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

Prof. Caio Azevedo

Exemplo 14: tempo de sobrevivências de bactérias

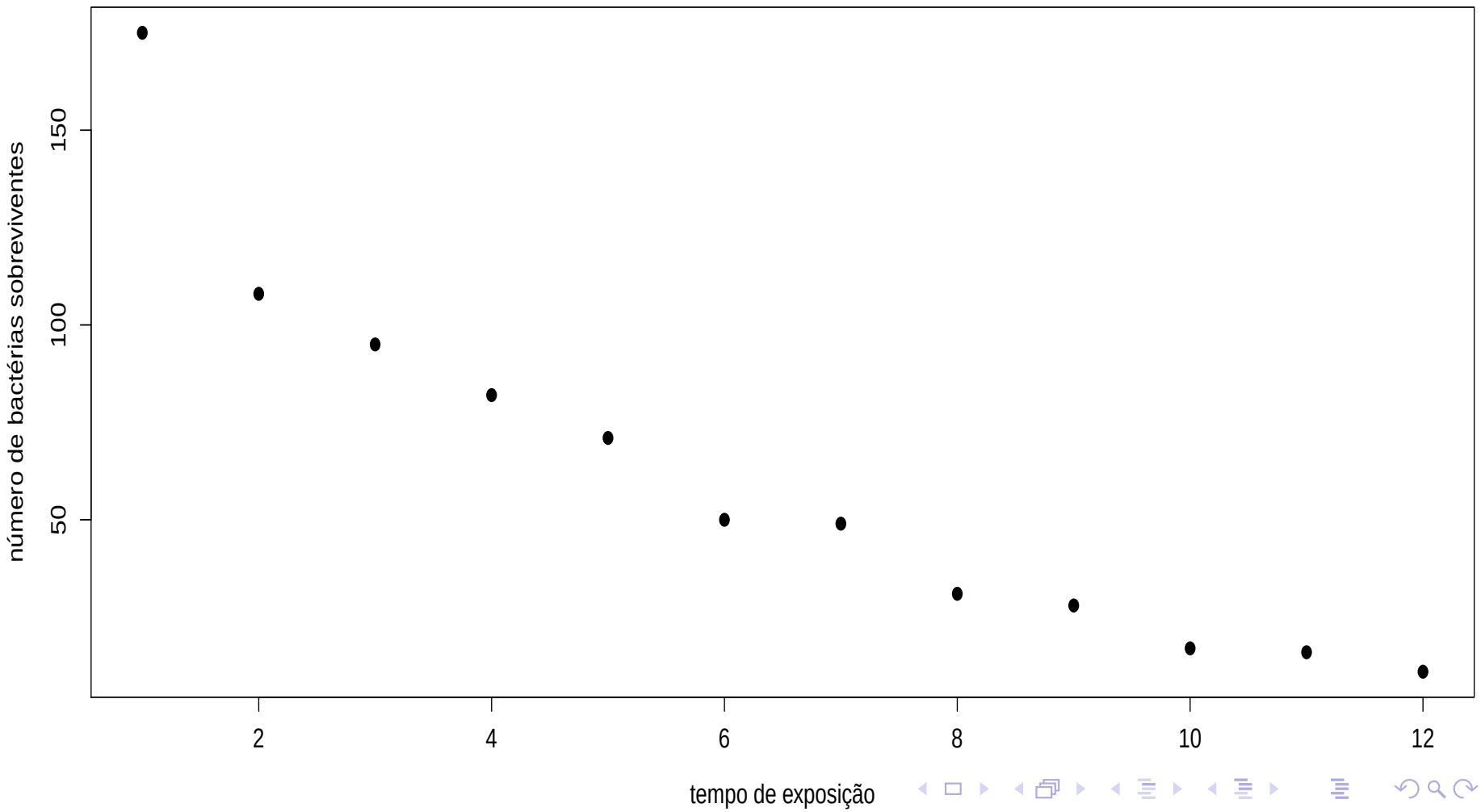
- Os dados correspondem ao número de bactérias sobreviventes em amostras de um produto alimentício segundo o tempo (em minutos) de exposição do produto à uma temperatura de $300^o F$.
- Nessas amostras de alimentos foram feitas 12 medições, a cada minuto, contabilizando a quantidade de bactérias vivas (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 $P(Y_i = k) > 0$ e $P(Y_i \in [r_1, r_2]), \forall k \in \{0, 1, ...\}, r_1, r_2 < 0, r_i \leq r_2, i = 1, 2, ...12$.
- 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 $\overline{x} = \frac{1}{12}\sum_{i=1}^{12} x_i$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. $\mathcal{E}(Y_i) = \mathcal{V}(Y_i) = \mu_i = e^{\beta_0 + \beta_1 x_i}$.
- 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 $300^o F$.
- Tem-se um modelo de regressão Poisson log-linear (função de ligação log).

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 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.
- 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.

Inferência para o modelo

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

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

Inferência para o modelo

- Como $\mu_i = e^{\eta_i}$, temos que a vogverossimilhança, 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

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

pois $\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{X}_i$ e $\mathbf{y} = (y_1, ..., y_n)'$ e $\boldsymbol{\mu} = (p_1, ..., p_n)'$ (**exercício**).

Inferência para o modelo

- Além disso, $\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i(\boldsymbol{\beta}) \dfrac{\partial \eta_i}{\partial \boldsymbol{\beta}}$, 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

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

Inferência para o modelo

- Logo, a informação de Fisher corresponde à

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

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

- **Exercício: obter a forma matricial de $\mathbf{I}(\beta)$.**

- Repetir os desenvolvimentos considerando $\sqrt{\mu_i} = \mathbf{X}_i'\beta$ (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:

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

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

Comentários

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

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

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 $\mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$.

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

Estimativas dos parâmetros (os testes se referem à nulidade de cada parâmetro)

Parâmetro	Estimativa	EP	Estat. Z_t	p-valor
β_0	3,818	0,048	79,258	<0,0000
β_1	-0,229	0,013	-18,023	<0,0000

Os dois parâmetros são significativos. Além disso, $D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}) = 8,42$, para $n - p = 12 - 2 = 10$ graus de liberdade, o que leva à um p-valor $= 0,5877$, sugerindo um bom ajuste do modelo.

