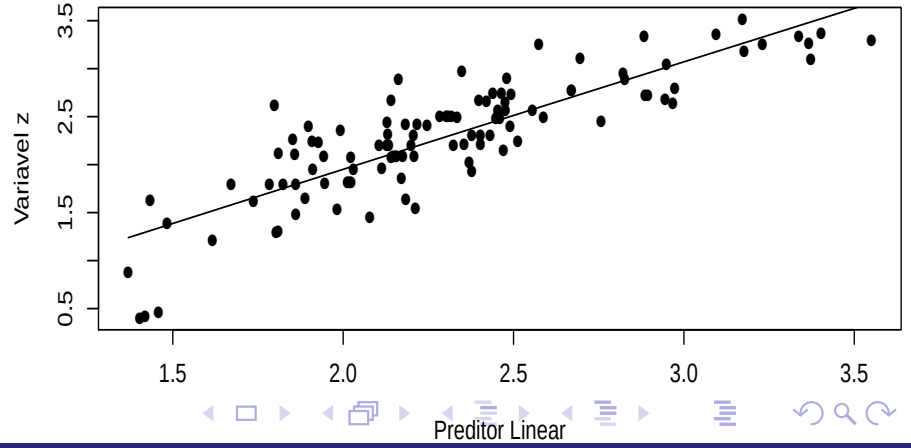
- e^{β_0} : número esperado de clientes para domicílios localizados em áreas com valor médio para cada uma das covariáveis.
- e^{β_j / s_j} : incremento (positivo ou negativo) no valor esperado do número de clientes, para o aumento em uma unidade no valor da covariável j, mantendo-se todas as outras 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

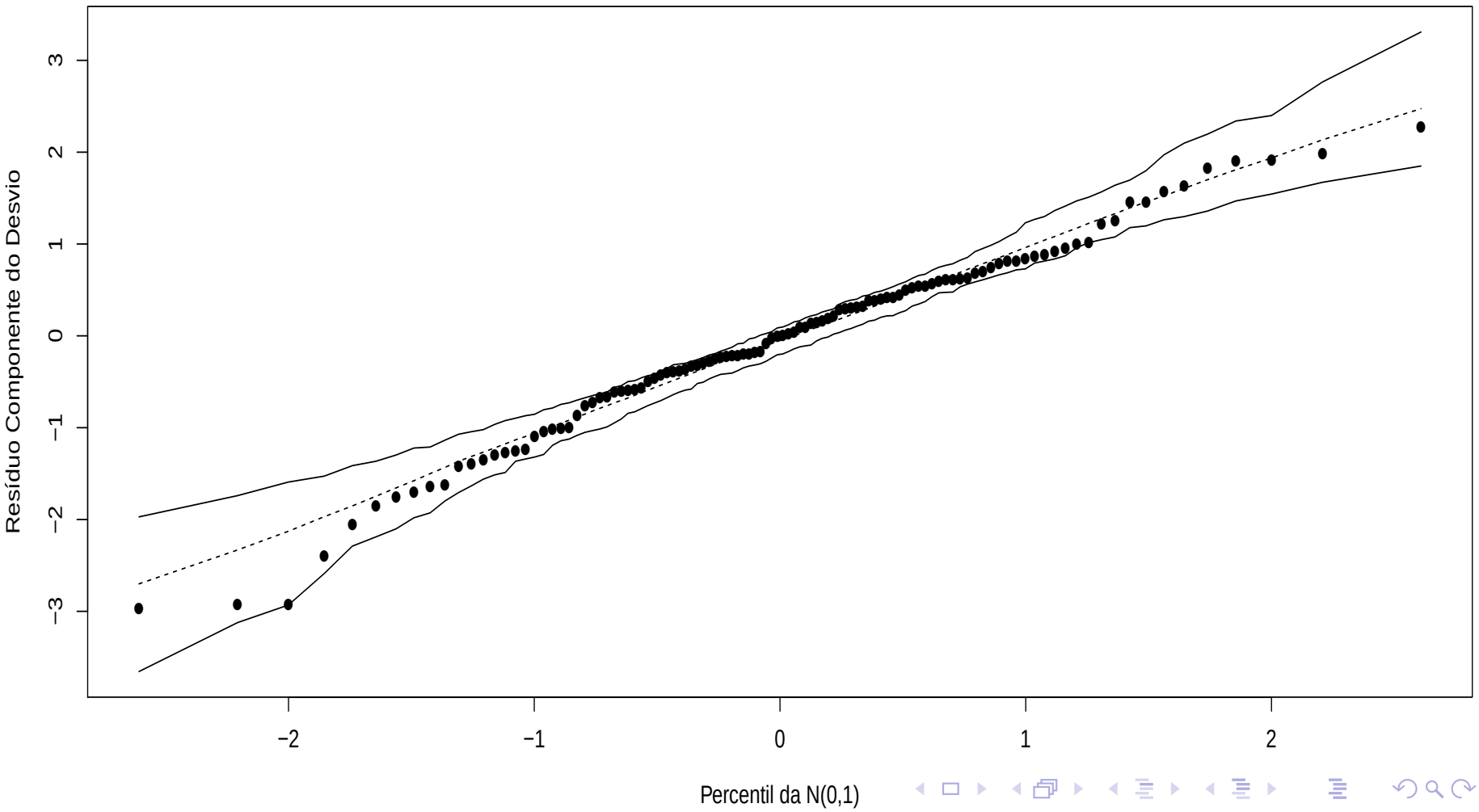








Gráficos de envelope

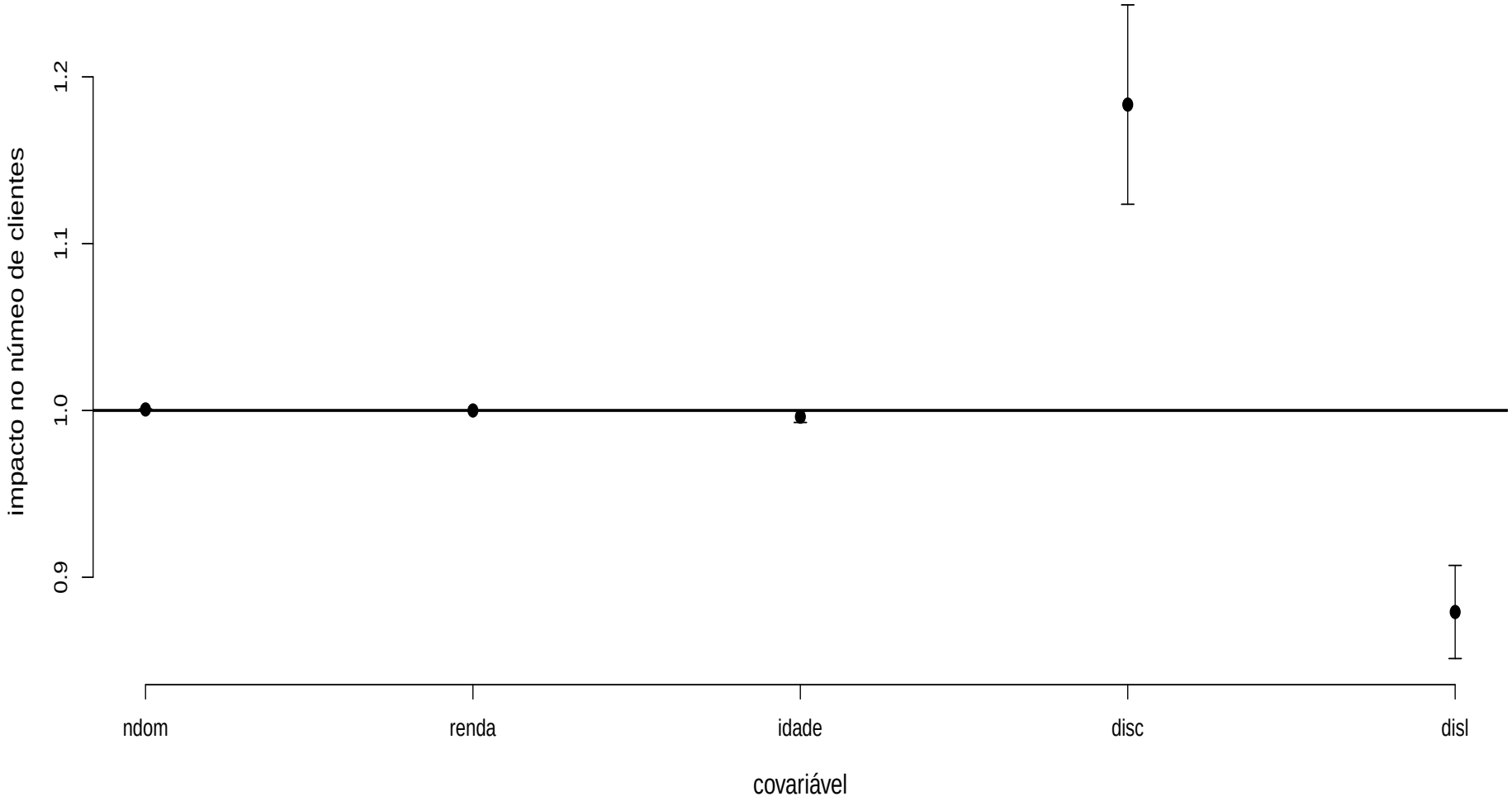


Estimativas dos parâmetros

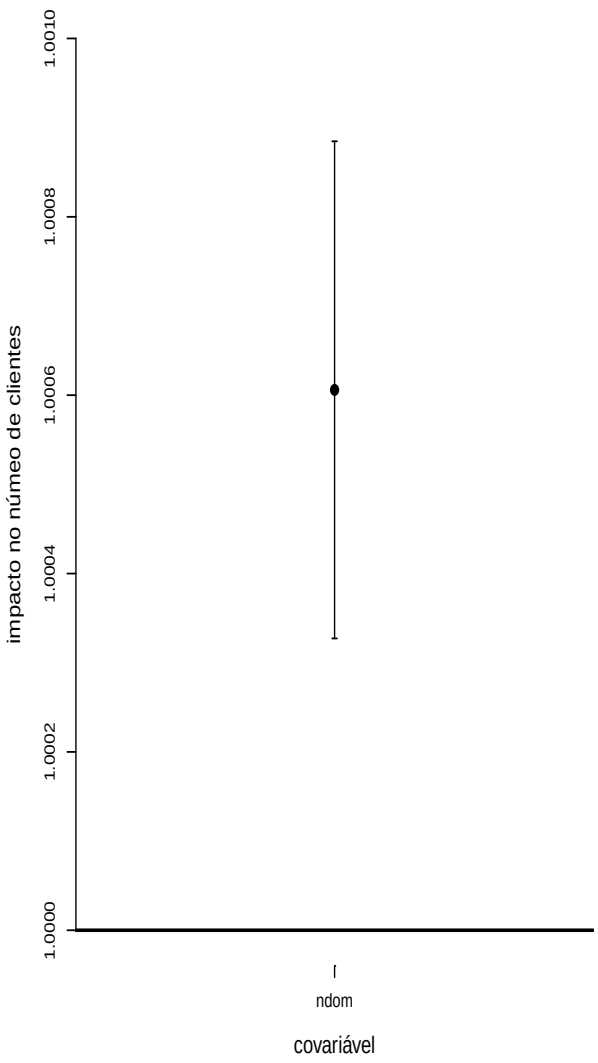
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	2,30	0,03	[2,24 ; 2,36]	72,92	$<$ 0,0001
β_1	0,16	0,04	[0,09 ; 0,23]	4,26	$<$ 0,0001
β_2	-0,22	0,04	[-0,29 ; -0,14]	-5,53	$<$ 0,0001
β_3	-0,062	0,030	[-0,120 ; -0,003]	-2,091	0,0365
β_4	0,25	0,04	[0,18 ; 0,33]	6,53	$<$ 0,0001
β_5	-0,30	0,04	[-0,37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

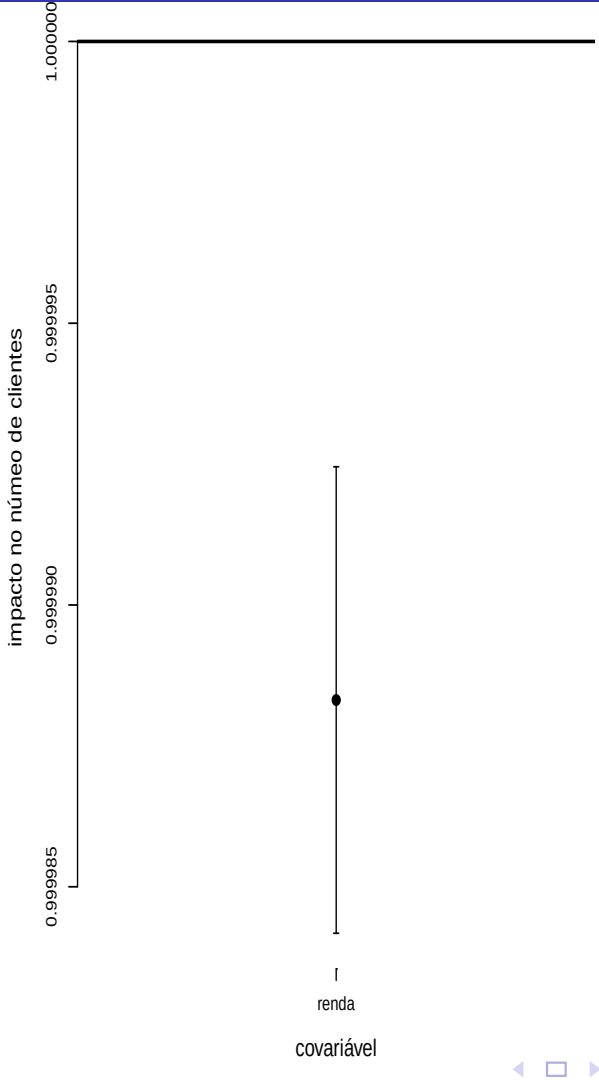
$D(\boldsymbol{y}, \widetilde{\boldsymbol{\mu}}) = 114,95$ (p = 0,2170) (considerando a aproximação pela distribuição $\chi^2_{(104)}$ adequada), o que indica que o modelo se adequou bem aos dados.

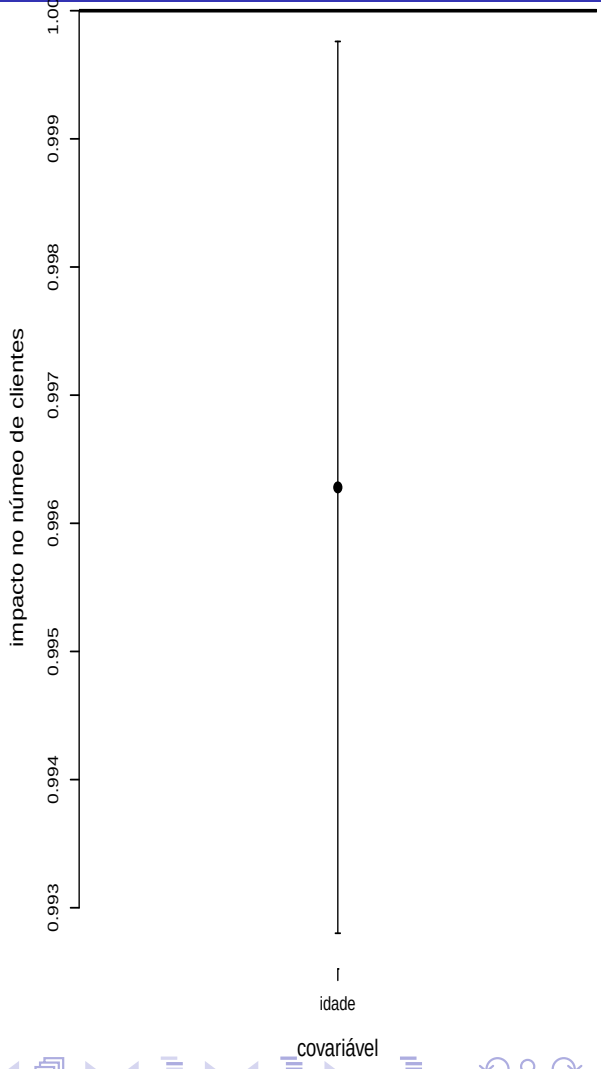
Estimativa do impacto que cada covariável (e^{β_j / s_j})