Gráfico de envelopes para os RCD's

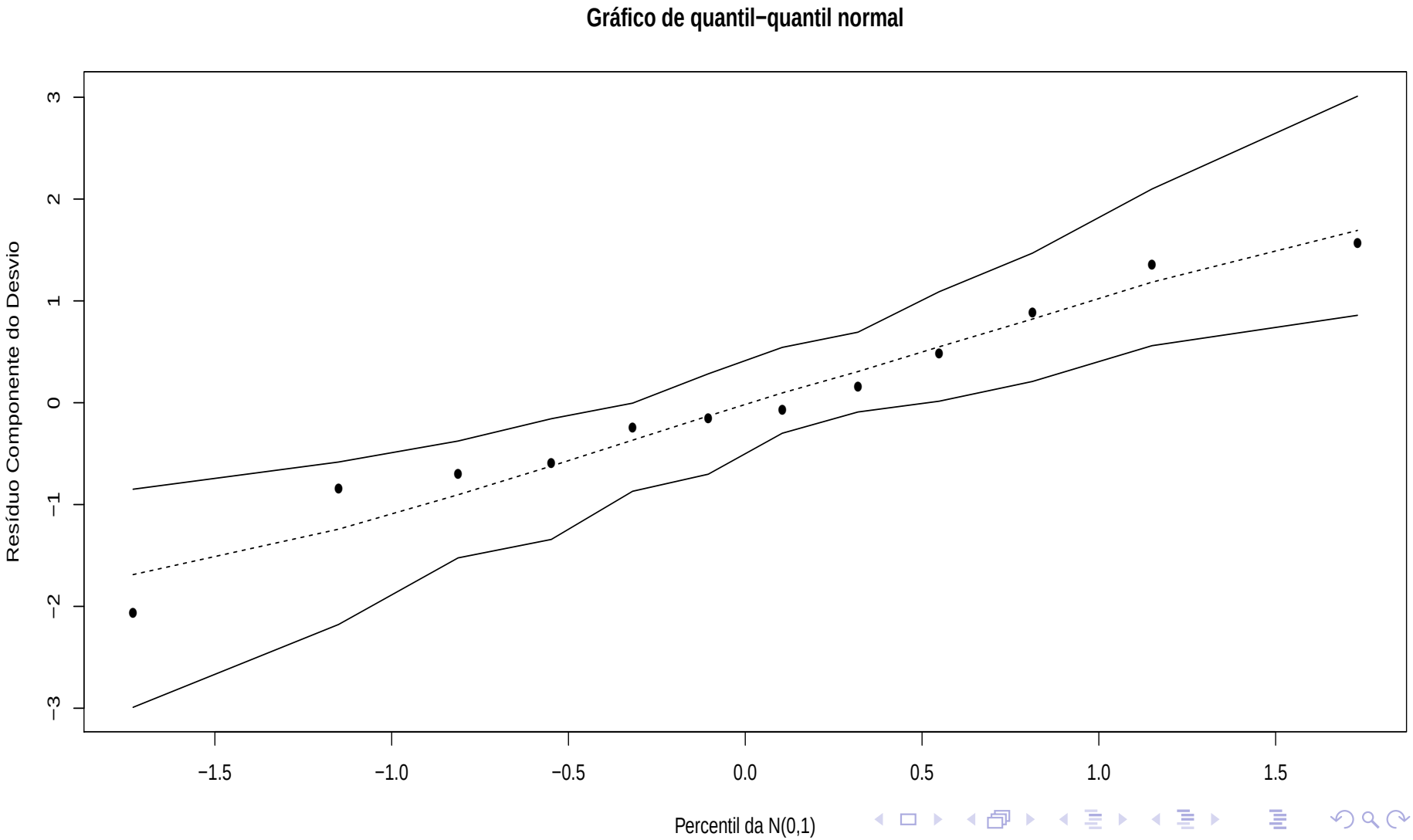
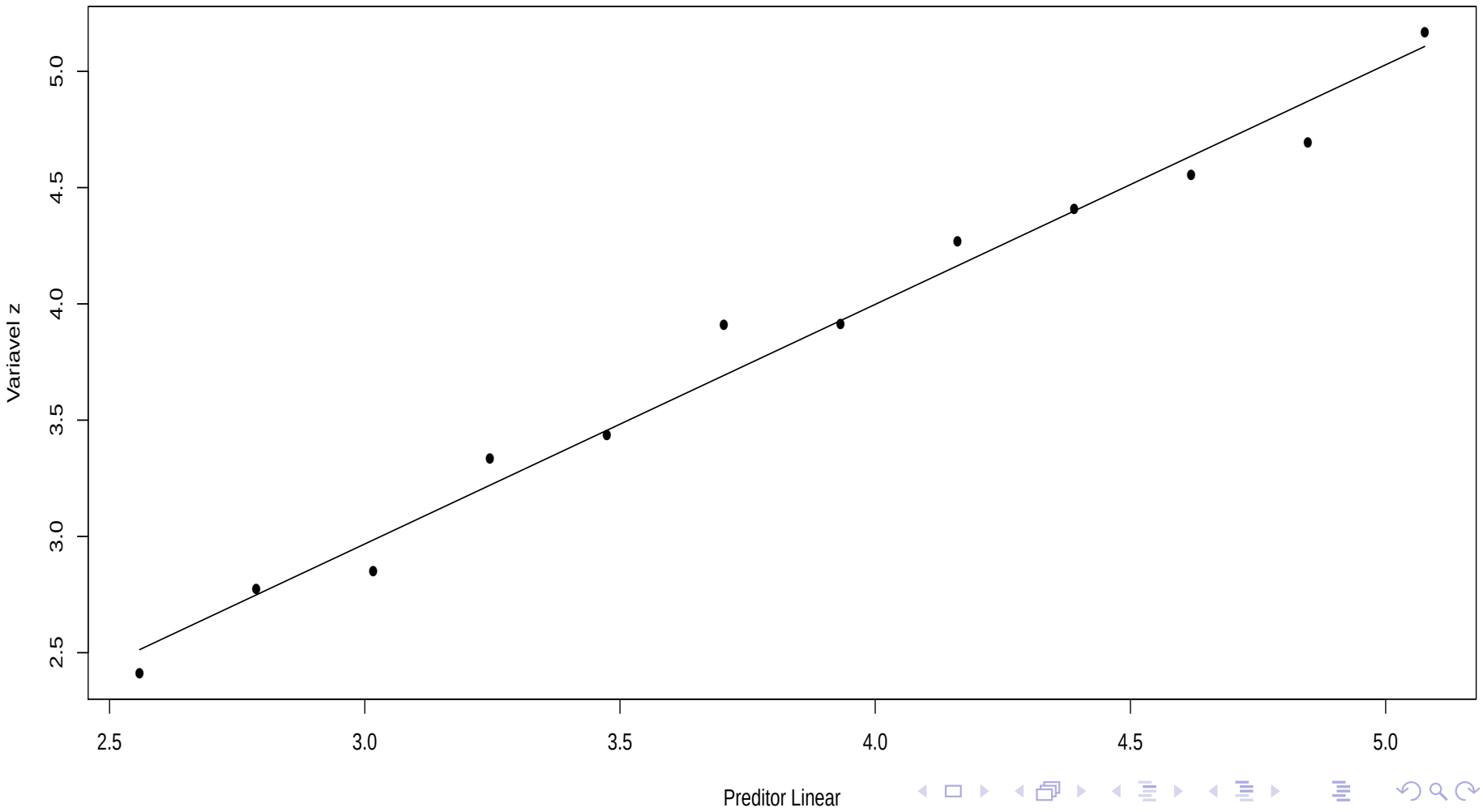
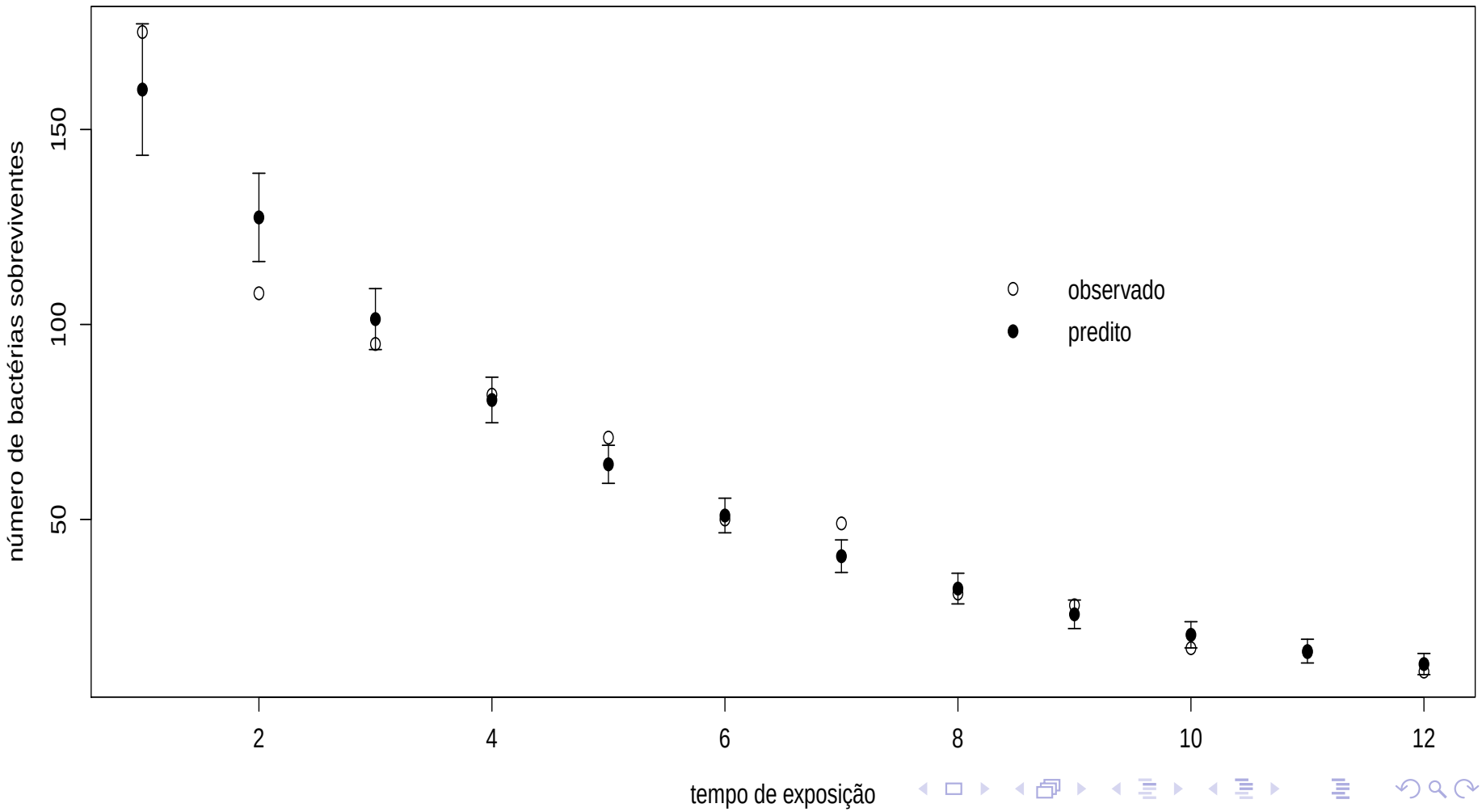