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



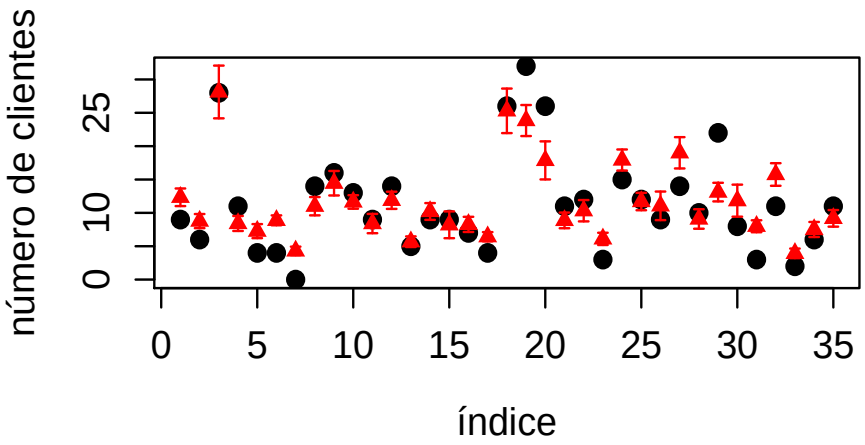


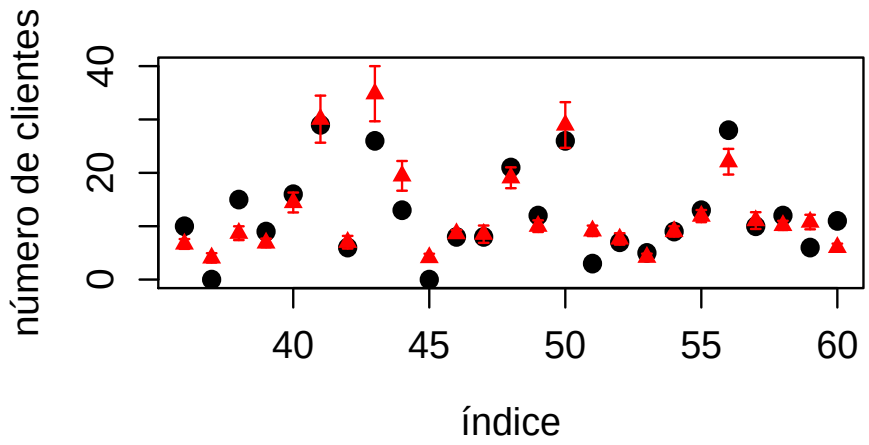


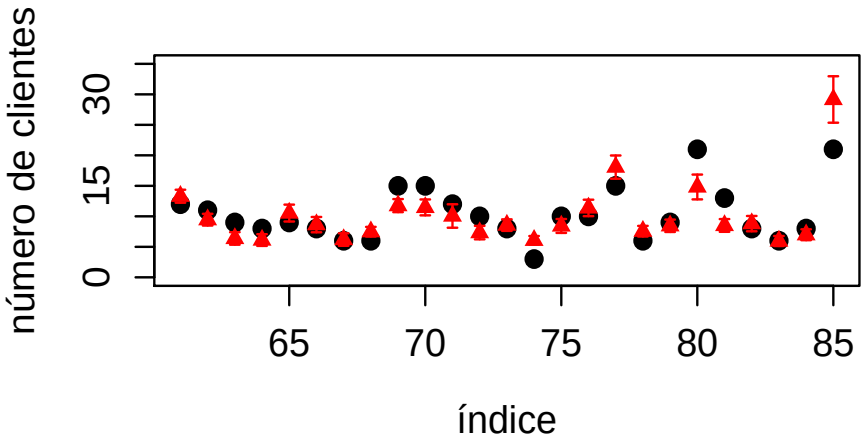
Mais sobre a escolha do modelo

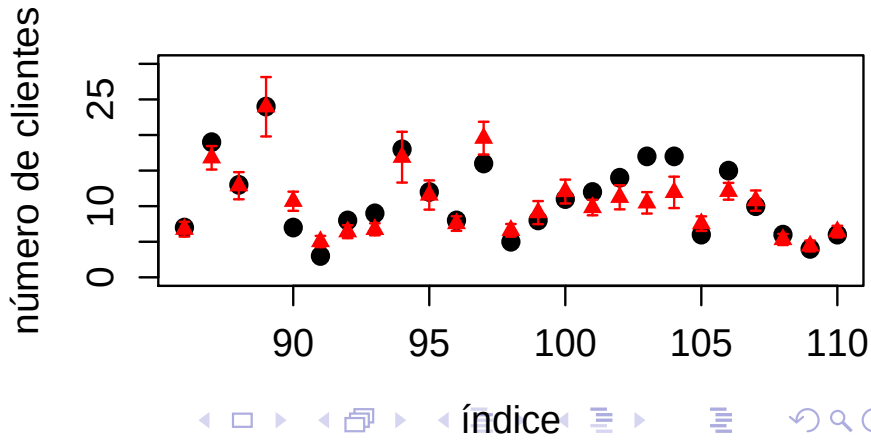
- A aplicação da metodologia stepwise, começando com o modelo só com o intercepto ou começando com o modelo completo, indicou, em ambos os casos, que todas as variáveis são significativas.
- AIC : modelo completo (571,02) sem a covariável idade (573,40).
- BIC : modelo completo (587,23) sem a covariável idade (586,91).

Valores preditos x observados

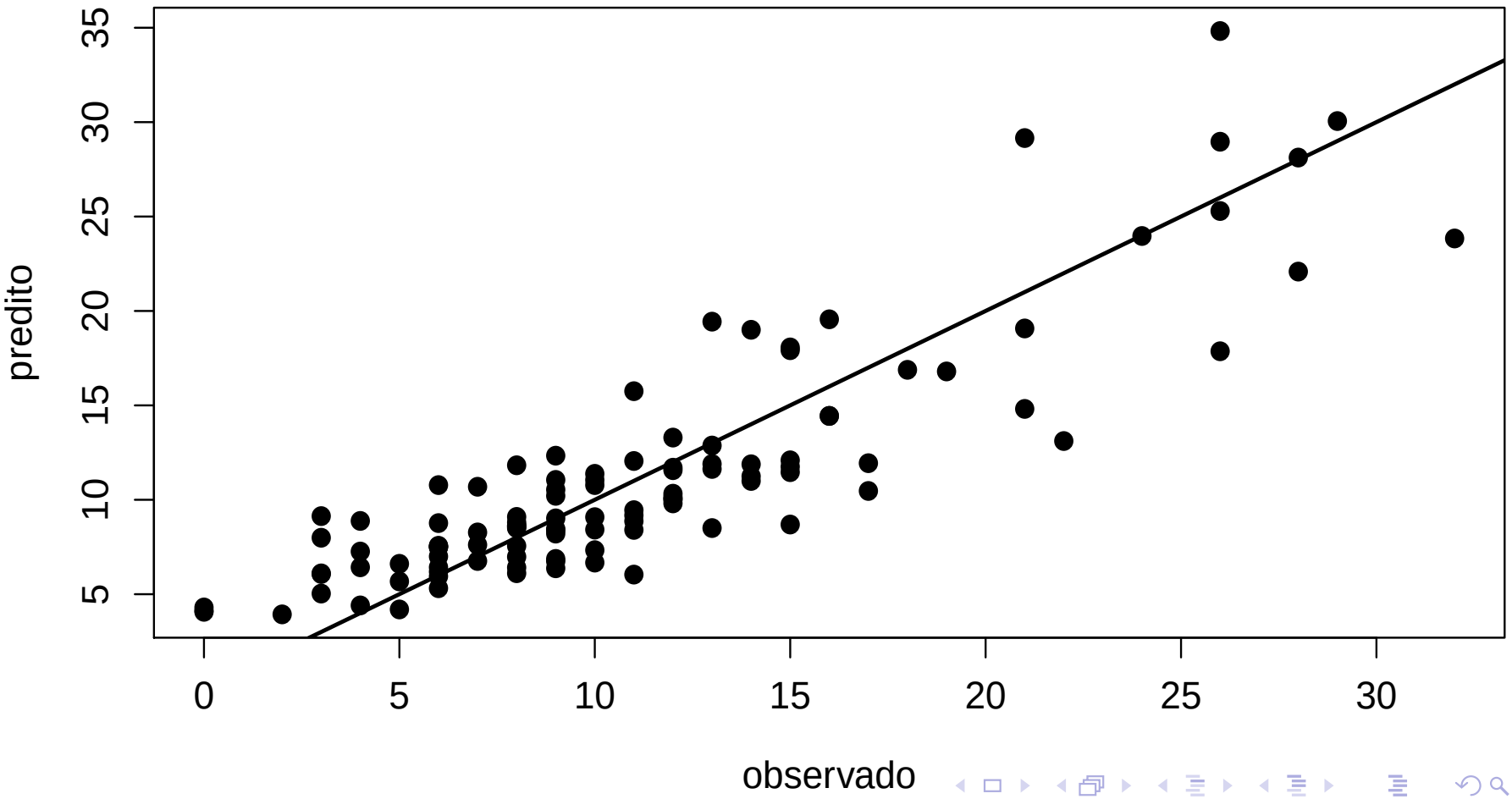








Valores preditos x observados

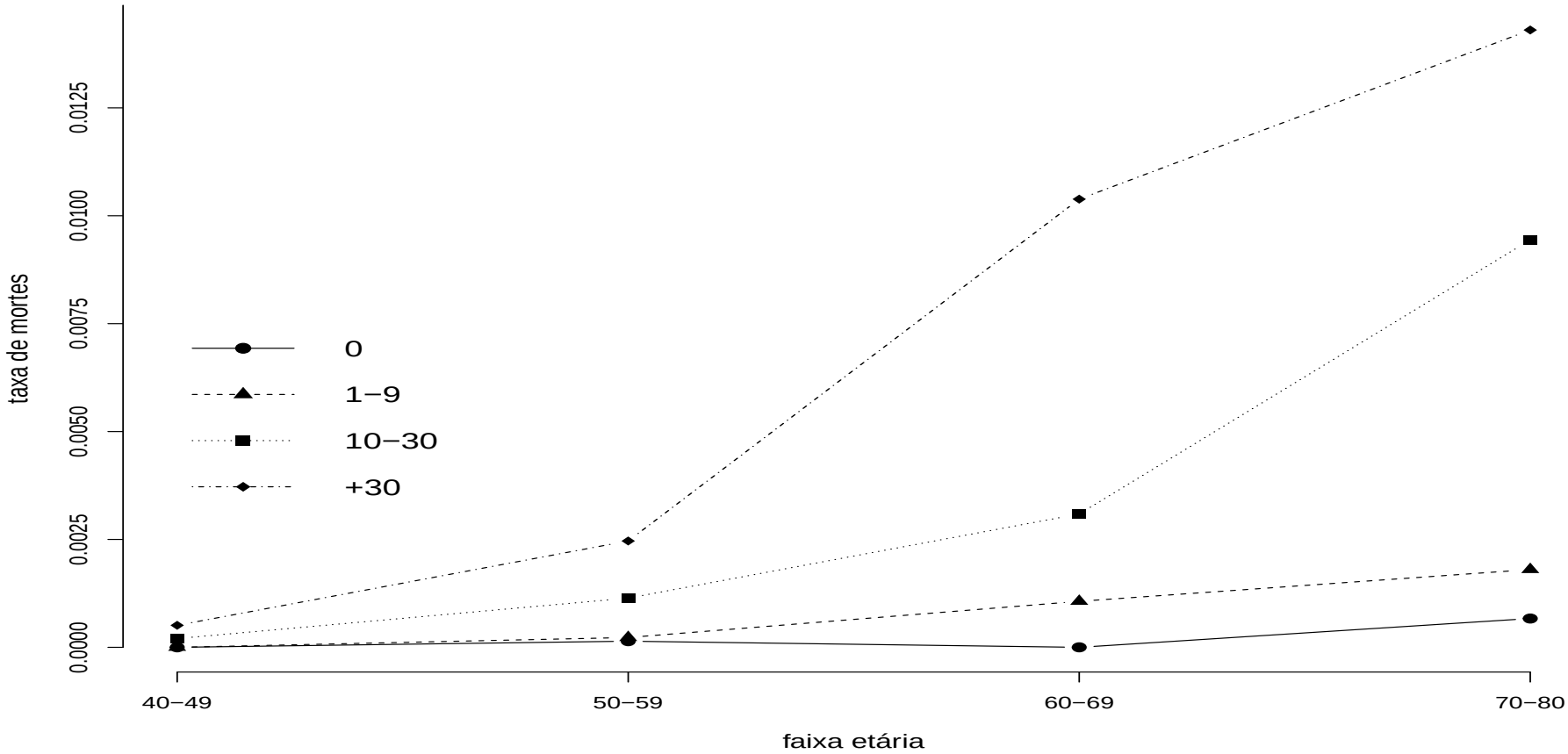


Exemplo 16: Mortalidade de câncer de pulmão

- A Tabela a seguir resume os resultados de um estudo de seguimento em que doutores Britânicos foram acompanhados durante a década de 50 e observado, em particular, a ocorrência de mortes por câncer de pulmão segundo o consumo médio diário de cigarros e a faixa etária por pessoas/anos (p-anos).
- Objetivo: avaliar como a faixa etária e o consumo médio diário de cigarros na mortalidade afetam a mortalidade por câncer de pulmão.

Consumo médio diário de cigarros		Faixa etária			
		40-49	50-59	60-69	70-80
0	morte	0	3	0	3
	p-anos	33679	21131,50	10599	4495,50
1-9	morte	0	1	3	3
	p-anos	6002,50	4396	2813,50	1664,50
10-30	morte	7	29	41	45
	p-anos	34414,50	25429	13271	4765,50
+30	morte	3	16	36	11
	p-anos	5881	6493,50	3466,50	769

Gráfico de perfis observados



Modelo 1

- Como levar em consideração o número de pessoas/anos na análise?

$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij}/t_{ij}) \\
\ln(\mu_{ij}/t_{ij}) &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
\Leftrightarrow \ln(\mu_{ij}) &= \ln(t_{ij}) + \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
\beta_i &= \gamma_j = (\beta\gamma)_{ij} = 0 \ i = j = 1
\end{aligned}
$$

- Y_{ij}: número de mortes para o i-ésimo nível de consumo e j-ésima faixa etária.
- μ_{ij} : taxa média de mortes por "unidade de tempo" (t_{ij}) para o consumo i e faixa etária j.
- β_i (consumo), γ_j (faixa etária).

Modelo 1

- Neste caso o termo $\ln(t_{ij})$ está sendo tratado como um "offset" (covariável cujo coeficiente é conhecido e igual a 1).
- Estimação: seja $\ln(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + \sum_{j=1}^{q} t_{ij}$ (em que t_j são "offsets"), então devemos maximizar

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i(\boldsymbol{X}_i'\boldsymbol{\beta} + \sum_{j=1}^{q} t_{ij}) - e^{\boldsymbol{X}_i'\boldsymbol{\beta}+\sum_{j=1}^{q} t_{ij}}]$$

com relação à $\boldsymbol{\beta}$.

- Podemos testar se os termos t_j podem, de fato, ser considerados como "offsets", ajustando um modelo com $\boldsymbol{X}_i'\beta + \sum_{j=1}^{q} \delta_j t_{ij}$ e testar se $H_0 : \delta_j = 1$ vs $H_1 : \delta_j \neq 1$, j=1,2,...,q.