Gráfico da variável Z vs preditor linear



Médias observadas e preditas pelo modelo



Comentários

- Apesar do modelo ter se ajustado bem ao conjunto de dados, o número médio de bactérias sobreviventes não foi apropriadamente predito (nem pontual nem intervalarmente) pelo modelo, para alguns instantes.
- Alternativa: manter o modelo substituindo-se o preditor linear (η_i) por um preditor não linear (modelos não lineares generalizados) e/ou considerar efeitos aleatórios.

Comentários

- De acordo com o modelo, há ma 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 $IC(e^{-\beta_1}, 95\%) = [0,775; 0,815] = [77,5\%; 81,5\%]$.

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

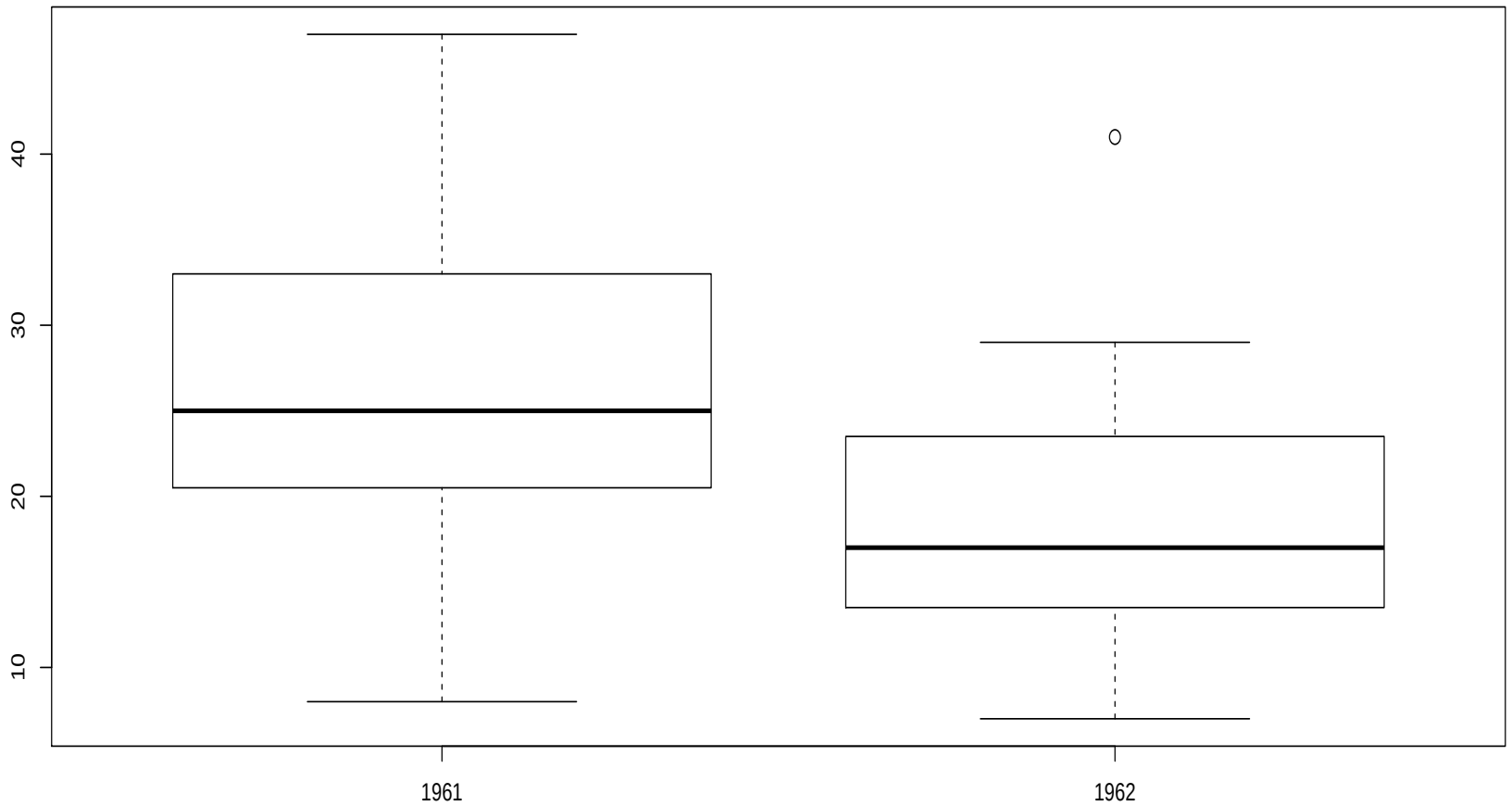
$$\begin{aligned} Y_{ij} &\overset{ind.}{\sim} \text{Poisson}(\mu_i), i = 1, 2, j = 1, ..., 43 \\ \ln \mu_i &= \mu + \alpha_i, \alpha_1 = 0 \end{aligned}$$

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 ($\mu_2 = \mu_1 e^{\alpha_2}$).

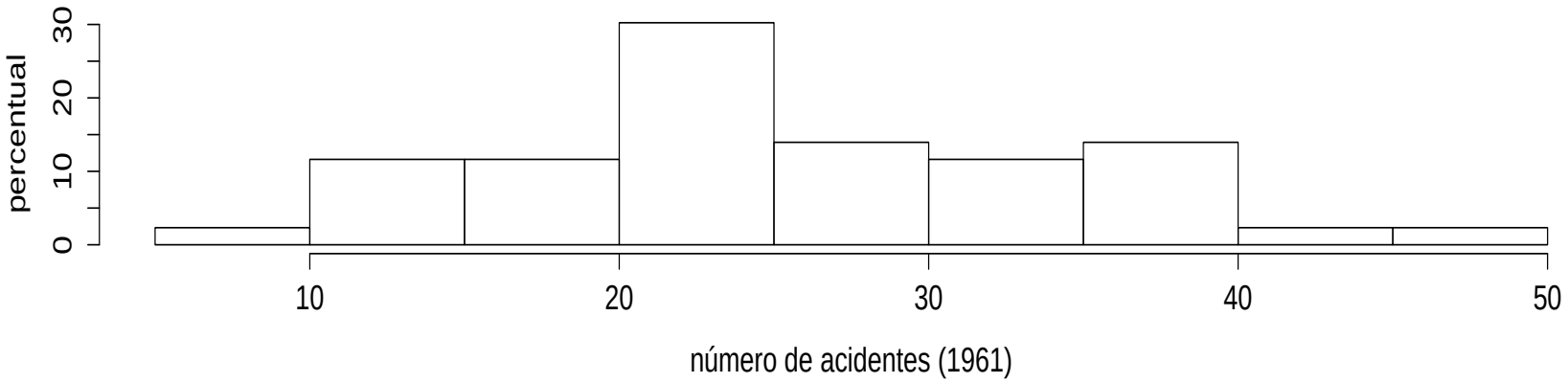
Medidas Resumo

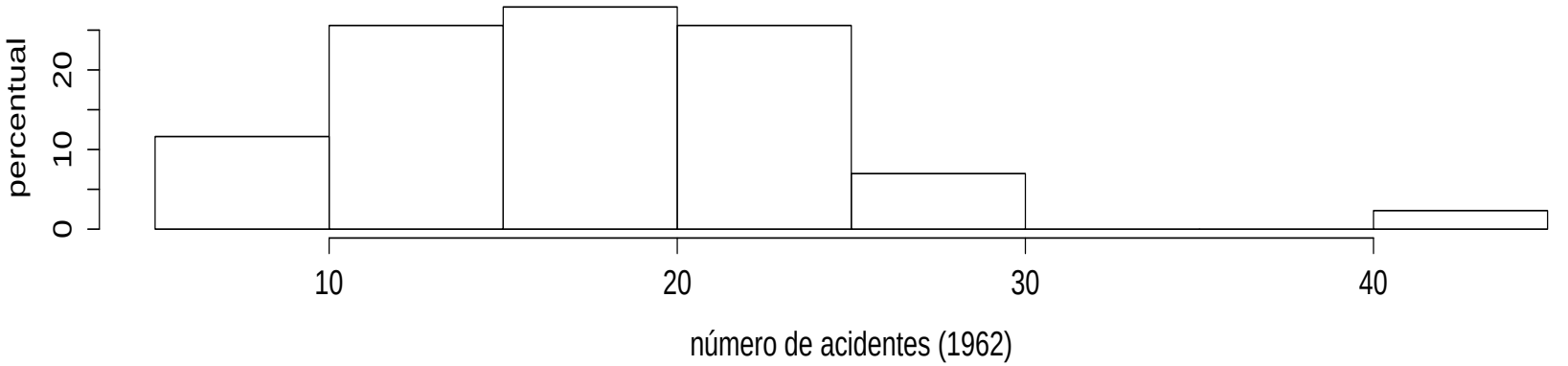
Ano	Média	Var.	DP	CV(%)	Mín.	Med.	Máx.
1961	26,05	82,66	9,09	34,91	8,00	25,00	47,00
1962	18,05	44,71	6,69	37,05	7,00	17,00	41,00

Boxplots do número de acidentes por ano



Histogramas do número de acidentes por ano





Estimativas dos parâmetros (os testes se referem à nulidade de cada parâmetro)

Parâmetro	Estimativa	EP	Estat. Z_t	p-valor
β_0	3,2599	0,0299	109,097	<0,0000
β_1	-0,3669	0,0467	-7,856	<0,0000

Os dois parâmetros são significativos. Entretanto, $D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}) = 235,17$, para $n - p = 86 - 2 = 84$ graus de liberdade, o que leva à um p-valor $< 0,0001$, indicando que o modelo não se ajustou bem aos dados.