Modelo 1

- As médias induzidas pelo modelo são:

$$
\begin{aligned}
\mu_{11} &= e^{t_{11}+\alpha}; \mu_{12} = e^{t_{12}+\alpha+\gamma_2}; \mu_{13} = e^{t_{13}+\alpha+\gamma_3}; \mu_{14} = e^{t_{14}+\alpha+\gamma_4} \\
\mu_{21} &= e^{t_{21}+\alpha+\beta_2}; \mu_{22} = e^{t_{22}+\alpha+\beta_2+\gamma_2+(\beta\gamma)_{22}} \\
\mu_{23} &= e^{t_{23}+\alpha+\beta_2+\gamma_3+(\beta\gamma)_{23}}; \mu_{24} = e^{t_{24}+\alpha+\beta_2+\gamma_4+(\beta\gamma)_{24}} \\
\mu_{31} &= e^{t_{31}+\alpha+\beta_3}; \mu_{32} = e^{t_{32}+\alpha+\beta_3+\gamma_2+(\beta\gamma)_{32}} \\
\mu_{33} &= e^{t_{33}+\alpha+\beta_3+\gamma_3+(\beta\gamma)_{33}}; \mu_{34} = e^{t_{34}+\alpha+\beta_3+\gamma_4+(\beta\gamma)_{34}} \\
\mu_{41} &= e^{t_{41}+\alpha+\beta_4}; \mu_{42} = e^{t_{42}+\alpha+\beta_4+\gamma_2+(\beta\gamma)_{42}} \\
\mu_{43} &= e^{t_{43}+\alpha+\beta_4+\gamma_3+(\beta\gamma)_{43}}; \mu_{44} = e^{t_{44}+\alpha+\beta_4+\gamma_4+(\beta\gamma)_{44}}
\end{aligned}
$$

Modelo saturado: 16 observações e 16 parâmetros.

Par.	Estimativa	EP	IC(95%)	Estat. Z	p-valor
α	-33,73	69653,80	[-136552,67 ; 136485,21]	>-0,01	> 0,9999
β_2	1,72	98505,35	[-193065,21 ; 193068,66]	< 0,01	> 0,9999
β_3	25,23	69653,80	[-136493,71 ; 136544,17]	< 0,01	> 0,9999
β_4	26,15	69653,80	[-136492,79 ; 136545,09]	< 0,01	> 0,9999
γ_2	24,87	69653,80	[-136494,07 ; 136543,81]	< 0,01	> 0,9999
γ_3	1,16	98505,35	[-193065,78 ; 193068,09]	< 0,01	> 0,9999
γ_4	26,41	69653,80	[-136492,53 ; 136545,36]	< 0,01	> 0,9999
$(\alpha\beta)_{22}$	-1,25	98505,35	[-193068,19 ; 193065,68]	>-0,01	> 0,9999
$(\alpha\beta)_{23}$	-23,14	69653,80	[-136542,08 ; 136495,80]	>-0,01	> 0,9999
$(\alpha\beta)_{24}$	-23,29	69653,80	[-136542,23 ; 136495,65]	>-0,01	> 0,9999
$(\alpha\beta)_{32}$	24,00	120643,92	[-236433,74 ; 236481,74]	< 0,01	> 0,9999
$(\alpha\beta)_{33}$	1,56	98505,35	[-193065,37 ; 193068,50]	< 0,01	> 0,9999
$(\alpha\beta)_{34}$	1,86	98505,35	[-193065,08 ; 193068,80]	< 0,01	> 0,9999
$(\alpha\beta)_{42}$	-0,73	98505,35	[-193067,67 ; 193066,21]	>-0,01	> 0,9999
$(\alpha\beta)_{43}$	-22,58	69653,80	[-136541,52 ; 136496,36]	>-0,01	> 0,9999
$(\alpha\beta)_{44}$	-23,08	69653,80	[-136542,02 ; 136495,86]	>-0,01	> 0,9999

Comentários

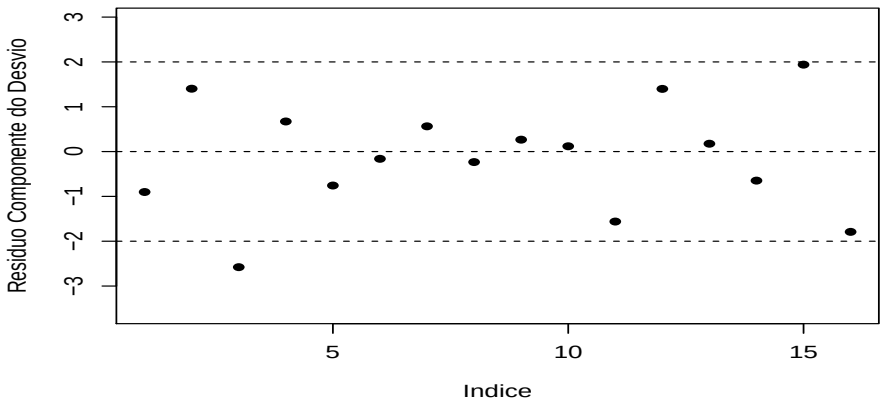
- Não foi possível produzir os gráficos de diagnóstico. Em geral, nos modelos saturados (n=p) a matrix "H" é singular.
- O teste de Wald para as hipóteses $H_0 : \boldsymbol{C\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{C\beta} \neq \mathbf{0}$ para testar a ausência vs a presença de interação ($H_0 : (\alpha\beta)_{ij} = \forall i \geq 2, j \geq 2$ vs H_1 : há pelo menos uma diferença) apresentou os seguites resultados $q_c = 4,70 (p = 0,8598)$, o que indica a ausência de interação.
- Vamos ajustar um modelo sem interação.

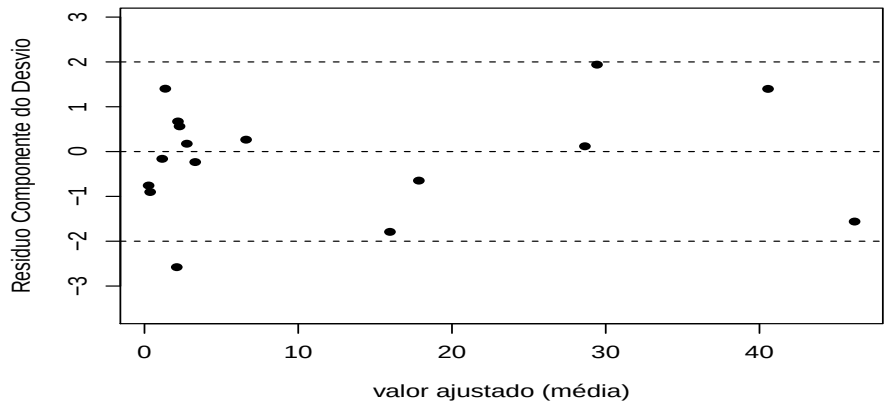
Modelo 2

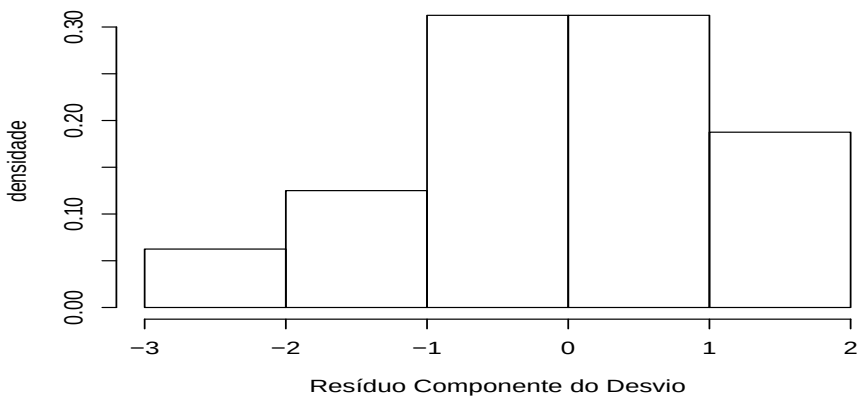
$$
\begin{aligned}
Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_{ij} t_{ij}) \\
\ln(\mu_{ij} t_{ij}) &= \alpha + \beta_i + \gamma_j + (\beta\gamma)_{ij} \\
\Leftrightarrow \ln(\mu_{ij}) &= \ln(t_{ij}) + \alpha + \beta_i + \gamma_j \\
\beta_i &= \gamma_j = (\beta\gamma)_{ij} = 0 \ \forall i, j
\end{aligned}
$$