RCD's $\times$ índice da observação

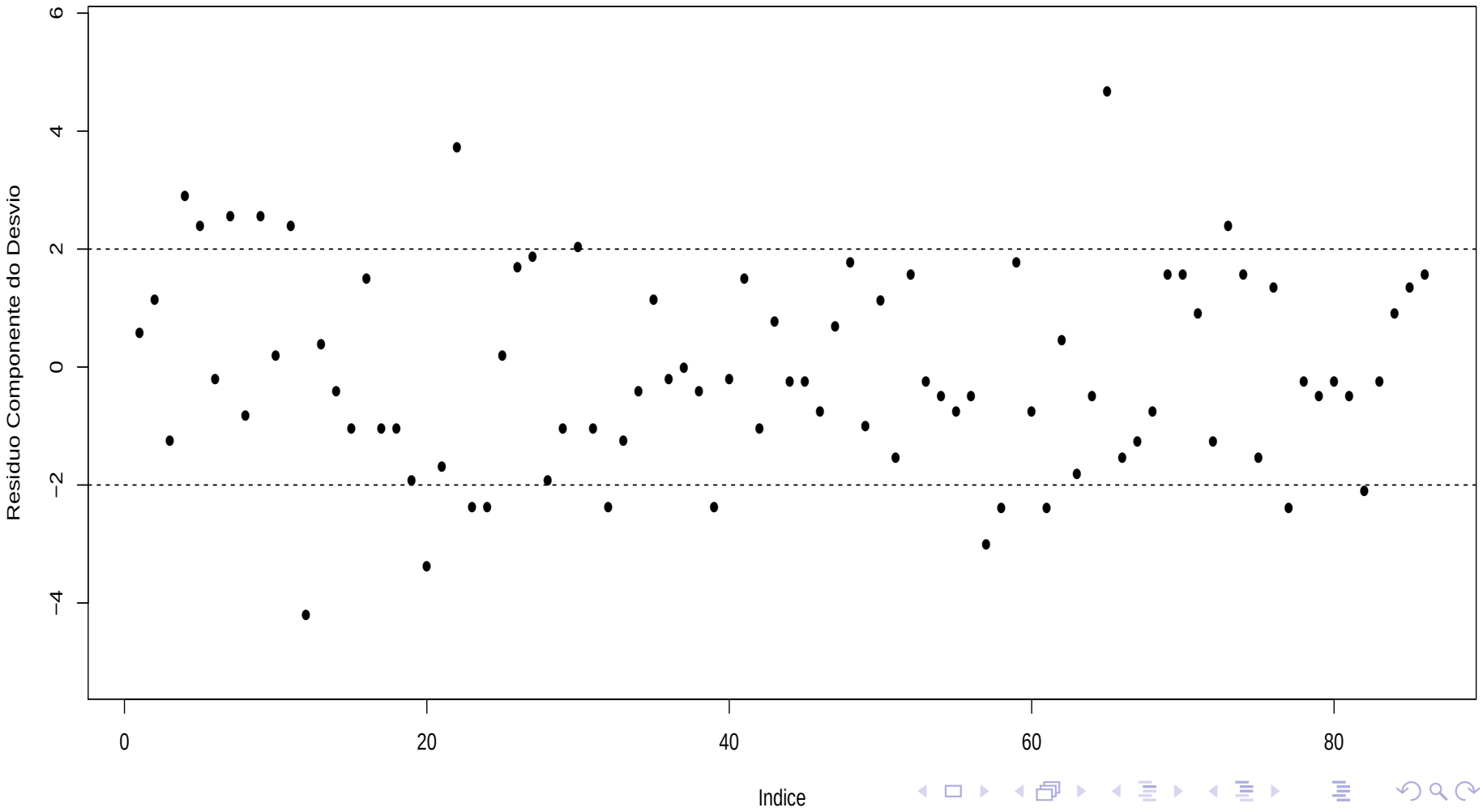
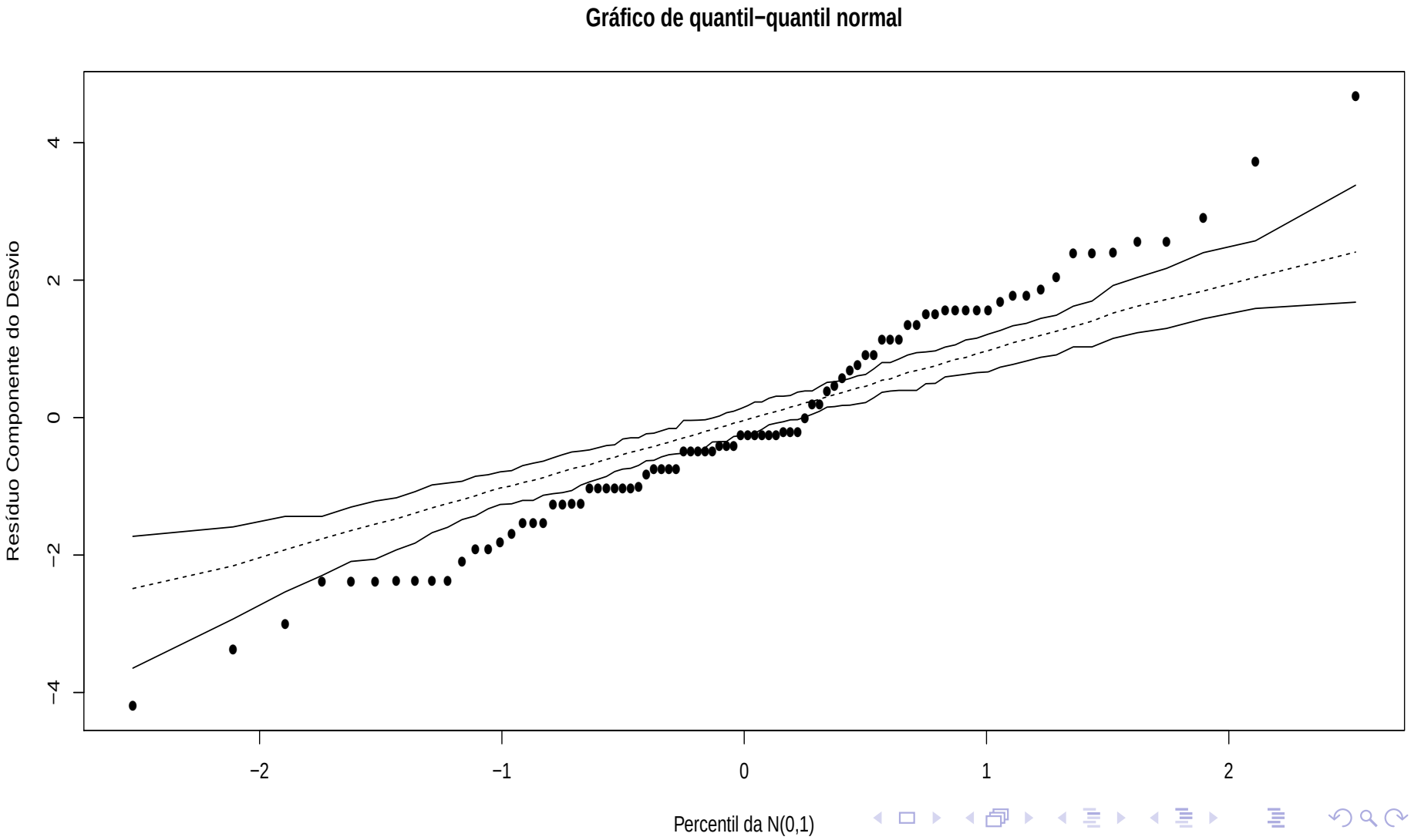
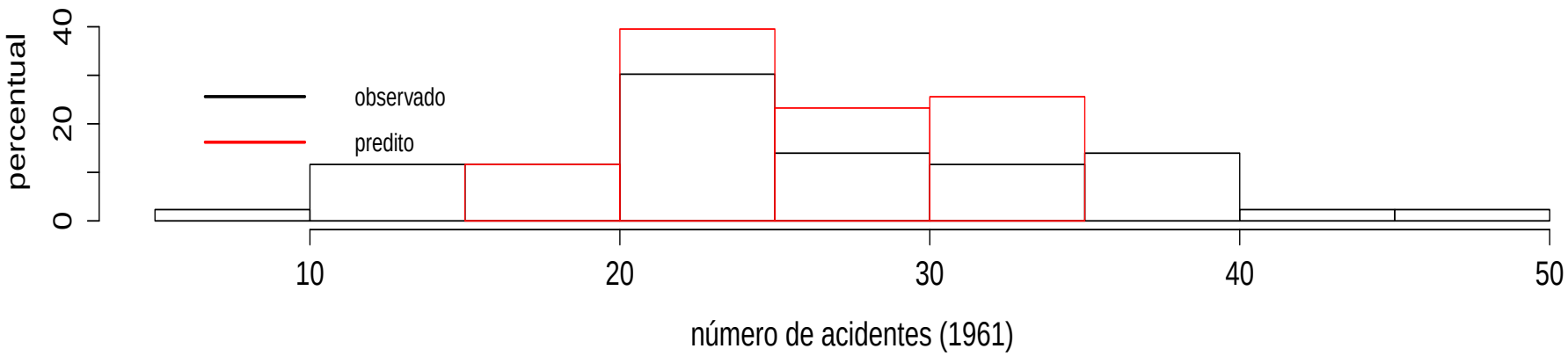
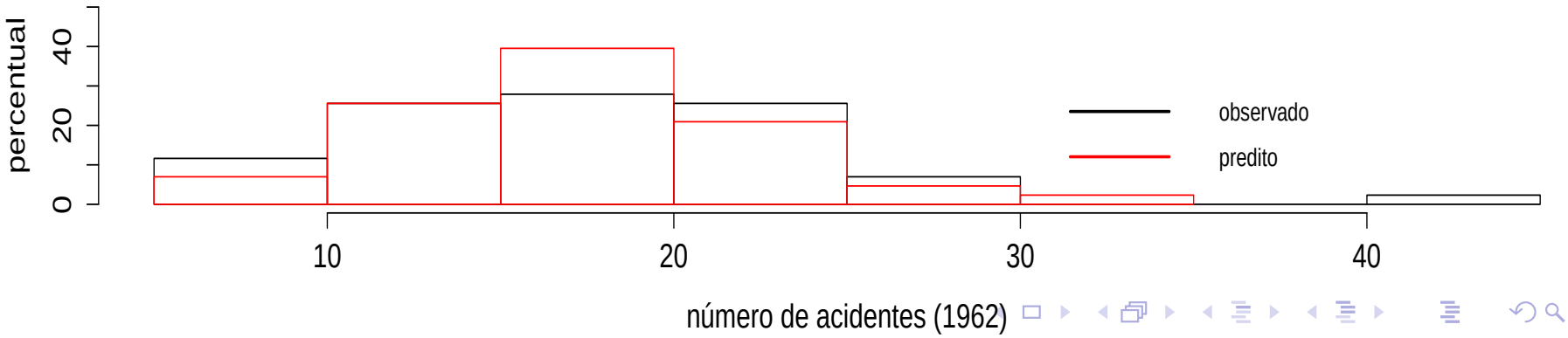


Gráfico de envelopes para os RCD's



Distribuições observadas e preditas pelo modelo





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 não paramétricos (sem supor alguma distribuição específica para a variável resposta) que contemple a superdispersão, outros modelos de regressão hetorocedásticos (modelos de superdispersão).

Médias ajustadas

- Caso o modelo tivesse se ajustado bem, deveríamos apresentar as estimativas pontuais e intervalares das médias de cada grupo.

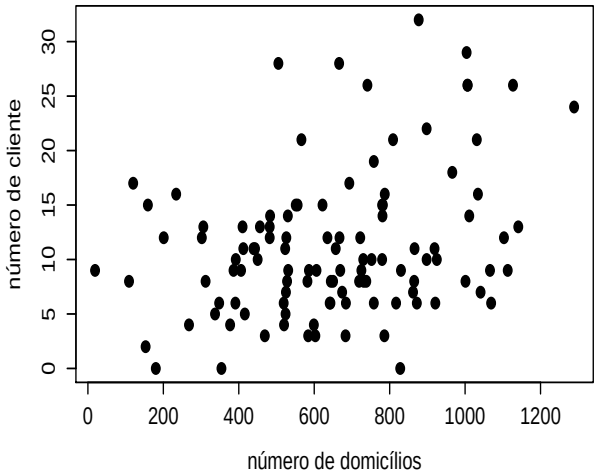
Ano	Est.	DP	IC(95%)
1961	26,05	0,78	[24,52 ; 27,57]
1962	18,05	0,65	[16,78 ; 19,32]

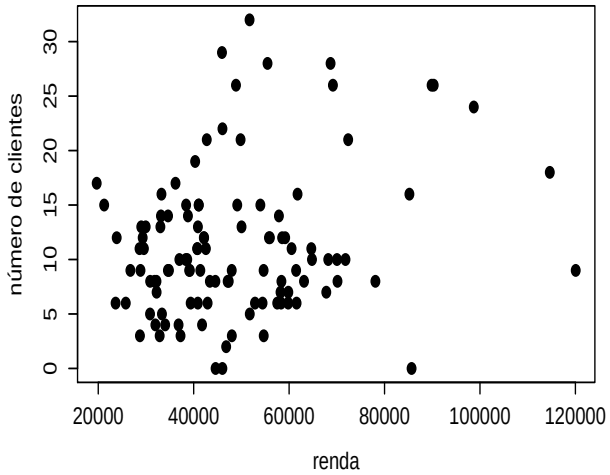
Neste caso, possivelmente, os intervalos de confiança apresentam uma amplitude menor do que deveriam (devido à superdispersão).

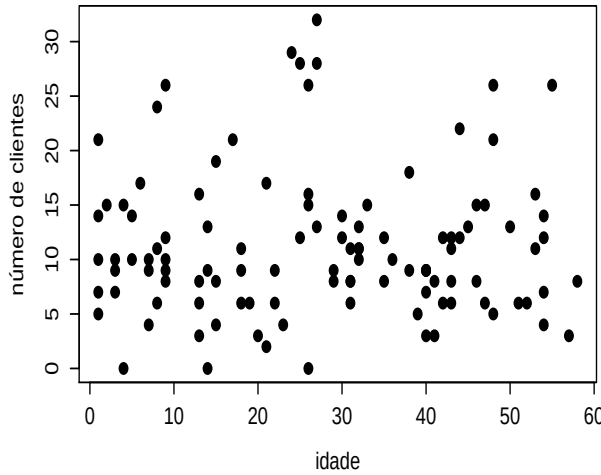
Exemplo 15: perfil dos clientes de uma loja

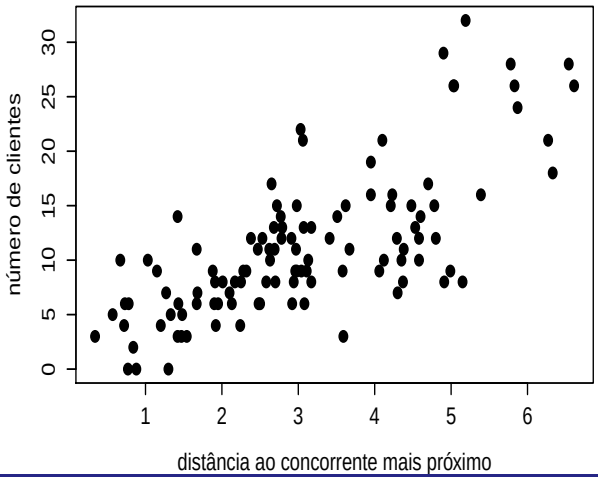
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável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

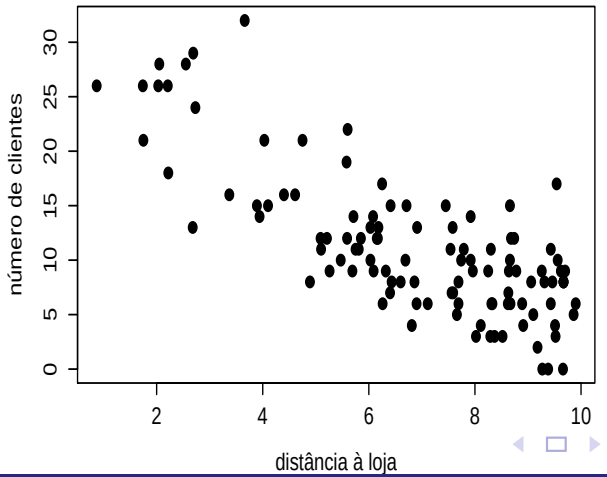
Gráficos de dispersão











Modelo (completo)

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

- x_{ji} : valor da variável explicativa j, associada à área i e $\overline{x}_j = \frac{1}{110}\sum_{i=1}^{110} x_{ji}$, $j = 1, 2, ..., 5$.
- β_j : parâmetro associado ao impacto da covariável j no valor esperado do número de clientes.
- e^{β_0} : número esperado de clientes quando cada uma das covariáveis for igual à sua respectiva média.

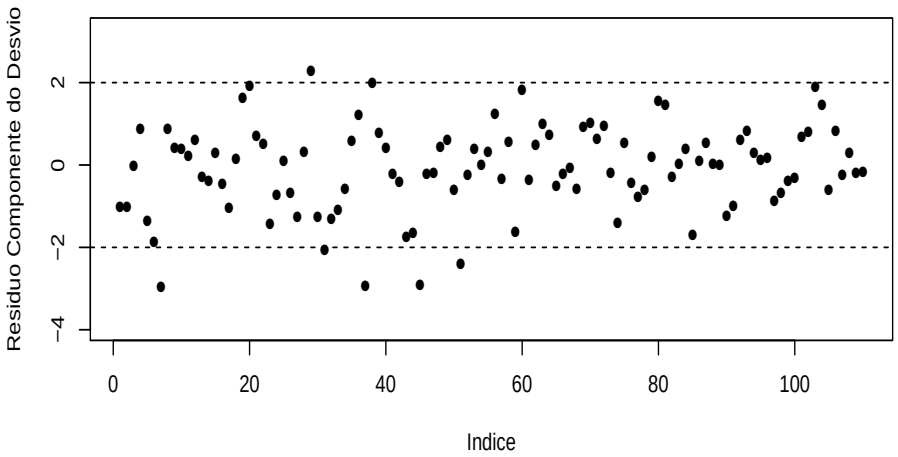
Modelo (completo) (cont.)

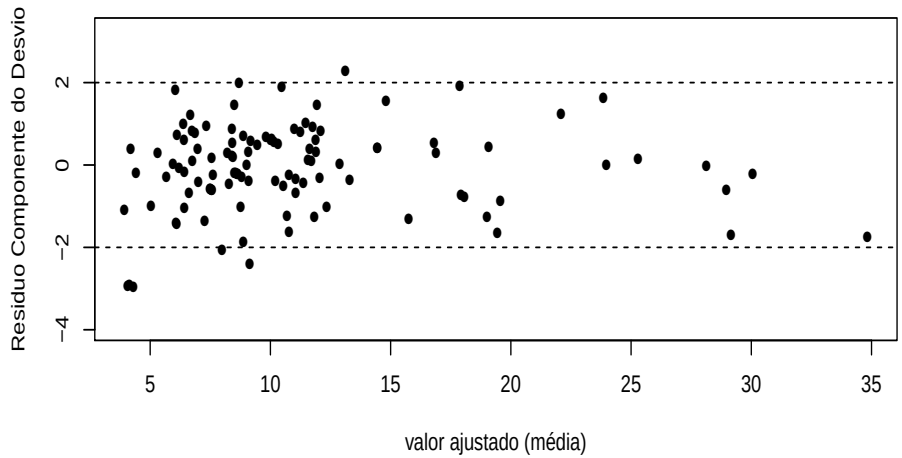
- e^{β_j} : incremento (positivo ou negativo) no número esperado de clientes para o aumento em uma unidade da covariável $j, j = 1, 2, ..., 5$.
- $\mathcal{E}(Y_i) = e^{\eta_{ij}}$ e $\mathcal{V}(e^{\eta_{ij}}) = e^{\eta_{ij}}$, em que $\eta_{ij} = \beta_0 + \beta_1(x_{1i} - \overline{x}_1) + \beta_2(x_{2i} - \overline{x}_2) + \beta_3(x_{3i} - \overline{x}_3) + \beta_1(x_{4i} - \overline{x}_4) + \beta_1(x_{5i} - \overline{x}_5)$.