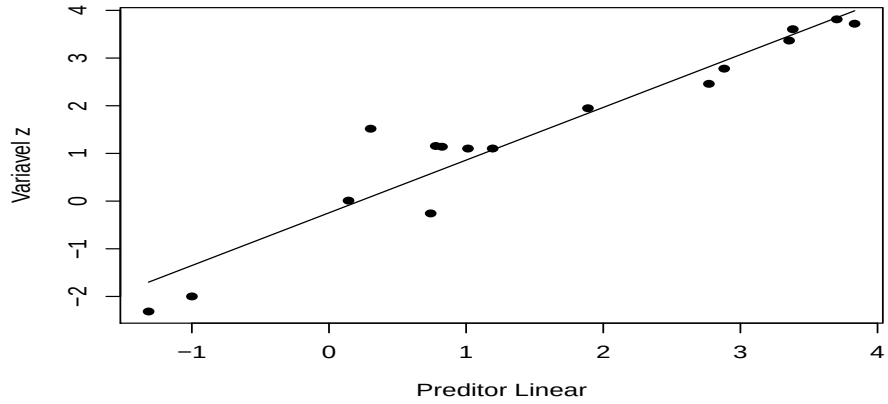
- Y_{ij}: número de mortes para o i-ésimo nível de consumo e j-ésima faixa etária.
- μ_{ij} : taxa média de mortes por "unidade de tempo" (t_{ij}) para o consumo i e faixa etária j.
- β_i (consumo), γ_j (faixa etária).

Gráficos de diagnóstico: modelo 2

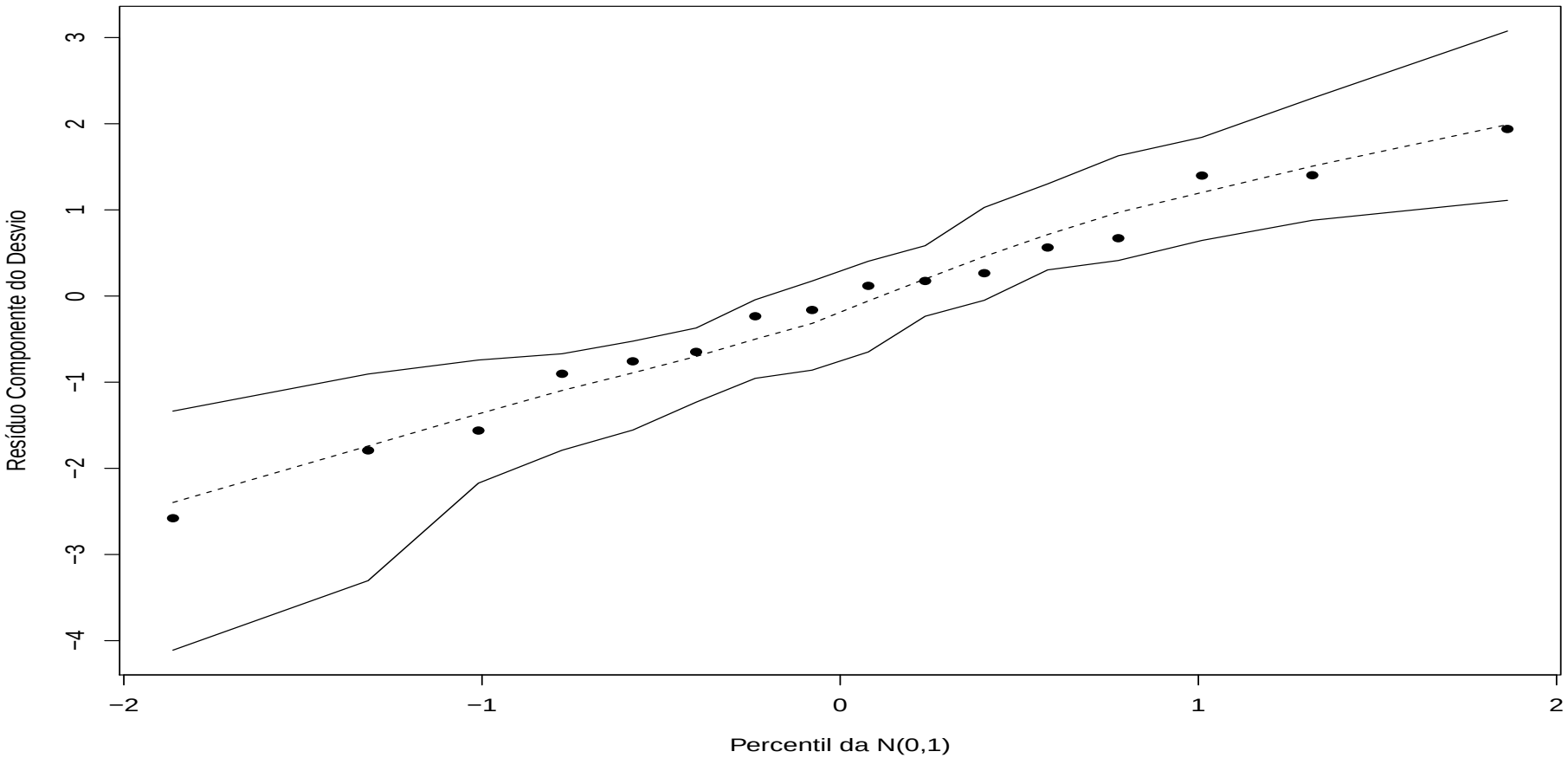






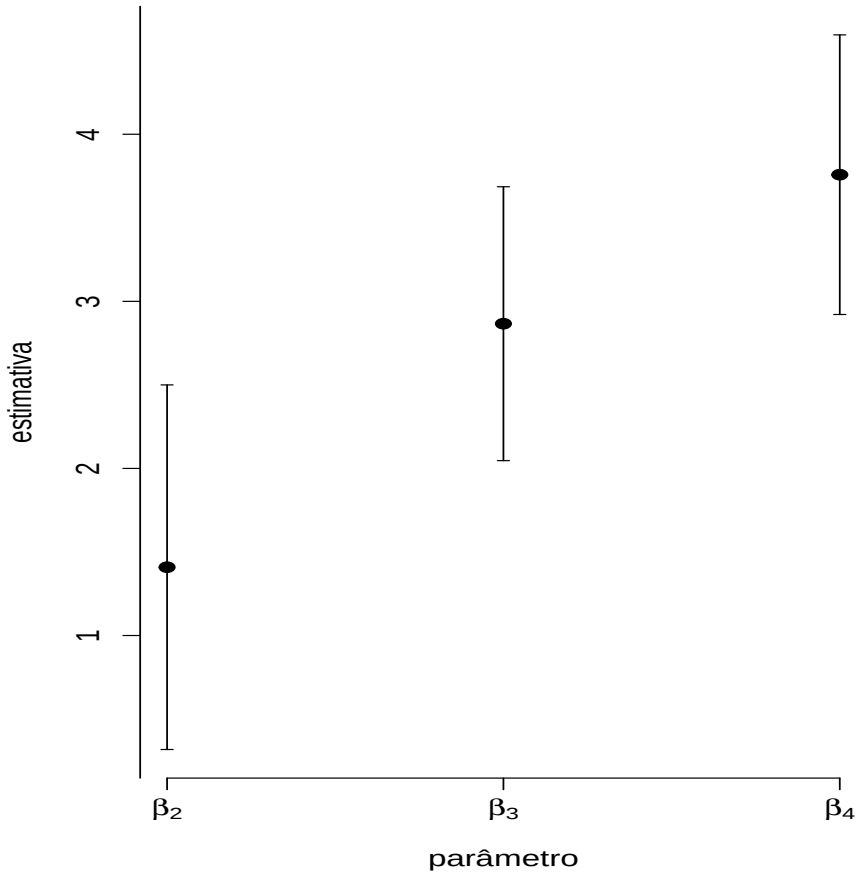


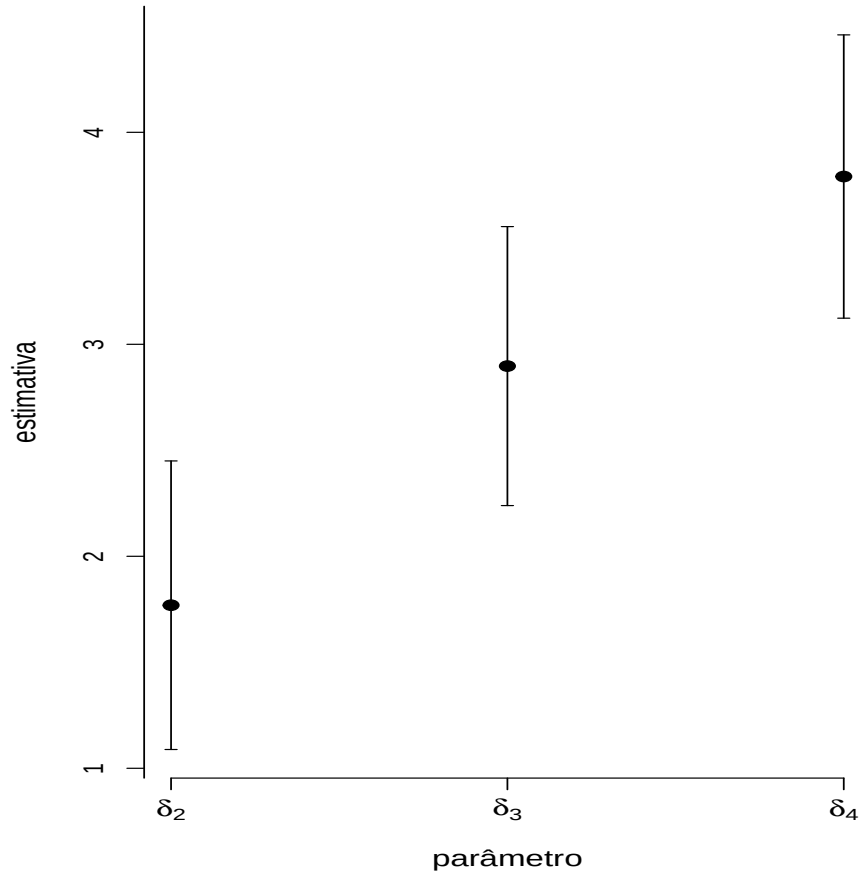
Gráficos de envelopes: modelo 2



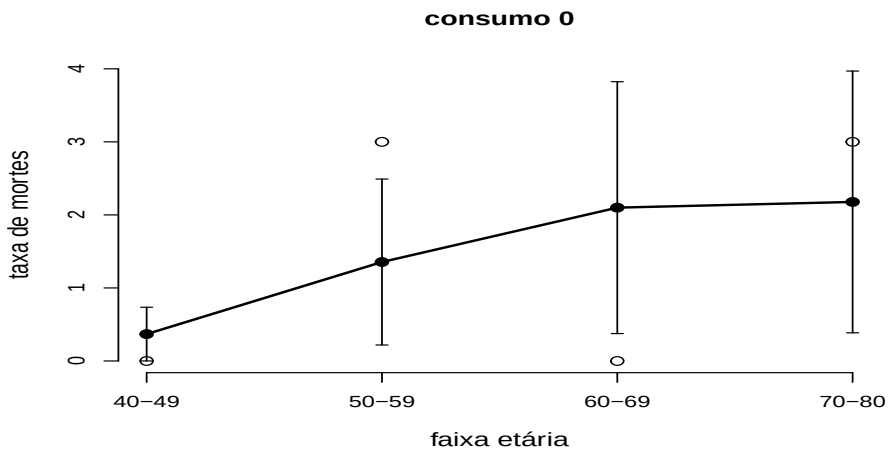
Par.	Estimativa	EP	IC(95%)	Estat. Z	p-valor
α	-11,42	0,51	[-12,42 ; -10,43]	-22,42	$<$ 0,0001
β_2	1,41	0,56	[0,32 ; 2,50]	2,53	0,0114
β_3	2,87	0,42	[2,05 ; 3,69]	6,85	$<$ 0,0001
β_4	3,76	0,43	[2,92 ; 4,59]	8,80	$<$ 0,0001
γ_2	1,77	0,35	[1,09 ; 2,45]	5,09	$<$ 0,0001
γ_3	2,90	0,34	[2,24 ; 3,56]	8,63	$<$ 0,0001
γ_4	3,79	0,34	[3,12 ; 4,46]	11,12	$<$ 0,0001

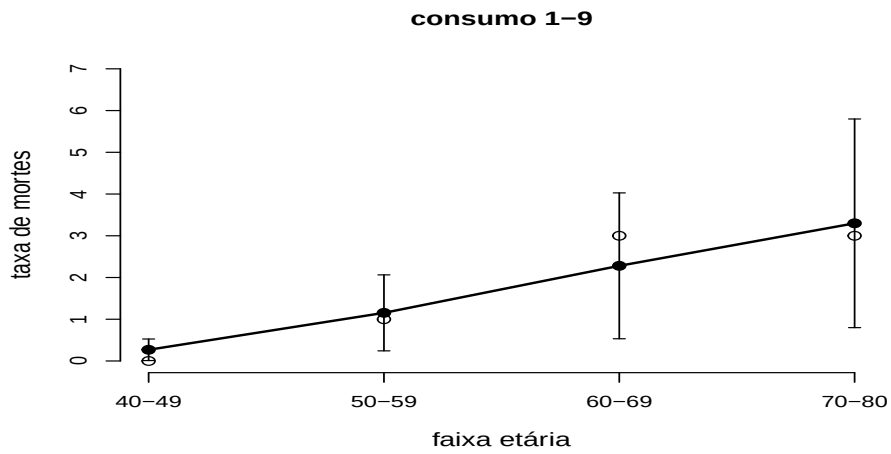
Est. pontuais e intervalares de alguns parâm.: modelo 2