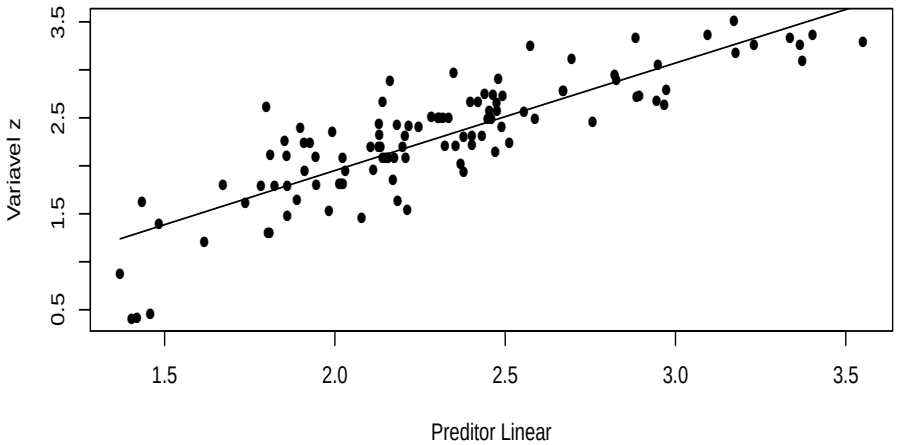
Abordagem adotada

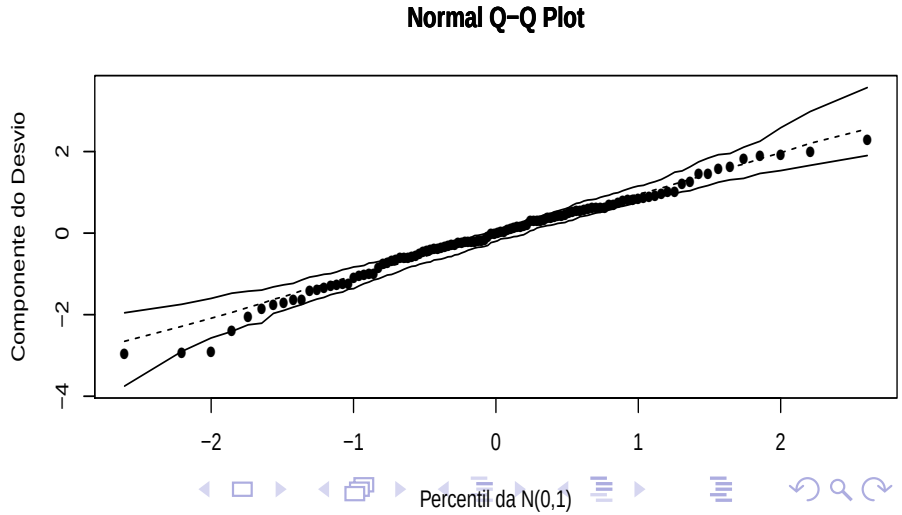
- Vamos utilizar o método stepwise para selecionar o modelo mais apropriado (em relação a selecionar parâmetros, associados às covariáveis, que seja significativos).
- Primeiramente, contudo, vamos avaliar a qualidade do ajuste do modelo completo.

Análise de diagnóstico









Estimativas dos parâmetros (os testes se referem à nulidade de cada parâmetro)

- O método de seleção stepwise, indicou que todas as variáveis contribuem para explicar a variabilidade do número médio de clientes entre os domicílios.
- Além disso, $D(\mathbf{y}; \widetilde{\boldsymbol{\mu}}) = 8,42$, para $n - p = 110 - 6 = 104$ graus de liberdade, o que leva à um p-valor $= 0,2170$, sugerindo um bom ajuste do modelo, o que corrobora as conclusões obtidas à partir dos gráficos de diagnóstico.

Estimativas dos parâmetros (os testes se referem à nulidade de cada parâmetro)

Parâmetro	Estimativa	EP	Estat. Z_t	p-valor
β_0	2,2988	0,0315	72,920	$<$ 0,0001
β_1 (ndom.)	0,0006	0,0001	4,262	$<$ 0,0001
β_2 (renda)	-1,1690 $\times 10^{-5}$	2,112 $\times 10^{-6}$	-5,534	$<$ 0,0001
β_3 (idade)	-0,0037	0,0018	-2,091	0,0365
β_4 (dist. conc.)	0,1684	0,0258	6,534	$<$ 0,0001
β_5 (dist. loja)	-0,1288	0,0162	-7,948	$<$ 0,0001

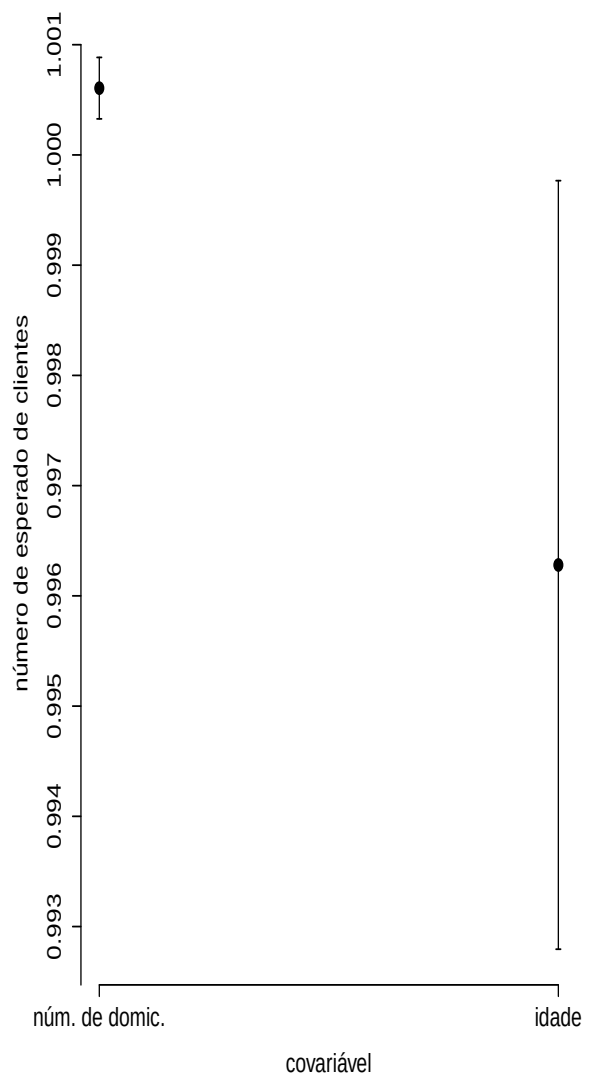
Todos os parâmetros são significativos, de fato.