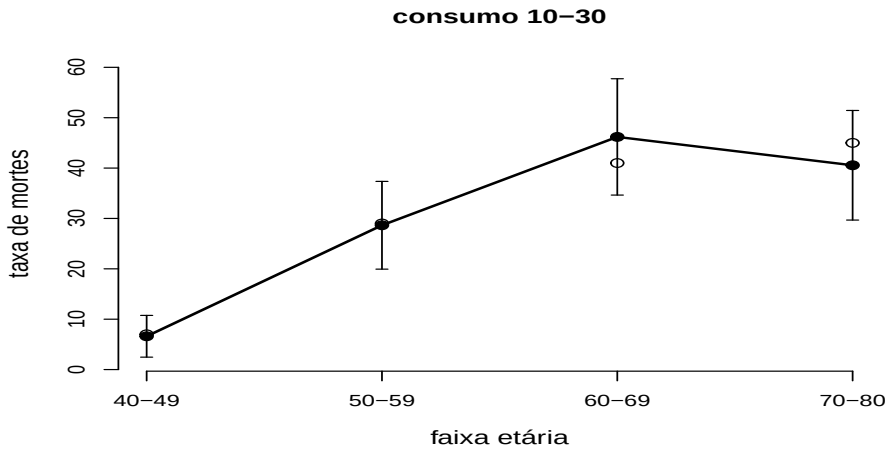


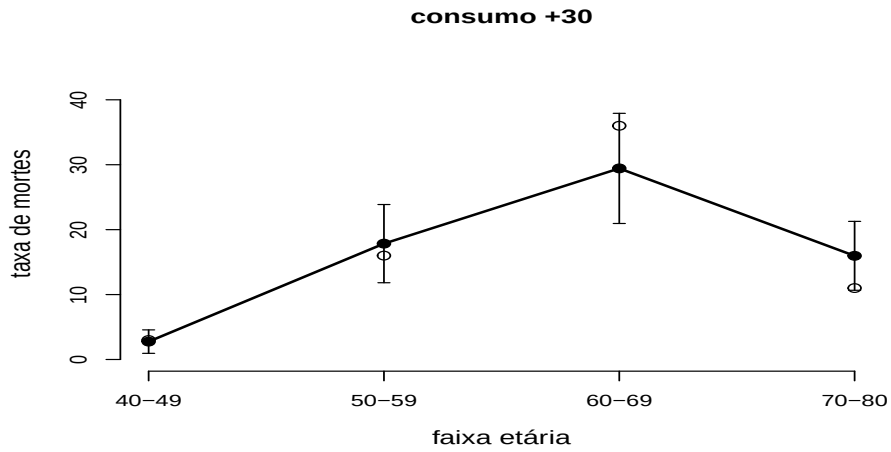


Médias preditas e observadas: modelo 2

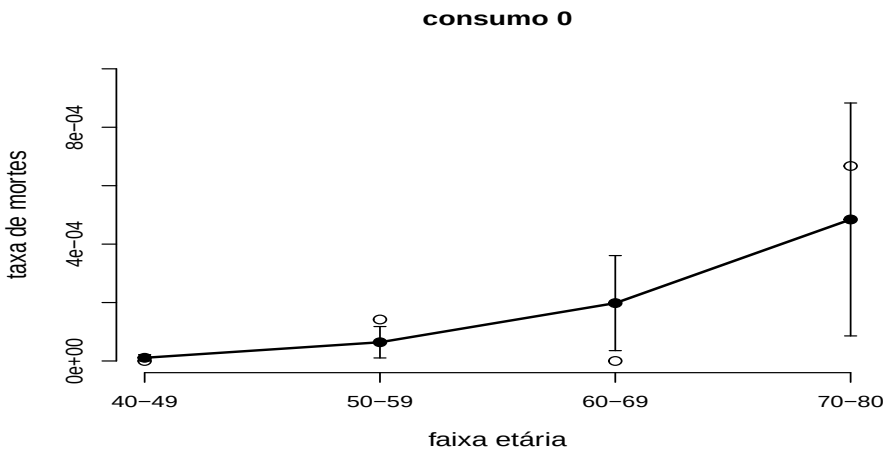


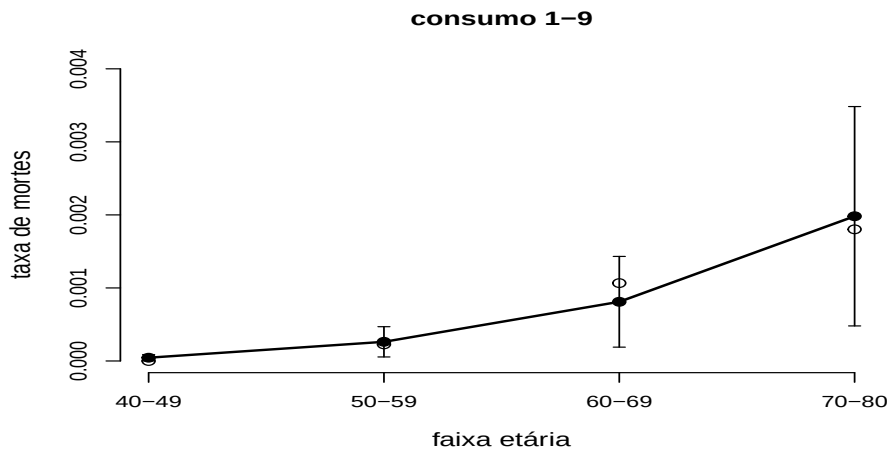


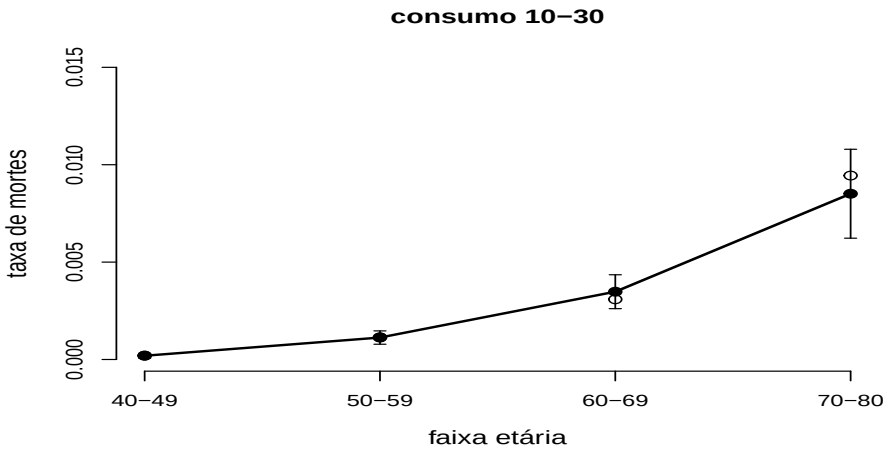


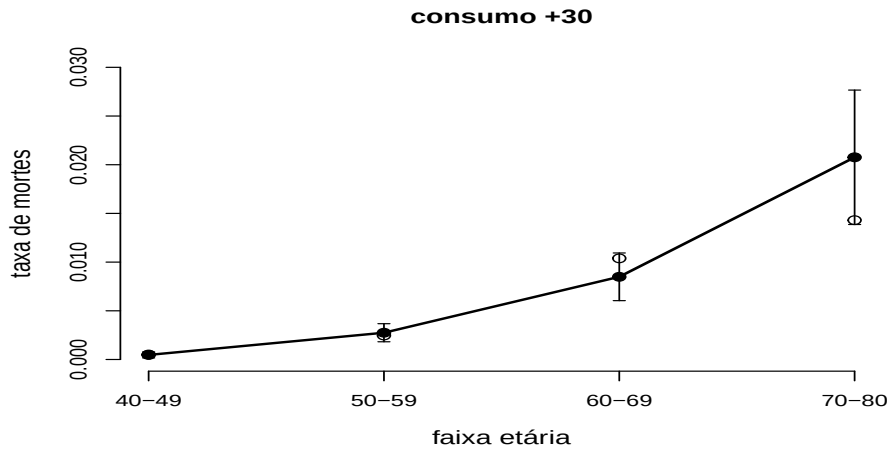


Taxas preditas e observadas: modelo 2









Consumo	Faixa etária	Observado	Estimativa	EP	IC(95%)
0	40-49	0,000	≈ 0,000	≈0,000	≈[0,000 ; 0,000]
0	50-59	0,000	≈0,000	≈0,000	≈[0,000 ; 0,000]
0	60-69	0,000	≈0,000	≈0,000	≈[0,000 ; 0,000]
0	70-80	0,001	≈0,000	≈0,000	≈[0,000 ; 0,001]
1-9	40-49	0,000	≈0,000	≈0,000	≈[0,000 ; 0,000]
1-9	50-59	0,000	≈0,000	≈0,000	≈[0,000 ; 0,000]
1-9	60-69	0,001	≈ 0,001	≈0,000	[0,000 ; 0,001]
1-9	70-80	0,002	0,002	0,001	≈[0,000 ; 0,003]
10-30	40-49	0,000	≈0,000	≈0,000	≈[0,000 ; 0,000]
10-30	50-59	0,001	0,001	≈0,000	[0,001 ; 0,001]
10-30	60-69	0,003	0,003	≈0,000	[0,003 ; 0,004]
10-30	70-80	0,009	0,009	0,001	[0,006 ; 0,011]
+30	40-49	0,001	≈0,000	≈ 0,000	≈[0,000 ; 0,001]
+30	50-59	0,002	0,003	≈0,000	[0,002 ; 0,004]
+30	60-69	0,010	0,008	0,001	[0,006 ; 0,011]
+30	70-80	0,014	0,021	0,004	[0,014 ; 0,028]

Conclusões

- O modelo indicou a inexistência de interação.
- Quanto maior o consumo de cigarros e/ou a idade, maior a taxa de morte por câncer de pulmão.
- Outra possibilidade: gerar uma estrutura multinomial e usar modelos para esse tipo de estrutura.
- O teste de Wald para testar (no modelo sem interação com um coeficiente δ para a variável t_{ij}) $H_0 : \delta = 1$ vs $H_1 : \delta \neq 1$, resultou em $q_c = 1,89(0,1689)$, o que sugere que a variavel pessoas/ano deve ser introduzida como um "offset".