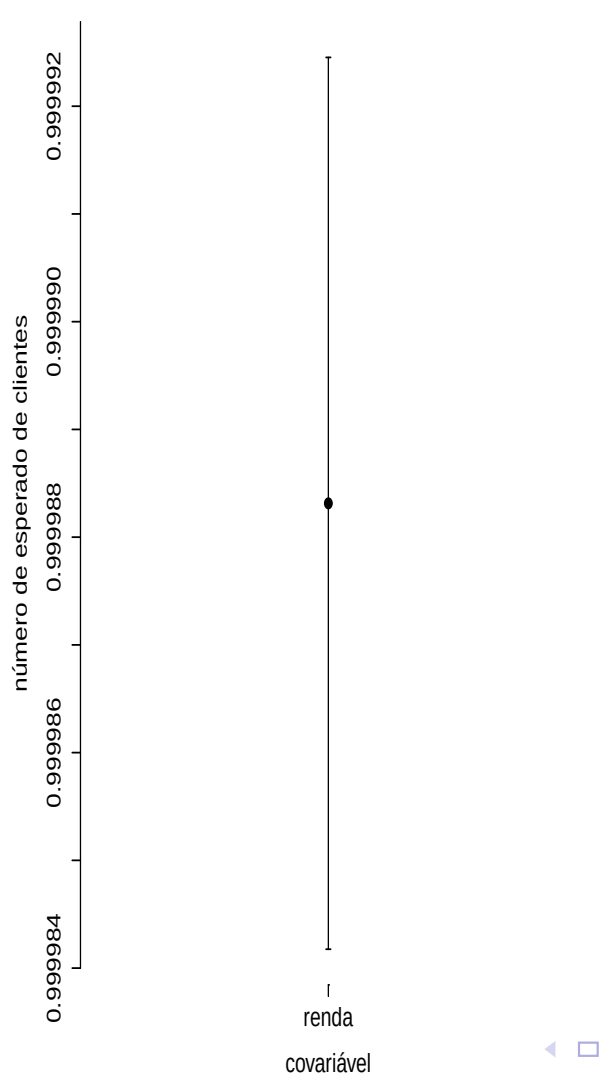
Número e incrementos esperados

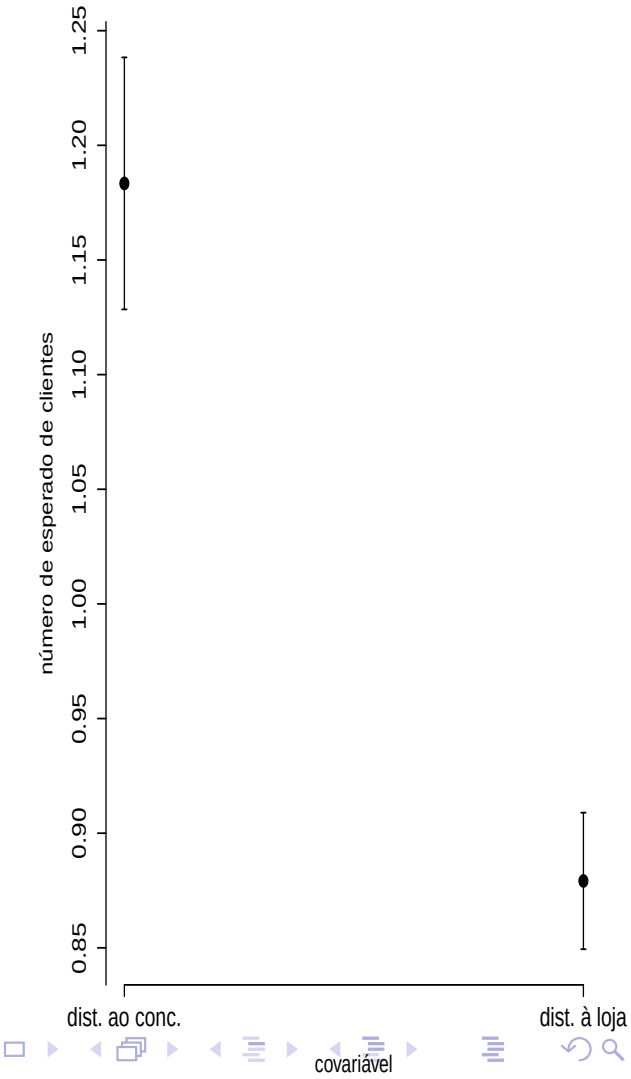
Parâmetro	Estimativa	EP	IC(95%)
e^{β_0}	9,9618	0,0995	[9,7669 ; 10,1569]
e^{β_1} (ndom.)	1,0006	0,0001	[1,0003 ; 1,0009]
e^{β_2} (renda)	0,9999	<0,0001	[0,9998 ; 1,0000]
e^{β_3} (idade)	0,9963	0,0018	[0,9927 ; 0,9997]
e^{β_4} (dist. conc.)	1,1834	0,0280	[1,1284 ; 1,238]
e^{β_5} (dist. loja)	0,8792	0,0152	[0,8494 ; 0,9089]

Em geral, espera-se que a loja tenha, aproximadamente, 10 clientes de uma região com as seguintes características (valores médios aproximados): ndom= 648; renda= 48837 ; idade=27; dist. conc. = 3; dist. loja = 7.

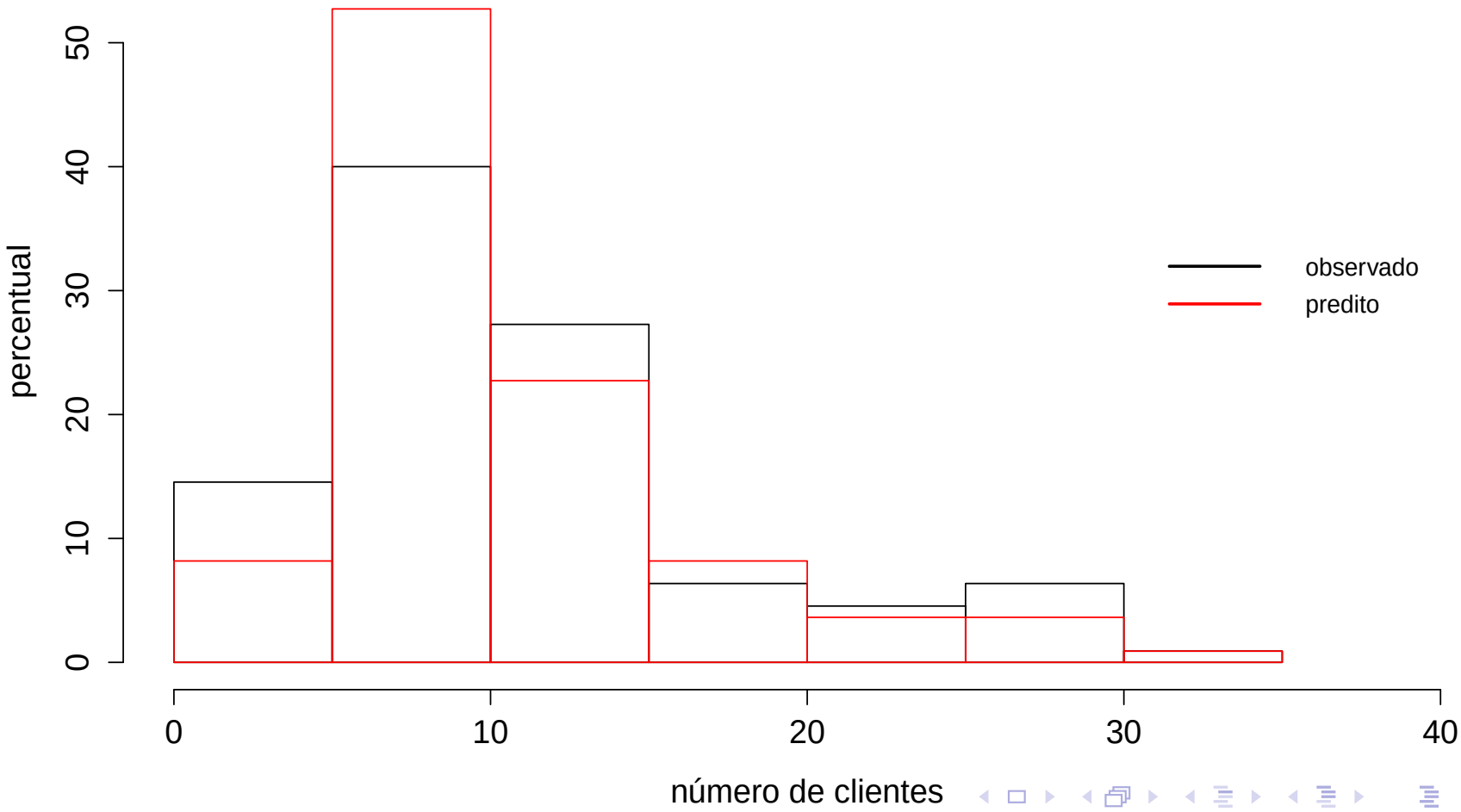
Número de clientes e incrementos esperados







Distribuição observada e predita



Comentários

- O modelo de regressão se ajustou bem aos dados.
- O poder preditivo do modelo se mostrou bom para áreas com números elevados de clientes e não muito bom para as outras áreas.
- O poder preditivo do modelo pode ser melhorado com a inserção de outras covariáveis e/ou com a utilização de estruturas de regressão segmentadas e não paramétricas.